Special Issue for International Conference of Advanced Materials Engineering and Technology (ICAMET 2013), 28-29 November 2013, Bandung Indonesia

Advances in Environmental Biology, 7(12) October Special Issue 2013, Pages: 3861-3869

Roles of Imputation Methods for Filling the Missing Values: A Review

[1,2,3]M.N. Norazian Ramli, [1]Yahaya, A.S., [1]Ramli, N.A., [1]Yusof, N.F.F.M., [3] Abdullah, M.M.A.

[1]Clean Air Research Group, Environmental and Sustainable Development Section, School of Civil Engineering, UniversitiSains Malaysia, Engineering Campus, NibongTebal, 14300, Pulau Pinang, Malaysia.
[2]Environmental Engineering Department, Universiti Malaysia Perlis, 02600 Jejawi, Perlis, Malaysia.
[3]Center of Excellence Geopolymer& Green Technology (CEGeoGTech), School of Material Engineering, Universiti Malaysia Perlis, 02600, Perlis, Malaysia.

ARTICLE INFO

Article history:
Received 11 September 2013
Received in revised form 21 November 2013
Accepted 25 November 2013
Available online 3 December 2013

Key words:
Imputation; single imputation; multiple imputations; missing data mechanism

ABSTRACT

Missing data are often encountered in many areas of research. Complete case analysis and indicator method can lead to serious bias. One of the comforting methods is implementation of imputation methods. The main purpose of this paper is to review the agreement of imputation methods as the most widely used method for filling missing observations. Single and multiple imputations had certain criteria to be satisfied before adoption. Single imputation methods works excellently in short gap length of missing data. Embracing single imputation method to the long gap of missing data will cause systematically error since the reflection of uncertainty is not covered. Multiple imputations were recognized as the superior method for missing-at-random (MAR) data set. Although the dominance of multiple imputations was known, the adoption of these imputations needs thorough understanding on the algorithms especially in designing a suitable method to perform the imputations. The reviews on assessment of available imputation software were also presented to compare the practicality of the software.

INTRODUCTION

Generally, the problem of missing data emerges in many areas of research such as environmental field [17,26,25], statistical survey [22,15], medical study [30,36] and indusrial databases [19]. Mondelo [23] described "missing data" as a generic term referring to two different situations. First situation is usually to describe indefinitely missing data those arising from a non-response in survey or an interview and also those arising from a breakdown in measurement instruments. At secondly situation is due to external (technical) reason, presumably due to change. Missing data is problematical. Discontinuities of dataset may lead to significant obstacle for analyzing the findings [17]. Sometimes the missingnessis due to unknown reasons, or error and omission when the data are recorded and transferred [27]. However, there are a number of evident reasons, including imperfect procedures of manual data entry, incorrect measurements, and equipment error [7].

Missing observations make it difficult for analysts to realize the data analysis. Types of problems that are usually associated with missing values are [12] 1) loss of efficiency; 2) complications in handling and analyzing the data; 3) bias resulting from differences between missing and complete data (bias estimates)and; 4) reduction of statistical power (inefficient estimates). Decision on selecting an appropriate method for handling missing observations on time series depends on the missing data pattern and on the missing-data mechanism [26]. However, if the observations were more than 60 percent missing, no method was found suitable to cure the data [7].

Embracing the imputation techniques are commonly used for the treatment of missing data. However there are few challenges in adoption of this technique that are 1) to maximize the use of available data in order to minimize the mean square error for univariate statistics and to preserve covariance structures in multivariate data sets [22]; 2) to include in the variance estimates of the uncertainty caused by the use of imputed data, i.e. synthetic (not really observed) data [21].

Corresponding Author: M.N. Norazian Ramli, Clean Air Research Group, Environmental and Sustainable Development Section, School of Civil Engineering, University Sains Malaysia, Engineering Campus, Nibong Tebal, 14300, Pulau Pinang, Malaysia.

The main purpose of this paper is to review the methods of imputation i.e. single and multiple imputationsand their limitation. Unlike other literatures that focused on specific imputation methods on solving the missingness, this manuscript reviewed on the adoption of single and multiple imputations in various field and enhancement for better prediction. The last section of this paper will present the types of software that are currently accessible for calculating missing data for the ease of readers' references.

