Missing data treatment

Overview of possible solutions

RA-MOW-2011-002

Wilmots B., Shen Y., Hermans E., Ruan D.

Onderzoekslijn Risicobepaling

DIEPENBEEK, 2011.
STEUNPUNT MOBILITEIT & OPENBARE WERKEN
SPOOR VERKEERSVEILIGHEID

Documentbeschrijving

Rapportnummer:	RA-MOW-2011-002
Titel:	Missing data treatment
Ondertitel:	Overview of possible solutions
Auteur(s):	Wilmots B., Shen Y., Hermans E., Ruan D.
Promotor:	Prof. dr. Tom Brijs
Onderzoekslijn:	Risicobepaling
Partner:	Universiteit Hasselt
Aantal pagina’s:	33
Projectnummer Steunpunt:	6.2
Projectinhoud:	In dit project staat de ontwikkeling van een set van indicatoren voor de verkeersveiligheid centraal. Dit rapport gaat dieper in op de aanpak ingeval van ontbrekende waarden.

Uitgave: Steunpunt Mobiliteit & Openbare Werken – Spoor Verkeersveiligheid, maart 2011.

Steunpunt Mobiliteit & Openbare Werken
Spoor Verkeersveiligheid
Wetenschapspark 5
B 3590 Diepenbeek

T 011 26 91 12
F 011 26 91 99
E info@steunpuntmowverkeersveiligheid.be
I www.steunpuntmowverkeersveiligheid.be

Samenvatting

Titel: Omgaan met ontbrekende data

Subtitel: Overzicht van mogelijke oplossingen

Datasets met werkelijke informatie gaan bijna altijd gepaard met ontbrekende gegevens omwille van verschillende onzekerheden. Dit beperkt onderzoekers in grote mate om klassieke analyses uit te voeren die volledige datamatrices vereisen in de meeste gevallen. Om dit veelvoorkomend probleem in data-analyse op te lossen,werden een aantal alternatieve methodes ontwikkeld gedurende de laatste vijf decennia.

Een eenvoudige en veelgebruikte strategie om ontbrekende informatie te behandelen, is het weglaten van cases die ontbrekende waarden bevatten en vervolgens de analyse uit te voeren op de overblijvende data. Ondanks het feit dat dit eenvoudig uit te voeren is en de standaardoptie is bij de grote statistische pakketten, heeft deze benadering toch ernstige beperkingen in termen van het elimineren van bruikbare informatie in de data en het resulteren in vertekening wanneer de gegevens niet volledig willekeurig ontbreken (i.e., not missing completely at random).

Later verschoof de interesse naar het uitvoeren van data-imputatie, het proces waarbij ontbrekende waarden in een dataset geschat worden door berekende waarden en waarbij dus een volledige dataset gecreëerd wordt. Enkele voorbeelden van deze werkwijze, gekend als traditionele enkelvoudige imputatie, zijn: onvoorwaardelijk gemiddelde imputatie, regressieimputatie, de indicatormethode, enz. Echter, zelfs wanneer de ontbrekende waarden op die manier worden geïmputeerd of ingevuld, blijft het probleem bestaan dat de onzekerheid die verbonden is aan ontbrekende data niet in rekening wordt gebracht. Daarom is vanaf de jaren 70 er sterke vooruitgang geboekt in het ontwikkelen van statistische procedures voor ontbrekende data en de twee belangrijkste benaderingen, i.e., maximum likelihood schatting en meervoudige imputatie, zijn beschikbaar geworden als bruikbare opties in de belangrijkste softwarepakketten.

Meer recent, met de ontwikkeling van computerwetenschap en technologie, zijn enkele artificiële intelligentie technieken ontstaan met betrekking tot het omgaan met ontbrekende informatie, zoals beslissingsbomen, neurale netwerken, fuzzy logic systemen, rough sets enzovoort, dewelke het onderzoek naar ontbrekende data naar een nieuwe fase brengen.

In dit rapport worden de belangrijkste ideeën van al deze benaderingen besproken evenals de sterktes en beperkingen van elke benadering. Verder staan we stil bij de beschikbare softwareprogramma’s en wordt er informatie geboden omtrent het selecteren van een bepaalde benadering in de praktijk.

English summary

Title: Missing data treatment

Subtitle: Overview of possible solutions

Abstract:

Real world data sets are almost always accompanied by missing data due to various uncertainties, which to a great extent restrict researchers from performing classical analyses as complete data matrices are required in most cases. To solve this pervasive problem in data analysis, a number of alternative methods have been developed during the last five decades.

Specifically, a simple and common strategy for handling missingness is to delete cases containing any missing values, and the analysis is then carried out on the data that remain. Although simple to implement and being the default for the major statistical packages, this approach has serious drawbacks in terms of elimination of useful information in the data and resulting in serious biases if data are not missing completely at random (MCAR).

Later, interest has centered on performing data imputation, the process by which missing values in a data set are estimated by appropriately computed values, thus constructing a complete data set. Unconditional mean imputation, regression imputation, the indicator method and so on are all related to this strategy, known as traditional single imputation. However, even if the missing values could be imputed in such a way, they still have a problem in accounting for missing data uncertainty. Therefore, from the late 70’s on, substantial progress has been made in developing statistical procedures for missing data, and two most important approaches, i.e., maximum likelihood estimation and multiple imputation, have become available, and are being included as useful options in the mainstream software programs.

More recently, with the development of computer science and technology, some artificial intelligence and machine learning techniques have arisen in the area of missing data treatment, such as decision trees, neural networks, fuzzy logic systems, rough sets, and so on, which push the missing data research forward to a new stage.

In this report, we outline the key ideas of all these approaches, address their main strengths and limitations, discuss the software programs currently available, and provide guidance on how to select such approaches in practice.

Inhoudsopgave

1. INTRODUCTION TO MISSING DATA

The performance of almost all kinds of analyses is largely based on the quality of the input data. In the real world, however, there is not any data collection system granting perfect data sets, and a certain risk in the form of missing values is always present. In longitudinal studies, subjects may drop out early or be unavailable during one or more data collection periods. When data are collected by questionnaire or interview, respondents may not infrequently leave particular items blank due to sensitivity, fatigue, lack of knowledge, or other factors. These types of missingness, although unintended and uncontrolled by the researchers, force them to decide whether to leave cases with missing data out of the analysis, or to replace the blank information by imputed values, as complete data matrices are required to perform classical analyses.