2.0 Missing Data Mechanisms:

There are 2 important types of missing data describe by Little and Rubin [21] known as ignorable and non-ignorable. Non-ignorable is where the probability of missing items is dependent to the values of observations whereas ignorable missing data is where the probability of missing items is not dependent upon the value of observations.

There are three types of missing data mechanisms that integrated with ignorable missing data. The missing-data mechanism describes the connection between missingness and the values of variables in the data matrix. As explained by Schafer [32], given an observed variable Y as Y_{obs} and a missing variable Y as Y_{mis}, it can be said that $Y = [(Y_{obs}, Y_{mis})]$.

The first type is missing completely at random (MCAR). It is defined as the missingness occurs at random across the whole data sets. Thus, the probability of missing value is independent of bothY_{obs} and Y_{mis}.

The second form is missing at random (MAR). This confusingly form of missing data occurs if the probability of a record having a missing value for an attribute that does not depend on the value of the missing data itself, but could be depend on the observed data [6]. Thus, the probability of missing value is independent of Y_{mis}. Generally, when the missing data are MAR all simple techniques for handling missing data i.e. complete and available case analyses, the indicator method and overall mean imputation, give biased results. Nonetheless, adoption of more sophisticated techniques like single and multiple imputations give unbiased results for MAR form of missing data [6,26,17]. Little and Rubin [21] reported that MAR and MCAR are able to be ignorable because it is possible to adjust for the missingness.

The third form is associated with sampling due to the impossibly to obtain data from a whole population, probability sampling is widely used to obtain data from representative population sample. This form is not considered further [12].

Non-ignorable is where the probability of a missing datum is dependent upon its value. Non-ignorable missing data occurs where the pattern of missingness is such that the missing value of Y cannot be reliably predicted from other dataset variable [12]. One form of non-ignorable is missing not at random (MNAR). It occurs if the probability of a record having a missing value for an attribute that depend on the value of the attribute. If missing data are MNAR, valuable information lost from the data and, there is no universal method of curing the missing data properly [6]. Figure 1 simplifies the important types of missing data.

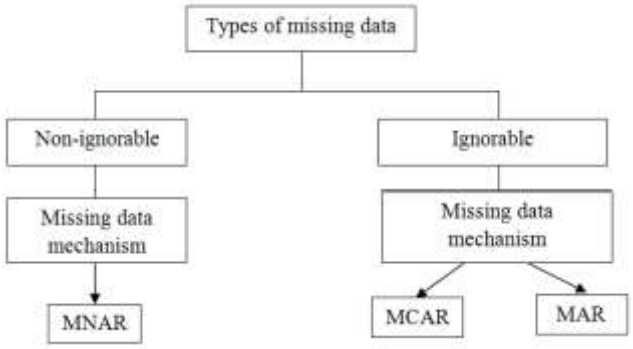

Fig. 1: Classification of missing data.

3.0 Single Imputation:

Imputation is a general and flexible method for handling missing-data problems. It are means or draws from a predictive distribution of the missing values, and require a method of creating a predictive distribution for the imputation based on the observed data [21]. Method of creating complete data via filling in missing value can be classified into single imputation and multiple imputation methods [14]. Single imputation is defined as filling in precisely one value for each missing one. Multiple imputations is a method of generating multiple simulated values for each missing item in order to reflect properly the uncertainty attached to missing data [17].

There are two general approaches to generate the predictive distribution of missing values namely explicit modelling and implicit modelling [21]. Explicit modelling is the predictive distribution that is based on a formal statistical model for example multivariate normal and hence the assumptions are explicit. This method includes mean imputation, regression imputation and stochastic regression imputation.

Mean imputation is filing the means from responding units into the missing data. Regression imputation is the method of replacing the missing values by predicted values from a regression of the missing item on the

items observed for the unit. The last form of explicit modelling is stochastic regression imputation where this method substitutes missing observations by a value predicted by regression imputation plus a residual, drawn to reflect uncertainty in the predicted value [21].

Implicit modelling focuses on algorithm, which implies underlying model [21]. This method includes hot deck imputation, substitution and cold deck imputation. Hot deck imputation involves substitution of individual values drawn from „similar' responding units. Hot deck is common in survey practice and can involve very elaborate schemes for selecting units that are similar for imputation.

Substitution is a method for dealing with unit non response at the fieldwork stage of survey, replaces nonresponding units with alternative units not selected into the sample where as cold deck imputation replaces a missing value of an item by a constant value from an external source, such as a value from a previous realization of the same survey. Figure 2 illustrates the flow of handling missing data via single imputation.

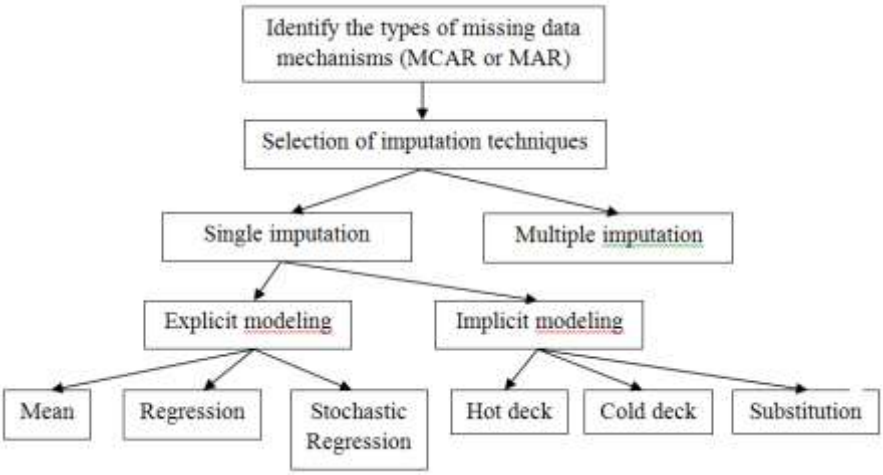

Fig. 2: Steps of managing missing values through single imputation.

3.1 Reviews on Adoption of Single Imputation Methods:

Implementation of single imputation technique had few advantages and drawbacks. The advantages of single imputation are [6]; 1) standard complete-data method can be applied to the filled-in data set; 2) the potentially substantial effort required to create sensible imputations need to be carried out only once, and; 3) this imputations can be incorporate the data collector's substance knowledge.

Omitting the missing data and substitution of mean values for missing data are commonly suggested when dealing with missing observations. This can direct to disorder of the inherent structure of the data, thus leading to tremendous errors in correlation matrix and degrade the performance of statistical modelling. Hawthorne and Elliot [12] conducted a comparative study on common techniques in imputing missing data in clinical and public health interventions. In this work, three methods were used (person mean, hot deck and regression) including the most widely reported methods i.e. mean substitution and list wise deletion. The results confirmed that list wise deletion and mean substitution performed poorly even in small scale of missing data. Hot deck method is recommended if the missing data are from a scale where more than the items are missing whereas person mean substitution is the method of choice due to its ease computational and more likely to be an option of statistical software packages.

Huisman [15] in his study on imputation of missing social network data uses some simple or single imputation procedures (imputation by reconstruction, unconditional mean, imputation using preferential attachment and hot deck imputation). The results agreed that ignoring missing data can have large negative effects on structural properties of the network and the simple imputations can correct the situation. It was reported that reconstruction method as the best method since it gave the smallest bias. However, simple procedure (single imputation) can still lead to statistical bias. As for solution, multiple imputation and improved single-imputation model are recommended.

On the other hand, Noor *et al.* [25] studied the assessment of various single imputation methods namely linear, quadratic, cubic interpolations, nearest neighbour, mean-before-after and mean after methods to the simulated missing air pollutant data (5%, 10%, 15%, 25% and 40%). Four performance indicators (prediction accuracy, coefficient of determination, mean absolute error and root mean squared error) were used to observe the goodness-of-fit for each method applied. The results show that mean-before-after gave the best results even for 40 percent missing data with the value of R^2 0.77 (which is identical to the result of 25% missing data). This is due to the way in which the simulated missing data were generated where the pattern of missing data for each percentage were identical. Hence, the effectiveness of the single imputation method of this study definitely can be debated.

Besides all the compensation of using single imputation technique, there is one most disadvantage of this technique. The drawback of single imputation is that it does not reflect extra uncertainty and display variation due to missing data [4]. Previous studies had shown that if the missing data was missing in short gap length, the adoption of single imputation technique caused no or tolerable values of bias depending on the numbers of missing observations. The increment in the percentage of missing values caused ineffective results as it increases the biasness. If the data mechanism is MAR, multiple imputation is superior compared to single imputation [35]. This is because of the multiple imputation generate multiple simulated values for each missing item, so that the uncertainty attached to missing data are reflected.