This report reviews a variety of approaches to handle missing data and introduces plenty of software currently available. We aim to familiarize researchers with strengths and limitations of all these possible solutions, provide them with some recent developments in this rapidly changing field, and give guidance on how to select such approaches in practice.

1.1 The missingness mechanisms

Any discussion of missing data must begin with the question of why data are missing in the first place, as the reasons for missing data play an important role in how those data will be treated. In this respect, Rubin (1976) defined a clear taxonomy of missingness that has become the standard for any discussion on this topic. It involves Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR).

1.1.1 Missing completely at random

There are several reasons why data may be missing. The term *'Missing Completely at Random*' refers to data where the missingness mechanism does not depend on the variable of interest, or any other observed variable in the data set, but occurs in a rather random fashion due to for example malfunctioning equipment, adverse weather, lack of personnel, etc. Typical examples of MCAR are when a questionnaire of a respondent is accidentally lost or a participant becoming ill thus not being able to appear for testing. In such cases, the reason for missingness is completely random, i.e., the probability that an observation is missing is not related to the value of this observation or to the value of any other variables. Pickles (2005) phrased the condition somewhat differently by saying that for MCAR the probability of missingness is a constant. Any observation on a variable is as likely to be missing as any other.

When missing data are MCAR, evidently the set of observations with no missing data is also a random sample from the source population. Hence, most simple techniques for handling missing data, including *listwise* and *pairwise deletion*, give unbiased results (see Chapter 2).

Little (1998) has developed a statistical test of the MCAR assumption, which is a *chi-square* test provided in for example the SPSS Missing Values Analysis (MVA) option [SPSS Inc., 2007]. If the *p* value for Little's MCAR test is not significant, then the data may be assumed to be MCAR.

1.1.2 Missing at random

Often, missing data are not completely at random [Rubin, 1976]. For example, in the self reported speeding behavior, if young drivers are more likely to omit reporting speed limit violation records than elders, we would not have data that are MCAR because missingness would be correlated with age. In such a case, if within each stage of age classification, the probability of missing records is unrelated to the value of records themselves, then a more general assumption, *Missing at Random*, can be defined (a more formal version of which is sometimes called the ignorability assumption). Theoretically, MAR means that the probability of missing data on a variable is not a function of its own value after controlling for other variables in the design. Under this assumption, missing data can indeed be considered 'random' conditional on the other observed data values in the data set that determined their missingness. However, when missing data are MAR, *listwise* or *pairwise deletion* is no longer based on a random sample from the source population and selection bias likely occurs. Taking the above self reported speeding behavior as an example, young drivers might be less inclined to report their speed limit violation records, and they might also have higher records in general. Thus, when we have a high rate of missing data among these individuals, the actual mean records might be higher than it would be without missing data.

Generally, when missing data are MAR, all simple techniques for handling missing data, e.g., *listwise* and *pairwise deletion*, *mean imputation*, and *the indicator method*, give biased results (see Chapters 2 and 3). However, some more sophisticated techniques like *maximum likelihood estimation* and *multiple imputation* give unbiased results (see Chapter 4).

Unfortunately, we generally cannot be sure whether data really are missing at random. The fundamental difficulty is that some potential 'lurking variables' are unobserved and so we can never rule them out. We generally must make assumptions, or check with reference to other studies (for example, surveys in which extensive follow-ups are done in order to ascertain the speeding behavior of non-respondents). In practice, we typically try to include as many predictors as possible in a model so that the MAR assumption is reasonable. For example, it may be a strong assumption that non-response to the speeding behavior question depends only on age, gender, and education, but this is a lot more plausible than assuming that the probability of non-response is constant, or that it depends only on one of these predictors.

1.1.3 Missing not at random

Data are classified as *Missing Not at Random* if either of the above two classifications are not met. Thus if the data are not at least MAR, i.e., if the probability that an observation is missing depends on information that is not observed, or the value of the observation itself, then they are MNAR, which is also known as non-ignorable missingness. Still taking the self reported speeding behavior as an example, the missing data are considered to be MNAR if people with high speed limit violation records are in fact more reluctant to report their behavior than people with lower records. If missing data are MNAR, valuable information is lost from the data and, there is no universal method of handling the missing data properly. One possible way to obtain an unbiased estimate of parameters is to model missingness. In other words, we would need to create a model that accounts for the missing data. Such a model could then be incorporated into a more complex model for estimating missing values. Unfortunately we rarely know what the missingness model is, so it is difficult to know how to proceed. In addition, incorporating a model of missingness is often a very difficult task and may be specialized for each application. See Dunning and Freedman (2008) for a useful example of a model dealing with missingness.

1.2 Techniques of missing data treatment

During the last five decades, various methods have been developed for handling missingness. The literature on the analysis of missing data is extensive now and still in rapid development. Generally, three different strategies for dealing with missing data can be identified, which are *deletion*, *imputation*, and *using as it is* [Kabak and Ruan, 2010; Rodríguez et al., 2010]. In this report, we mainly focus on the first two strategies. The comprehensive techniques within each strategy are illustrated in Figure 1-1.