Junninen *et al.* [17] had used a few methods of single imputation that are linear, spline and nearest neighbour interpolation and studied the performances of these methods by filling in the simulated missing values (10 and 25 percent of simulated missing data). Since the missing mechanism was MAR, the results show that all methods dropped significantly for length of gap higher than 10 values (10 hours). Among all the methods, linear interpolation is the best method for short length of missing data.

This finding was supported by the study conducted by van der heijden*et al.* (2006). In this study, single imputation technique (i.e. of unconditional and conditional mean), multiple imputation technique (MI), complete case analysis (discard the missing values) and missing-indicator method were filled to the simulated data of diagnosis of pulmonary embolism. The results indicate that single imputation methods performed equally well compared to multiple imputation methods that are known to be superior. However, this is because of the low overall number of missing values.

Consequently, from the studies done by Junninen*et al.,* [17], Plaia and Bondi [26] and van der heijden *et al.* [35], it is proven that univariate methods (single imputation) performed excellently in the short gaps of missing data.

4.0 Multiple Imputation:

In history, Rubin had proposed multiple imputations in 1987. Rubin argued that an important limitation of single imputation methods is that „standard variance formulas' applied to filled-in data systematically underestimated the variance of estimates, thus he proposed multiple imputation [21]. In this method, the first step is to specify one encompassing multivariate model for the entire data set. There are four different types of multivariate complete data models that are [33]; i) normal model, which perform imputation under a multivariate normal distribution; ii) log linear model, which has been traditionally used by social scientists to describe associations among variables in cross-classified data; iii) general location model, which combines a loglinear model for the categorical variables with multivariate normal regression model for the continuous variables and; iv) two level linear regression model, which is commonly applied to multi-level data. The chosen imputation model should be compatible with the subsequent analysis or to be precise, the model should be flexible enough to preserve the relationships among variables that will be the focal point of later analysis.

Multiple imputations are similar to single imputation in that it imputes a set of likely values from a distribution for each missing variable. In general, adoption of multiple imputation technique involve three steps i.e. imputation, analysis and pooling. First steps, it imputes several values, through at less or more then 2 method, $m \geq 2$, for each missing datum. Then, each individual m is analysed using standard complete-data procedures that can be done using SPSS, LISREL, AMOS or any other statistical software. Lastly, pooling m complete datasets is the step of integrating the m analyses to produce overall estimates and standard errors. This step consists of calculating the mean over the m repeated analyses, its variance and its confidence interval or P value. Pooling data from a number of imputations allows multiple imputations to produce more accurate. Figure 3 shows an example of multiple imputation where $m = 3$.

Multiple imputations are an attractive choice as a solution to missing data problems because it represents a good balance between quality of results and ease of use. The performance of multiple imputations in a variety of missing data situations has been well studied and it has been shown to perform favourably. Multiple imputations introduced appropriate random error into the imputation process makes it possible to get approximately unbiased estimates of all parameters. There is no deterministic imputation method can do this in general settings. Further, multiple imputation has been shown to be robust to departures from normality assumptions and provides adequate results in the presence of low sample size or high rates of missing data.

Undoubtedly, certain requirements must be met for multiple imputations to meet these desirable properties [2]. First, the data must be missing at random (MAR) meaning that the missing data are dependent on the observed variables not the missing observations. Secondly, the selection of model used to generate the imputed values must be well suited with the subsequent analysis so that it can conserve the associations among variables that will be the focus of later analysis. Thirdly, the model used for the analysis must agree with the model used in the imputation. All these conditions have been described thoroughly by Rubin.