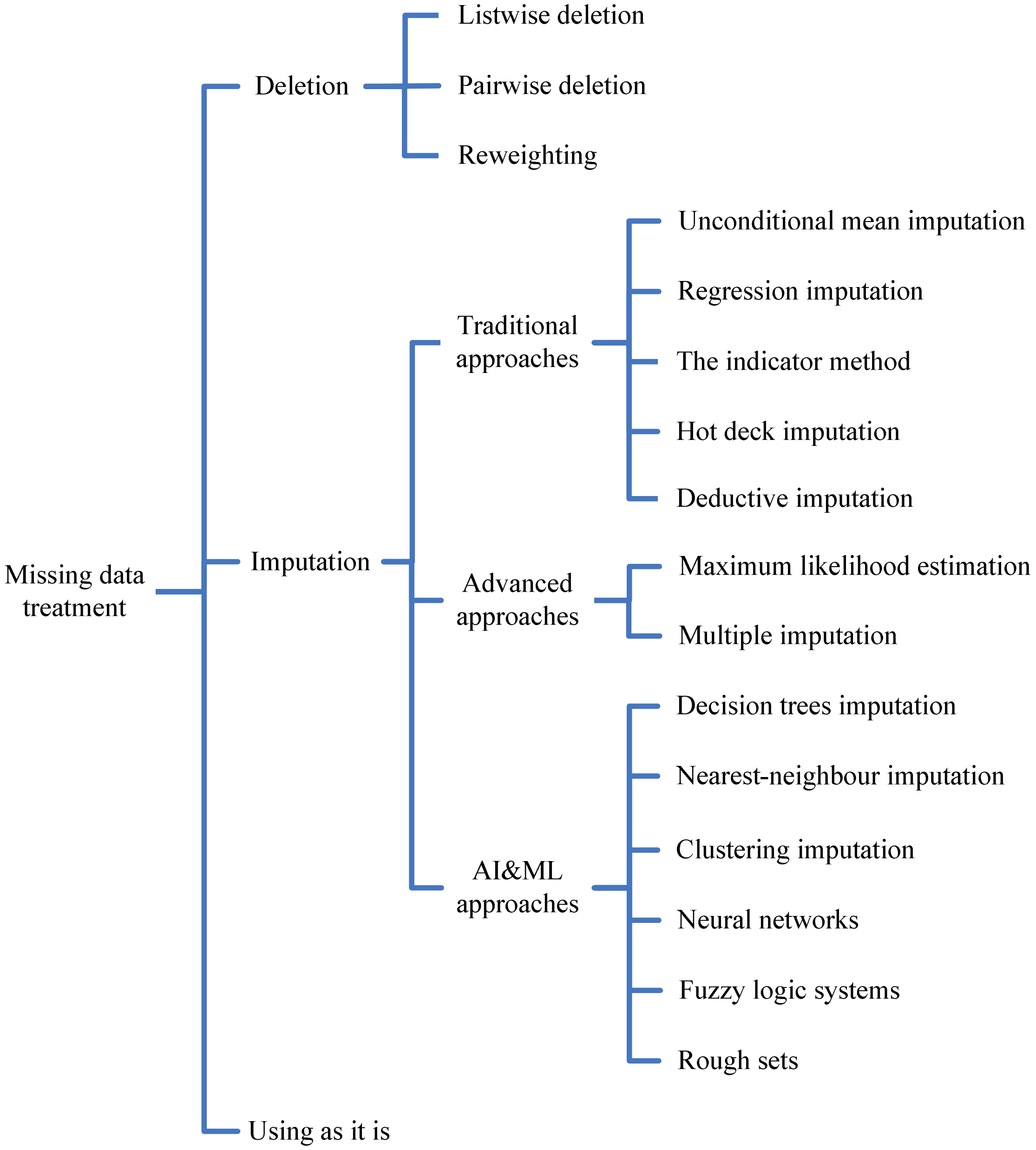

Figure 1-1 The overview of the techniques for missing data treatment

Specifically, the first strategy simply deletes cases containing any missing observations *listwise* or *pairwise*. The analysis is then carried out on the data that remain. These ad hoc methods, although simple to implement, have serious drawbacks in terms of elimination of useful information in the data and resulting in serious biases if the subjects who provide complete data are unrepresentative of the entire sample (i.e., the missing data are not MCAR) [Little and Rubin, 1987; Schafer and Graham 2002; Howell, 2008; Oltman and Yahia, 2008].

The second strategy to deal with missing data is to perform data imputation, defined as the process by which missing values in a data set are estimated by appropriately computed values, thus constructing a complete data set [Rubin, 1987]. Currently, most of the models dealing with missing data use this strategy. See also Chen and Shao (2000); Schafer and Graham (2002); Farhangfar et al. (2004); Molnar et al. (2008); Howell (2008); Jiang and Gruenwald (2008); Pospiech-Kurkowska (2008); Wang and Wang (2009); Silva-Ramírez et al. (2010), and their references. Some traditional techniques developed in the imputation approach are *mean imputation*, *regression* *imputation*, and so on. However, simple mean substitution will seriously dampen relationships among variables, and substituting regression predictions will artificially inflate correlations (see Chapter 3). Even if the missing values could be imputed in such a way that the distributions of variables and relationships among them were perfectly preserved, the imputed data set would still fail to provide accurate measures of variability for the following reason: subsequent analyses would fail to account for missing data uncertainty. Regardless of the imputation method, imputed values are only estimates of the unknown true values. Any analysis that ignores the uncertainty of missing data prediction will lead to standard errors that are too small, *p*-values that are artificially low, and rates of Type I error that are higher than nominal levels [Schafer and Olsen, 1998]. To solve the problems given by the traditional imputation approaches, substantial progress has been made in developing statistical procedures for missing data. In the last three decades, interest has centered on two advanced approaches, which are *maximum likelihood estimation* using for example the *EM (Expectation-Maximisation) algorithm* and *multiple imputation*. More recently, with the development of computer science and technology, some artificial intelligence and machine learning (AI&ML) techniques have arisen in the area of missing data, such as *decision trees*, *neural networks*, *fuzzy logic systems*, *rough sets,* and so on, which push the missing data research forward to a new stage.

Although the imputation of missing values requires additional efforts, it has an added value in terms of making use of all available information. When missing values are imputed then classical models can easily be applied to the completed data. However, when the missing data are not at random it becomes very hard to impute reliable values. The last strategy for missing data is to use the data as it is without any treatment. In this approach, original data sets with missing values are not preprocessed, i.e., data sets are not preliminarily converted into complete data sets. Thus, the models to apply on the data should be capable of using incomplete data. However, it is very difficult to realize since complete data matrices are in most cases the prerequisite of performing classical analyses. This strategy, therefore, is still rare in literature and is used only in limited application areas until now [Li, 2006; Grzymala-Busse, 2008; Kabak and Ruan, 2010]. Consequently, in this report, we mainly discuss the approaches for dealing with the missing data problem within the first two strategies. It should be noted that most of these approaches apply only when the data are at least *missing at random*.

1.3 Structure of the report

In the remaining chapters, a variety of missing data approaches within the strategies of *deletion* and *imputation* are presented, their main strengths and limitations discussed and available software listed. The structure of the whole report is illustrated in Figure 1-2.

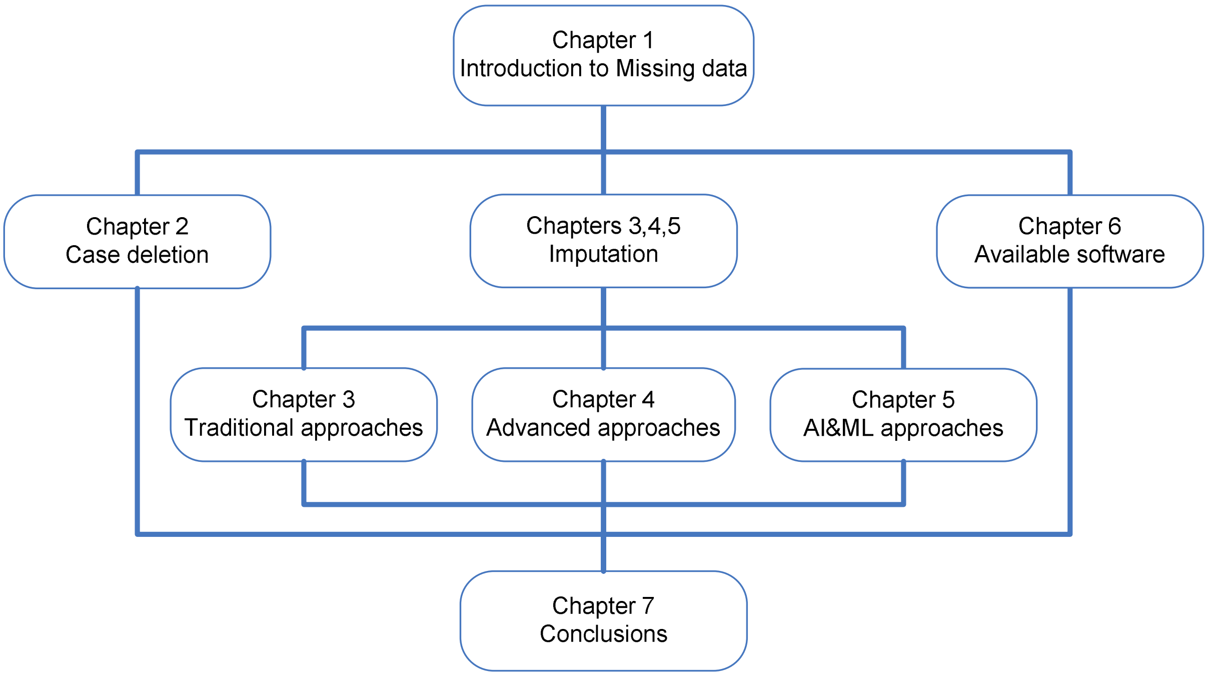

Figure 1-2 The structure of this report

2. CASE DELETION

Until recently, many missing data approaches simplified the problem by throwing away data. We discuss in this Chapter how these approaches work, and their main advantages and disadvantages.

2.1 Listwise deletion

By far the most common approach for missing data is to exclude them. Using this approach we simply drop from the analysis all cases that include any missing observation. The analysis is then carried out on the data that remain. Thus if 5 subjects in group one don't show up to be tested, that group is 5 observations short. Or if 5 individuals have missing scores on one or more variables, we simply omit those individuals from the analysis. This approach is usually called *listwise deletion*, which is also known as *complete-case analysis* in the regression context [Schafer and Graham, 2002].

In many situations, *listwise deletion* is an appropriate option. First of all, it is easy to implement, and is usually the default analysis for most statistical software, such as SPSS. If a missing-data problem can be resolved by discarding only a small part of the sample, then this approach can be quite effective. In particular, under the assumption that data are MCAR, it leads to unbiased parameter estimates. However, *listwise deletion* also has serious drawbacks which have been well documented (see, e.g., Little and Rubin (1987); Schafer and Graham (2002); Howell (2008); Oltman and Yahia (2008)). For multivariate analyses involving a large number of items, this approach can be very inefficient, discarding an unacceptably high proportion of subjects; even if the per-item rates of missingness are low, few subjects may have complete data for all items. Consequently, *listwise deletion* often results in a substantial decrease in the sample size available for the analysis. Moreover, when the data are not MCAR, this approach may lead to biased estimates, because the complete cases can be unrepresentative of the full population. For example, when the drivers with high speed limit violation records are less likely to report their speeding behavior, the resulting mean is biased in favor of lower records. In addition, standard errors will, in general be larger in a reduced sample given that less information is used.

2.2 Pairwise deletion

Another case deletion approach is called *pairwise deletion*, also commonly known as *available-case analysis* [Schafer and Graham, 2002]. The basic idea of this approach is that data are kept or deleted on the basis of pairs of scores. In computing the overall covariance or correlation matrix, a pair of scores contributes to the correlation if both scores are present, but does not contribute if one or both of them are missing. For example, if one participant reports his age and speeding behavior, but not his education level, he is included in the correlation of age and speeding behavior, but not in the correlations involving education level. Thus, all available observations would be used in estimating means and standard deviations of the variables.

This method has the advantage that it makes use of all available data and thus estimates parameters on the maximum sample size. But this is its only advantage. The major disadvantage is that each correlation, mean and standard deviation is estimated on a somewhat different subset of the data and they will not necessarily be consistent with each other. Moreover, it is also possible that the covariance or correlation matrices resulting from this approach and needed for the analysis are not positive definite. This implies that it is impossible to calculate a normal inverse of either matrix, which is likely to bring the whole analysis to a stop [Schafer and Graham, 2002; Howell, 2008].

In general, *pairwise deletion* may be necessary when overall sample size is small or the number of cases with missing data is large. It is believed that the underlying principle of this approach—to make use of all the available data—is eminently sensible, but deleting cases is a poor way to realize it, and misinterpretation may result when data are not MCAR.

2.3 Reweighting

As discussed previously, *listwise deletion* can yield biased estimates because the sample of observations that have no missing data might not be representative of the full sample, i.e., data may not be missing completely at random. In some non-MCAR situations, it is possible to reduce biases from case deletion by the judicious application of weights [Schafer and Graham, 2002]. Specifically, after incomplete cases are removed, the remaining complete cases are weighted so that their distribution resembles that of the full sample or population more closely with respect to auxiliary variables. Weights are derived from the probabilities of response, which must be estimated from the data (e.g., by a logistic or probit regression). Reweighting can eliminate bias due to differential response related to the variables used to model the response probabilities, but it cannot correct for biases related to variables that are unused or unmeasured. For a review of reweighting in the context of sample surveys, we refer to Little and Rubin (1987).

Reweighting is nonparametric, requiring no model for the distribution of the data values in the population. It does, however, require some model for the probabilities of response. Suppose, for instance, that only one variable has missing data. We could build a model to predict the non-response in that variable using all the other variables. The inverse of predicted probabilities of response from this model could then be used as survey weights to make the complete-case sample representative of the full sample. This method becomes more complicated when there is more than one variable with missing data. Moreover, as with any weighting scheme, there is the potential that standard errors will become erratic if predicted probabilities are close to 0 or 1. However, a resurgence of interest in reweighting arose during the last two decades, with new methods for parametric and semi-parametric regression appearing in biostatistics (see also Robins et al. (1994), (1995), and (1998)).