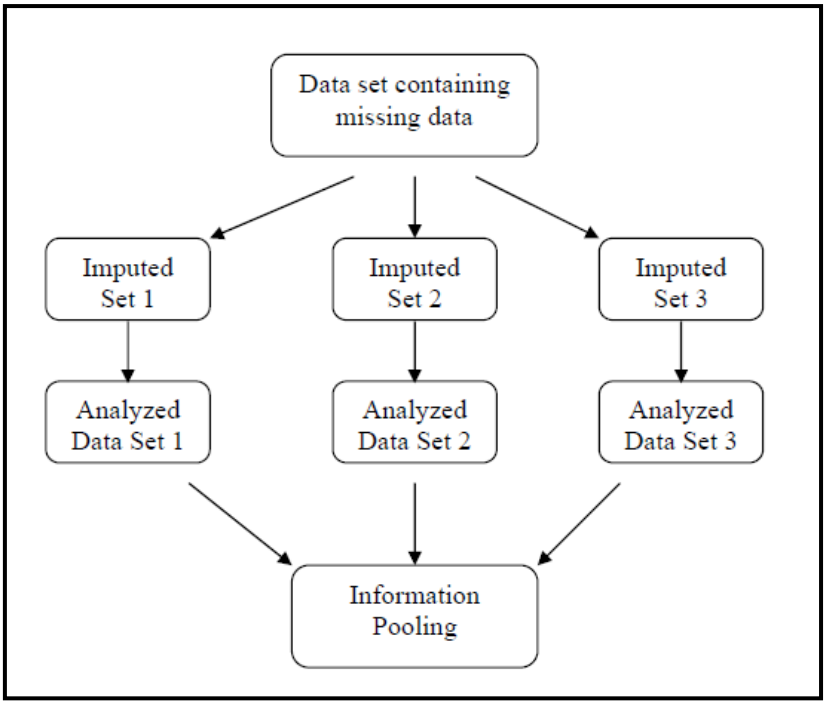

Fig. 3: The steps of implementing multiple imputations.

At the moment, broad applications of multiple imputations in estimating the missing data had been acknowledged especially in health discipline [3,20,9,30]. Adoption of multiple imputations on various missing data on serum cholesterol observations [3], otoneurologic data [20], body mass index (BMI) data and major lung resection outcomes from 1980-2006 [9] had agreed that multiple imputations as the method of choice compared to others.

Besides, huge applications of multiple imputations on health discipline, it was also adopted in environmental data set [18,1] industrial databases [16,19] survey data [11,34] and data mining that is the process of extracting pattern from large data set by combining methods from statistics and artificial intelligence with database management [16].

As described above, the application of multiple imputations were known to be superior compared to single imputation as long as the missing data mechanism is MAR. However, the battle of adopting multiple imputations lies in designing a suitable method to perform the imputations [38]. As most of imputations are executed using software, the awareness of the perfect prediction problem is essential in order to reduce systematically error.

5.0 Enhancement On Imputation Methods:

Since imputation had been the method of choice for replacing the missing values in various fields of research, a lot of studies had been conducted to explore the best method of all. It is hard to choose specifically one method that suits missing observations of all fields of research because of the good replacement method depend on the missing data pattern and mechanism for each data [21]. Nevertheless, the research on creating new method or enhancing the established methods needs to be done so that analysts had the reliable data for analysis. As explained above, single or multiple imputations had their own tales of advantages and drawbacks. In this section, the reviews on enhancement of these two techniques were revealed.

Although single imputation methods were known to cause moderate to serious bias in large amount of missing data, the simplicity and ease on adoption of single imputation methods had made this method a method of choice. Hence, Alireza *et al*., had done comprehensive review on the representative imputation techniques including single imputation methods (hot deck and Naïve-Bayes) and four multiple imputation models to develop a unified framework supporting a host of imputation methods. The framework consists of 3 modules that are i) mean pre-imputation; ii) application of confidence intervals and iii) boosting. Combining the proposed framework with hot deck and Naïve-Bayes methods did implementation of the framework. The results proved that by implementing the proposed framework with the low quality single imputation method had increased the imputation accuracy comparably to some other advanced imputation techniques. It is also demonstrated, both theoretically and experimentally, that the application of the proposed framework leads to a

linear computational complexity and hence, does not change the asymptotic complexity of associated imputation method.

A new single imputation method i.e. site-dependant effect method (SDEM) for environmental pollution data sets was proposed by Plaia and Bondi [26]. The proposed method was compared to other common single imputation technique (hour mean method, row mean method and last and next method) and a multivariate model-based multiple imputation method (MI). The results show that the performances of the proposed method were greater than other single imputation techniques and MI independently on the gap length and on the number if stations with missing data.

As for multiple imputations, most of the researches in health, environment and mining fields were fairly agreed that it is a superior method. Besides adoption of multiple imputations from Rubin, Verboven e*t al.* [36] proposed a new method for gene expression data that can be embedded in a multiple imputation. The single imputation methods once adopted to the gene expression data caused inaccurate results since some of them used limited set of genes to estimate the missing values whereas other utilize consumed long computational time. Other replacing methods are i) searching for the closest values for the incomplete data; and ii) iterative approach in which the missing data are estimated iteratively until satisfied convergence criterion. The performance of the newly proposed method namely SEQ impute was compared to KNN impute (K-nearest neighbour), SKNN impute (Sequential K-nearest neighbour), IA impute (iterative algorithm) and BPCA (Bayesian Principle Component Analysis). As the result, the proposed method is shown to be superior in terms of accuracy and computation speed.