2.4 Summary

This Chapter introduces three main missing data approaches with respect to *case deletion*. They are listwise deletion, pairwise deletion and reweighting. The properties of these three methods are summarized in Table 2-1.

Table 2-1 Summary information on the three case deletion methods

	Method description	Main advantages	Main disadvantages	Assumptions	When to use
Listwise deletion	Drop all cases that include any missing observation	Easy to implement	May result in a substantial decrease in the sample size Elimination of useful information in the data Lead to biased estimates when data are not MCAR	Data are MCAR	Only a small number of data are missing
Pairwise deletion	Data are kept or deleted on the basis of	Use all available	Estimates are obtained from a different	Data are MCAR	Overall sample size is small or

	pairs of scores	data	subset of the data and may not be consistent with each other		the number of cases with missing data is large
Reweighting	Use weights to make the complete-case sample representative of the full sample	Reduce biases of estimates	Complex when there is more than one variable with missing data	Requires some model for the probabilities of response to get weights	Use listwise deletion but data are not MCAR

3. TRADITIONAL IMPUTATION APPROACHES

As a rule of thumb if a variable has more than 5% missing values, cases cannot be deleted [Little and Rubin, 1987]; many researchers are much more stringent than this. In such a case, rather than removing variables or observations with missing data, another possible strategy is to perform data imputation, defined as the process by which missing values in a data set are estimated by appropriately computed values, thereby producing a complete data set [Rubin, 1987].

Imputation has several desirable features. It is potentially more efficient than case deletion, because it uses 'expensive to collect' data that would otherwise be discarded, which helps to prevent loss of power resulting from a diminished sample size. Moreover, if the observed data contain useful information for predicting the missing values, an imputation procedure can make use of this information and maintain a high level of precision. Imputation also produces an apparently complete data set that may be analyzed by standard methods and software. To a data user, the practical value of being able to apply a favorite technique or software product can be immense. Finally, when data are to be analyzed by multiple persons or entities, imputing once, prior to all analyses, helps to ensure that the same set of units is being considered by each entity, facilitating the comparison of results. On the negative side, imputation can be difficult to implement well, particularly in multivariate settings. Some ad hoc imputation methods can distort data distributions and relationships. In the words of Dempster and Rubin (1983): "The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to real and imputed data have substantial bias."

A variety of imputation approaches exist ranging from extremely simple to rather complex. In this Chapter, we classify and review some traditional *single imputation* methods. See also Little and Rubin (1987); Schafer and Graham (2002); Howell (2008); Jiang and Gruenwald (2008), and their references.

3.1 Unconditional mean imputation

One of the easiest ways to impute is to replace each missing value with the mean of the observed values for that variable, which is known as *unconditional mean imputation*. In addition, the median (the value that divides in two equal parts the distribution of the random variable) and the mode (the value with the highest frequency) of the distribution could also be calculated on the available sample and be used to substitute missing values. The mean imputation has the dubious advantage of using all the cases, but it has several disadvantages. First of all, although we have added cases, we have not added new information (the overall mean, with or without replacing missing data, is the same), and any change is in some way spurious. Moreover, the standard error is biased downward (holding the numerator constant while increasing the denominator automatically reduces the standard error), leading to an inappropriate test on coefficient and incorrect confidence interval for the population mean. In general, *unconditional mean imputation* is not a particularly good way to proceed when you have missing data.

3.2 Regression imputation

One additional fairly simple approach for the treatment of missing data is to regress the variable that has missing observations on the other independent variables (or even variables not used in the study), thus producing a model for estimating the value of a missing observation. We then use the regression equation to impute a value for that variable whenever an observation is missing. Suppose to have a number $p > 1$ of independent variables X_1, X_2, $\cdots$, X_p, then the missing values are commonly predicted from the regression as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \qquad (3\text{-}1)$$

where Y denotes the dependent variable or response, X_1, X_2, $\cdots$, X_p are the independent or predictor variables, and β_0, β_1, $\cdots$, β_p are unknown coefficients.

Like most imputation procedures, *regression imputation* assumes missing values are MAR (but not necessarily MCAR). The regression method also assumes homogeneity of regression, meaning that the same model explains the data for the non-missing cases and for the missing cases. If this assumption is false, the imputed values may be quite different from what the values would be if we had been able to measure them. In general, the *regression imputation* has at least one advantage over the *unconditional mean imputation*, which is that the imputed value is in some way conditional on other information we have. So this approach is also occasionally called *conditional mean imputation*. Particularly, when there is a strong relationship between the variable that has missing observations and other independent variables, *regression imputation* is thought to work reasonably well. However, regression imputation will increase the correlations among items because some of the items will have been explicitly calculated as a linear function of other items. This will affect the regression coefficients that result from the analysis. Moreover, the problem of error variance remains. By imputing a value that is perfectly predictable from other variables, it would be expected to have less error than if the value was not missing. Thus *regression imputation* is likely to underestimate the standard error of the regression coefficients by underestimating the variance in the imputed variable, which means the inference based on the entire data set (including the imputed data) cannot fully account for imputation uncertainty. As a result, another alternative solution, i.e., *stochastic regression imputation*, is proposed, which imputes a *conditional draw* instead of imputing the *conditional mean*. Specifically, it is realized by deliberately adding a random error to the imputed observation as shown below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \qquad (3\text{-}2)$$

where ε is a random variable with normal distribution $N(0, \sigma^2)$ and σ^2 is the residual variance from the regression. Its presence represents the absence of an accurate relationship between dependent and independent variables, and it is used by default in the SPSS MVA option. Although this additional error does not completely solve the problem, it is an attempt to reduce the negative bias in the estimated standard errors (see Acock, 2005).

In addition, the above regression models (3-1) and (3-2), however, only consider the situation with quantitative variables. In case a qualitative variable is taken into account, suppose it is an independent variable X_i, dummy dichotomous variables will then be introduced into the model; While if the dependent variable Y is dichotomous or categorical variable with more than two categories, *logistic regression* or *multinomial logistic regression* will be performed to express the corresponding relationship.