For the time being, multiple imputations are the favourable method because of the accountability of covering the uncertainty surrounding the real data or in the simpler words, it reduces bias between observed and unobserved data. Furthermore, the availability of multiple imputations application for filling the missing values in statistical packages increases the favourability of this method. In the next section, several types of software were discussed to offer information on selecting the most appropriate package for the analysis.

6.0 Types Of Imputation Software:

Over the last three decades, there has been extensive progress in developing software to handle missing data. The needs of reliable complete datasets had been justified for the basis of this progress. Several types of software are able to estimate missing values in datasets. In this sub section, five software will be briefly described that are; i) SPSS: MVA; ii) SOLAS: MDA; iii) Stata: ICE; iv) SAS: MI; and v) NORM and related programs.

Missing Values Analysis (MVA) is an optional module for SPSS. Varieties of methods are offered including complete case analysis (list wise deletion method), missing-indicator method (pair wise method) and multiple imputation method (EM algorithm and multiple linear regressions). According to Hippel [13], none of these four methods is completely satisfactory when the missing data mechanism is MAR. List wise and pair wise are well known to be biased. For the regression imputation, the regression parameters are noticed to be biased because the derivations are using pair wise deletion. The last method, expectation maximization (EM) produces asymptotically unbiased estimates but the application is limited to point estimates of means, variances and covariance (without standard errors). MVA can also fill in the values using the EM algorithm, but values are imputed without residual variation, so analyses that use the imputed values can be biased.

Secondly, Statistical Solution (SOLAS) for Missing Data Analysis (MDA) was referred as an easy and validated application of imputing missing values. SOLAS-MDA offers 6 different imputation techniques principled approaches to analysed data with missing values. Types of available imputation methods in SOLAS are single imputations methods (hot deck, predicted mean using regression, last value carried forward and group mean) and multiple imputation methods (non parametric approach based on propensity scores and approximate Bayesian bootstrap). Allison [1] reported that multiple imputations using a propensity score classifier with the approximate Bayesian bootstrap (SOLAS-MDA) produces badly biased estimates of the regression coefficients when the missing data mechanism was MAR or MCAR. In this study, the algorithm was compared with a quite different method for generating multiple imputations, described by Schafer [32] and embodied in NORM, his free software for Windows. The results indicate that a regression-based method employing the data augmentation algorithm (NORM) produces estimates with little or no bias compared to SOLAS-MDA.

Royston [28] created the application of multiple imputations of missing values in Data analysis and statistical software (Stata). Then in 2005, Royston introduced an update to this application that is imputation with chained equations. This function provides single and multiple imputations using a model based approach based on chained regression equations. The user had a choice of selections for a regression model based on the level of measurement of the variables. A supplied program namely MICOMBINE allows rolling up the estimates from multiple imputed data sets for many types of regression-based methods.

Statistical Analysis System (SAS) Release 8.2 (Chapter 9) embedded the experimental procedure of imputing missing values via single imputation and multiple imputations methods. For data sets with monotone missing patterns, either a parametric regression method (Rubin 1987) that assumes multivariate normality or a

nonparametric method that uses propensity scores is appropriate. For data sets with arbitrary missing patterns, a Markov Chain Monte Carlo (MCMC) method [32] that assumes multivariate normality is used to impute all missing values or just enough missing values to make the imputed data sets have monotone missing patterns. The users have the relieved of selectivity on a number of options regarding the details of the imputation process and the program can produce a lot of diagnostics information to assess the satisfactoriness of the imputations for the data. This experimental package also includes the option to produce an EM for single-imputed dataset. On the other hand, this implementation is differing from the SPSS-MVA (for EM algorithm) since it does properly include error estimation in the imputed values.