3.3 The indicator method

Another method for handing missing data is the so-called *indicator method*. For each independent variable with missing values a new dummy or indicator (0/1) variable is created with '1' indicating a missing on the original variable and '0' indicating an observed value. For the original variable the missing values are recoded as '0'. For (original) categorical variables this in fact means creating an extra value category for the missing values. When estimating the association between the independent variable and the outcome in a multivariable analysis, the indicator is always included together with the original (though recoded) variable. The main advantage of the *indicator method* is that all subjects are used in the multivariable analysis. However, Jones (1996) has shown that coding for missingness when we have multiple independent variables can lead to bias in both the regression coefficients and their standard errors.

3.4 Hot deck imputation

A different way to impute is through matching, i.e., for each unit with a missing *Y*, find a unit with similar values of *X* in the observed data and take its *Y* value. This is one of the earliest methods of imputing missing data, which is known as *hot deck imputation* (in contrast to *cold deck imputation*, where the imputations come from an external source, e.g., a previously collected data set of the same survey). Scheuren (2005) provided an interesting discussion on how the hot deck procedures were developed within the US Census Bureau. In the 1950s most citizens seemed to feel that they had an obligation to respond to government surveys, and the non-response rate was low. In an effort to deal with unit non-response, data cards for respondents were duplicated, and non-responders were replaced by a random draw from these duplicate cards. Thus if you were missing a respondent of a certain gender from a certain census track, a draw was made from the data of respondents of that gender residing in that census track and the analysis continued. The method worked well when only a small amount of data was missing, and the variance properties of the method were understood [Hansen et al., 1953]. Unfortunately, the response rate to any survey or census has fallen over the years, and as we replace more and more data, the properties of our estimators, particularly their standard errors, become a problem. *Hot deck imputation* is not common today, although it is apparently useful in some settings.

3.5 Deductive imputation

There are still some imputation methods in which the missing data are deduced from existing information or rules.

3.5.1 Using information from related observation.

In some cases, missing observations can be imputed by using information from related observations. Suppose we are missing data regarding the income of fathers of children in a data set. Then we can fill these values in with the values of the mother. This is a plausible strategy, although these imputations may propagate measurement error. Also, we must consider whether there is any incentive for the reporting person to misrepresent the measurement for the person about whom he or she is providing information.

3.5.2 Imputation based on logical rule.

Sometimes we can impute using logical rules: for example, the Social Indicators Survey [Meyers and Garfinkel, 1999] includes a question on 'number of months worked in the previous year,' which all 1501 respondents answered. Of the persons who refused to answer the earnings question, 10 reported working zero months during the previous year. Thus we could impute zero earnings to them. This type of imputation strategy does not rely on particularly strong assumptions since, in effect, the missing data mechanism is known.

3.5.3 Last value carried forward.

This technique can be used when the data are longitudinal (i.e. repeated measures have been taken per subject) [Molnar et al., 2008]. The last observed value is used to fill in missing values at a later point in the study. Therefore, one makes the assumption that the response remains constant at the last observed value, rather than declining or improving further. Moreover, it also assumes that missing data are MCAR.

3.6 Summary

In this Chapter, we present some traditional *single imputation* methods which are still widely adopted in many real data analyses nowadays. Their main properties are summarized in the following Table 3-1.

Table 3-2 Summary information on the traditional imputation methods

	Method description	Main advantages	Main disadvantages	Assumptions	When to use
Unconditional mean imputation	Replace each missing value with the mean of the observed values for that variable	Easy to implement	No change on the overall mean Standard error is biased downward	Data are at least MAR	If no other methods available
Regression imputation	Regress the variable that has missing observations on the other independent variables	Conditional on other available information Possible for both numerical and categorical data	Artificially increase the correlations Cannot fully account for imputation uncertainty	Data are at least MAR The same model explains the data for the missing and the non-missing cases	Often, especially when the related independent variables are found
The indicator method	Create a new dummy or indicator variable for missing values	Use all subjects in the multivariable analysis	Leads to biased estimates when there are multiple independent variables Redefine the variables	Data are at least MAR	If no other methods available
Hot deck imputation	Fill in the missing attribute value with a value from an estimated distribution for the missing value from the current data	Easy to implement	May not work well when a large amount of data are missing	Data are at least MAR	Not common today
Using information from related observation	Data are imputed by using information from related observations	Easy to understand	May propagate measurement error	Data are MCAR	If we can find such observations and no other methods available
Imputation based on logical rule	Data are imputed by using logical rules	Easy to understand	Difficult to find such rules	Data are MCAR	If only we can find such rules

Last value carried forward	The last observed value is used to fill in missing values at a later point in the study	Easy to implement	May result in biases	Data are MCAR	When the data are longitudinal and no other methods available

4. ADVANCED APPROACHES

During the last three decades, substantial progress has been made in developing statistical procedures for missing data. Amongst others, two most important solutions --- one relies on *maximum likelihood estimation (MLE)* using *expectation/maximization* (known as the *EM algorithm*) [Dempster et al., 1977] and the other involves *multiple imputation (MI)* [Rubin, 1987] --- have become the mainstream approaches in recent literature.

4.1 Maximum likelihood estimation

Suppose that *X* denotes the data set. In the likelihood based estimation the data are assumed to be generated by a model described by a probability or density function $f(X/\theta)$, where θ is the unknown parameter vector lying in the parameter space Ω_θ (e.g., the real number for means, the positive real number for variances, and the interval [0,1] for probabilities). The probability function captures the relationship between the data set and the parameter of the data model, and describes the probability of observing a data set for a given $\theta \in \Omega_\theta$. Since θ is unknown while the data set is known, it makes sense to reverse the argument and look for the probability of observing a certain θ given the data set *X*, which is the likelihood function. Therefore, given *X*, the likelihood function $L(\theta/X)$ is any function of $\theta \in \Omega_\theta$ proportional to $f(X/\theta)$:

$$L(\theta/X) = k(X)f(X/\theta) \tag{4-1}$$

where $k(X) > 0$ is a function of *X* (not of θ). The log-likelihood is then the natural logarithm of the likelihood function. In the case of *M* independent and identically distributed observations $X = (x_1, x_2, \cdots x_M)^T$, from a normal population with mean μ and variance σ^2 the joint density is

$$f(X/\mu,\sigma^2) = (2\pi\sigma^2)^{-M/2}\exp(-\frac{1}{2}\sum_{i=1}^{M}\frac{(x_i-\mu)^2}{\sigma^2}) \tag{4-2}$$

For a given sample *X* the log-likelihood is a function of (μ,σ^2):

$$\begin{aligned} l(\mu,\sigma^2/X) &= \ln[L(\mu,\sigma^2/X)] \\ &= \ln[k(X)f(X/\mu,\sigma^2)] \\ &= \ln k(X) - \frac{M}{2}\ln 2\pi\sigma^2 - \frac{1}{2}\sum_{i=1}^{M}\frac{(x_i-\mu)^2}{\sigma^2} \end{aligned} \tag{4-3}$$

Maximizing the likelihood function corresponds to the question of which value of $\theta \in \Omega_\theta$ is mostly supported by a given sampling realization *X*. This implies solving the likelihood equation:

$$D_l(\theta/X_{obs}) \equiv \frac{\partial \ln L(\theta/X_{obs})}{\partial\theta} = 0 \tag{4-4}$$

The principle of maximum likelihood is fairly simple, but the actual solution is computationally complex. When a closed-form solution of equation (4-4) cannot be found, iterative methods need to be applied.