The last but not least software described here is NORM and related programs. Schaefer and his group from Department of Statistics, The Pennsylvania State University, created these set of programs. The set of programs consist of NORM (Multiple imputations of multivariate continuous data under a normal model), CAT (Multiple imputations of multivariate categorical data under log linear models), MIX (Multiple imputation of mixed continuous and categorical data under the general location model) and PAN (Multiple imputation of panel data or clustered data under a multivariate linear mixed-effects model). They are written for use in S-Plus but there is a version of NORM that can be run in Windows. John Graham has developed a set of utilities to facilitate the use of NORM in SAS and SPSS. Most of the imputation models used in NORM and related programs are similar to those in PROCMI function in SAS. Darmawan [31] in his article on reviewing NORM software suggested that this software is user-friendly and capable on dealing the data sets with high percentage of incomplete cases. Table 1 shows the summarization of methods in various software packages.

Table 1: Methods in common statistical packages for missing values.

SOFTWARE	METHODS			
	SINGLE IMPUTATION	MULTIPLE IMPUTATION	COMPLETE CASE ANALYSIS	INDICATOR METHOD
SPSS – MVA (Missing Value Analysis)	-	• EM algorithm • Multiple Linear Regression	• List wise	• Pairwise
SOLAS – MDA (Missing Data Analysis)	• Hot Deck • Predicted mean • Last value carried Forward (LVCF) • Group mean	• Predictive model based • Propensity score based	-	-
Stata: ICE (Imputation with Chained Equations)	• Model based approach using chained equation	• Model based approach using chained equation	-	-
SAS: MI (Multiple Imputation)	-	• Parametric regression method by Rubin (1987) • Nonparametric method • Markov chain Monte Carlo (MCMC) by Schafer (1997)	-	-
S-Plus: NORM SAS: NORM SPSS: NORM Windows: NORM	-	• Model based multiple imputation	-	-

Conclusions:

Imputation methods had been the method of choice for replacing missing data in broad areas of research. Imputation methods namely single and multiple imputations had certain criteria to be fulfilled before adoption. In order to understand the philosophy of imputation, the missing mechanisms were firstly described. Single imputation, usually simple statistical method, works very well in missing-completely-at-random (MCAR) data set. Adoption of single imputation method to the missing-at-random (MAR) data will cause serious bias since the reflection of uncertainty is not covered. Multiple imputations were recognized as the best method for MAR data set. Although the superiority of multiple imputations was known, the adoption of these imputations indulged thorough understanding on the algorithms especially in designing a suitable method to conduct the imputations. The reviews on assessment of available imputation software were also presented to compare the practicality of the software.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support by Ministry of Higher Education, Malaysia for Fundamental Research Grant Scheme (FRGS 9003-00272).

REFERENCES

[1] Allen, R.J. and A.T. DeGaetano, 2001. Estimating Missing Daily Temperature Extremes using an Optimized Regression Approach. *International Journal of Climatology,* 21: 1305-1319.

[2] Allison, P.D., 2000. Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research,* 28(3): 301-309.

[3] Barzi, F. and M. Woodward, 2004. Imputations of Missing Values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology,* 160: 34-45.

[4] Bono, C., L.D. Ried, C. Kimberlin, B. Vogel, 2007. Missing data on the Center for Epidemiologic Studies Depression Scale: A comparison of 4 imputation techniques. *Research in Social and Administrative Pharmacy,* 3: 1-27.

[5] Darmawan, I.G.N., 2002. NORM software review: handling missing values with multiple imputation methods. *Evaluation Journal of Australasia,* 21(1): 51-57.

[6] Donders, A.R.T., G.J.M.G. van der Heijden, T. Stijnen, K.G.M. Moons, 2006. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology,* 59: 1087-1091.

[7] Farhangfar, A., L. Kurgan, J. Dy, 2008. Impact of imputation of missing values on classification error for discrete data.*Pattern Recognition,* 41: 3692-3705.

[8] Farhangfar, A. and L.A. Kurgan, 2007. A Novel Framework for Imputation of Missing Values in Databases*. IEEE Transactions on Systems, man and Cybernatics- Part A: Systems and Humans,* 37(5): 692-709.

[9] Ferguson, M.K., J. Siddique, T. Karrison, 2008. Modeling major lung resection outcomes using classification tress and multiple imputation techniques. *European Journal of Cardio-thoracic Surgery,* 34: 1085-1089.

[10] Fu, T.C., 2010. A review on time series data mining.*Engineering Applications of Artificial Intelligence,* 24: 164-181.

[11] Ginkel, J.R.V., L.A. Van der Ark, K. Sijtsma, J.K. Vermut, 2007. Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis,* 51: 4013-4027.