One of the most commonly used iterative methods to obtain maximum likelihood estimators is called the *Expectation-Maximization algorithm*, abbreviated as the EM algorithm. The issue is that *X* contains both observable and missing values, i.e., $X = (X_{obs}, X_{mis})$. Thus one has to find both the unknown parameters and the unknown observations of the model. As Schafer (1999) have noted: "If we knew the missing values, then estimating the model parameters would be straightforward. Similarly, if we knew the parameters of the data model, then it would be possible to obtained unbiased predictions for the missing values." Here we are going to do both. Assume that missing data are MAR (or MCAR), the EM consists of two components, the expectation (E) and maximization (M) steps. Each step is completed once within each algorithm cycle. Cycles are repeated until a suitable convergence criterion is satisfied. Specifically, in the M step, the maximum likelihood estimation of θ is computed just as if there were no missing data (thus missing values are replaced by estimated values. Especially, in the first round of maximization, we would just take estimates of the variances, covariances and means, perhaps from *listwise deletion*). In the E step, the missing data are estimated by their expectations given the observed data and current estimated parameter values (in order to deal with the problem of underestimating the error in choosing estimates, the EM algorithm gets around this by adding a bit of error to the variances it estimates). In the following maximization step, the parameters in θ are re-estimated using maximum likelihood applied to the observed data augmented by the estimates of the unobserved data (coming from the previous round). The whole procedure is iterated until convergence (absence of changes in estimates and in the variance-covariance matrix) is reached. Effectively, this process maximizes, in each cycle, the expectation of the complete data log-likelihood. On convergence, the fitted parameters are equal to a local maximum of the likelihood function (which is the maximum likelihood in the case of a unique maximum).

There are alternative maximum likelihood estimators that will probably be better than the ones obtained by the EM algorithm, such as the *Newton-Raphson algorithm* and the *Fisher scoring method*. However, both involve a calculation of the matrix of second derivatives of the likelihood, which, for complex patterns of incomplete data, can be a very complex function of θ, thus often require algebraic manipulations and complex programming. As a result, in certain classes of models—finite mixtures, for example—EM is still the method of choice [McLachlan and Peel, 2000].

The advantage of EM is its broadness (it can be used for a broad range of problems, e.g., factor analysis), its simplicity (EM algorithm is often easy to construct conceptually and practically), and each step has a statistical interpretation and convergence is reliable. So, in theory, maximum likelihood estimation (using the EM algorithm) is more attractive than ad hoc techniques of case deletion and traditional single imputation. However, it still rests on a few crucial assumptions. The most important one is that it assumes that the sample is large enough for the maximum likelihood estimates to be approximately unbiased and normally distributed. In missing data problems the sample may have to be larger than usual, because missing values effectively reduce the sample size, and moreover, with a large fraction of missing information, convergence may be very slow. The user should also be sure that the maximum found is indeed a global maximum and not a local one. To test this, different initial starting values for each θ can be used.

4.2 Multiple Imputation

The imputation methods mentioned in the previous Chapters, also including the *maximum likelihood estimation*, are all related to *single imputation*, i.e., each missing value in a data set is replaced with one imputed value. If the simplicity is its main appeal, an important limitation of these methods is that they systematically underestimate the variance of the estimates. One solution was the one used in the EM algorithm, where we altered the calculational formulae by adding in a bit of error in the calculation. However, this doesn't solve the problem completely. A valuable way is to repeat the imputation several times, generating multiple sets of new data whose coefficients vary from set to set. We then capture this variability and add it back into our estimates. This technique is known as *multiple imputation* (MI).

MI, proposed by Rubin (1987), has emerged as a flexible alternative to likelihood methods for a wide variety of missing-data problems. MI retains much of the attractiveness of single imputation from a conditional distribution but solves the problem of understating uncertainty. It is now becoming the dominant approaches to the treatment of missing data. A discussion of this method can be found in [Rubin, 1996, Schafer and Olsen, 1998, Allison, 2001, and Howell, 2008].

The interesting thing about MI is that the word 'multiple' refers not to the iterative nature of the process involved in imputation but to the fact that we impute multiple complete data sets and run whatever analysis is appropriate on each data set in turn. We then combine the results of those multiple analyses using fairly simple rules put forward by Rubin (1987). In a way it is like running multiple replications of an experiment and then combining the results across the multiple analyses. But in the case of MI, the replications are repeated simulations of data sets based upon parameter estimates from the original study. Graphically, each missing value is replaced by a list of N>1 simulated values as depicted in Figure 4-1. Substituting the *j*th element of each list for the corresponding missing value, $j = 1, 2, \cdots, N$, produces *N* plausible alternative versions of the complete data. Each of the *N* data sets is analyzed in the same fashion by a complete-data method. The results, which may vary, are then combined by simple arithmetic to obtain overall estimates and standard errors that reflect missing-data uncertainty as well as finite sample variation.