[12] Hawthorne, G. and P. Elliott, 2005. Imputing cross-sectional missing data: Comparison of common techniques. *Australian and New Zealand Journal of Psychiatry,* 39: 583-590.

[13] Hippel, V. and T. Paul, 2004. Biases in SPSS 12.0 Missing Values Analysis. *The American Statistician,* 58: 160-164.

[14] Hopke, P.K., C.H. Liu and D.B. Rubin, 2001. Multiple Imputations for Multivariate Data with Missing and Below-Threshold Measurements: Time Series Concentrations of Pollutants in the Arctic, *Biometrics,* 57: 22-33.

[15] Huisman, M., 2007. Imputation of missing network data: Some simple procedures. In: Proceedings of Sunbelt XXVII International Sunbelt Social Network Conference, pp: 1-20.

[16] Jagannathan, G. and R.N. Wright, 2008. Privacy-preserving imputation of missing data. *Data & Knowledge Engineering,* 65: 40-56.

[17] Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, 2004. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment,* 38: 2895-2907.

[18] Kotsiantis, S., A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias, 2006. Filling Missing Temperature Values Weather Data Banks. In: Proceedings of 2nd IEE International Conference on Intelligent Environments, 327-334.

[19] Lakshminarayan, K., S.A. Harp, T. Samad, 1999. Imputation of Missing Data in Industrial Databases.*Applied Intelligence,* 11: 259-275.

[20] Laurikkala, J., E. Kentala, M. Juhola, I. Pyykko, S. Lammi, 2000. Usefulness of imputation for the analysis of incomplete otoneurologic data.*International Journal of Medical Informatics58-59,* pp: 235-242.

[21] Little, R.J.A. and D.B. Rubin, 1987. *Statistical Analysis with Missing Data*. Wiley, New York.

[22] Marco, D.Z., U. Guarnera, O. Luzi, 2007. Imputation through finite Gaussian mixture models.*Computational Statistics & Data Analysis,* 51: 5305-5316.

[23] Mondelo, D., 2006. *Imputation Strategies for Missing Data in Environment Time Serial for an Unlucky Situation*.Springer Berlin Heidelberg.

[24] Moore, L., J.A. Hanley, A.F. Turgeon, A. Lavoie, M. Emond, 2009. A Multiple Imputation Model for Imputing Missing Physiologic Data in the National Trauma Data Bank.*American College of Surgeons,* 209(5): 572-579.

[25] Noor, N.M., A.S. Yahaya, N.A. Ramli, M.M.A. Abdullah, 2008. Estimation of missing values in air pollution data using single imputation techniques*. Science Asia,* 34: 341-345.

[26] Plaia, A. and A.L. Bondi, 2006. Single Imputation method of missing values in environmental pollution data sets.*Atmospheric Environment,* 40: 7316-7330.

[27] Qin, Y., S. Zhang, X. Zhu, J. Zhang, C. Zhang, 2009. POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert System with Application,* 36(2): 2794-2804.
[28] Royston and Patrick, 2004. Multiple imputation of missing values.*The Stata Journa,* l4: 227-241.
[29] Royston and Patrick, 2005. Multiple imputation of missing values: Update. *The Stata Journal,* 5: 1-14.
[30] Sartori, N., A. Salvan, K. Thomaseth, 2005. Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose.*Computational Statistics & Data Analysis,* 49: 937-953.
[31] SAS OnlineDOC^{TM} (Version 8): Chapter 9: MI Procedure.
[32] Schafer, J.L., 1997. *Analysis of incomplete multivariate data.*Monographs on Staistics and Applied Probability, No. 72. Chapman and Hall, London.
[33] Schafer, J.L. and M.K. Olsen, 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research,* 33(4): 545-571.
[34] Schenker, N. and J.M.G. Taylor, 1995. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis,* 22: 425-446.
[35] Van der heijden, G.J.M.G., A.R.T. Donders, T. Stijnen, K.G.M. Moons, 2006. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A Clinical example. *Journal of Clinical Epidemiology,* 59: 1102-1109.
[36] Verboven, S., K.V. Branden, P. Goos, 2007. Sequential imputation for missing values.*Computational Biology and Chemistry,* 31: 320-327.
[37] Wasito, I. and B. Mirkin, 2005. Nearest neighbor approach in the least-squares dat imputation algorithms. *Information Science,* 169: 1-25.
[38] White, I.R., R. Daniel and P. Royston, 2010. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis,* 54: 2267-2275.