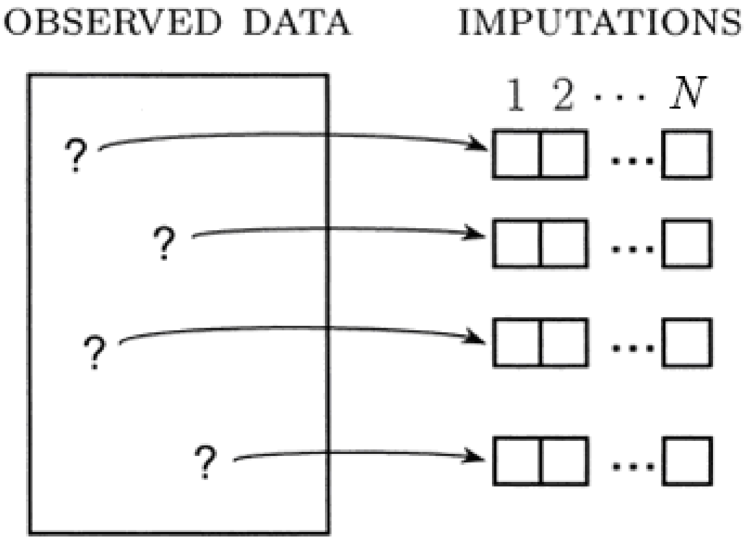

Figure 4-1 Schematic representation of multiple imputation (Source: [Schafer and Olsen, 1998])

There are a number of ways of performing MI, though they all involve the use of random components to overcome the problem of underestimating the standard errors. One of the most general models is the Markov Chain Monte Carlo (MCMC) method (see Figure 4-2). The Markov chain is a sequence of random variables in which the distribution of the actual element depends on the value of the previous one. It assumes that data are drawn from a multivariate normal distribution and requires MAR or MCAR assumptions.

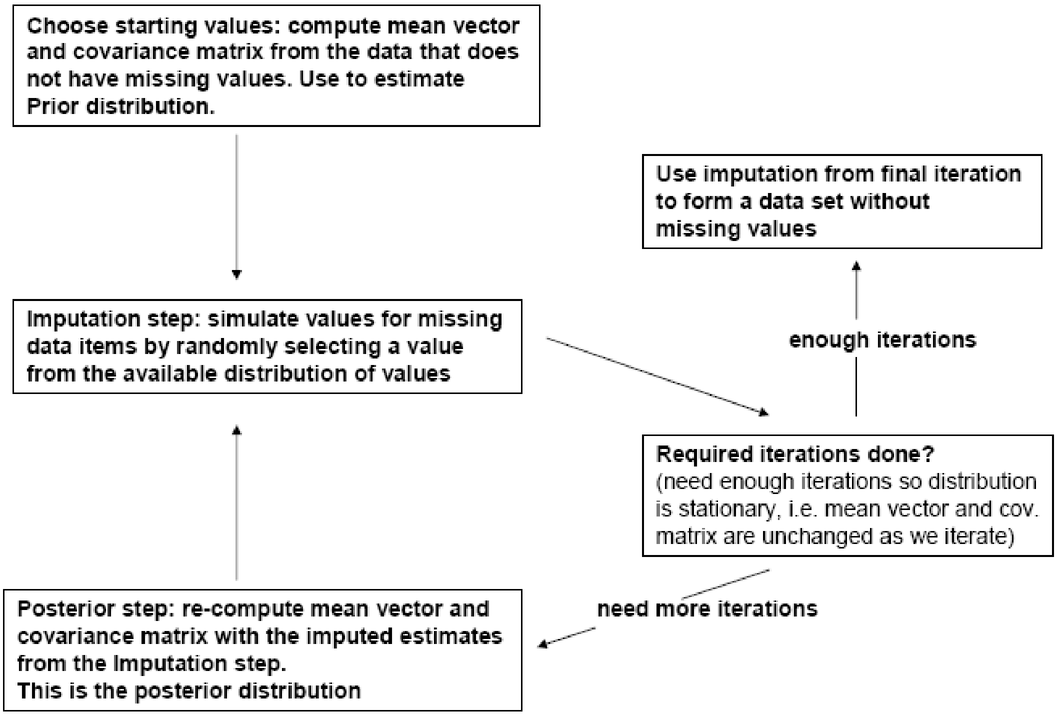
Choose starting values: compute mean vector and covariance matrix from the data that does not have missing values. Use to estimate Prior distribution.
Imputation step: simulate values for missing data items by randomly selecting a value from the available distribution of values
Posterior step: re-compute mean vector and covariance matrix with the imputed estimates from the Imputation step.
This is the posterior distribution
Required iterations done?
(need enough iterations so distribution is stationary, i.e. mean vector and cov. matrix are unchanged as we iterate)
enough iterations
need more iterations
Use imputation from final iteration to form a data set without missing values

Figure 4-2 Functioning of MCMC method (Source: [OECD, 2008])

The process of MI, at least as carried out through data augmentation, involves two random aspects. First, the imputed value contains a random component from a standard normal distribution. Second, the parameter estimates used in imputing data are a random draw from a posterior probability distribution of the parameters.

The process of MI via data augmentation with a multivariate normal model is relatively straightforward. The first step involves the imputation of a complete set of data from parameter estimates derived from the incomplete data set. We could obtain these parameters directly from the incomplete data using *listwise* or *pairwise deletion*; or, as suggested by Schafer and Olsen (1998), we could first apply the EM algorithm and take the parameter estimates from the result of that procedure.

Under the multivariate normal model, the imputation of an observation is based on regressing a variable with missing data on the other variables in the data set. Assume, for simplicity, that *Y* was regressed on only one other variable (*X*). Denote the standard error of the regression as s_{YX}. In standard regression imputation the imputed value of *Y*, i.e., $\hat{Y}$, would be obtained as:

$$\hat{Y} = \beta_0 + \beta_1 X \tag{4-5}$$

When discrete data are taken into account, the polytomous regression [Farhangfar et al., 2007, 2008] will then be utilized.

Moreover, for data augmentation we will add random error to our prediction by setting

$$\hat{Y} = \beta_0 + \beta_1 X + u s_{YX} \tag{4-6}$$

where *u* is a random draw from a standard normal distribution. This introduces the necessary level of uncertainty into the imputed value. Following the imputation procedure just described, the imputed value will contain a random error component. Each time we impute data we will obtain a slightly different result.

But there is another random step to be considered. The process above treats the regression coefficients and the standard error of regression as if they were parameters, when in fact they are sample estimates. And parameter estimates have their own distribution. So the second step will be to make a random draw of these estimates from their posterior distributions—the distribution of the estimates given the data, or pseudo-data, at hand.

Having derived imputed values for the missing observations, MI now iterates the solution, imputing values, deriving revised parameter estimates, imputing new values, and so on until the process stabilizes. At that point we have our parameter estimates and can write out the final imputed data file.

However, the MI process does not stop yet. Having generated an imputed data file, the procedure continues and generates several more data files. We do not need to generate many data sets, because Rubin (1987) has shown that in many cases three to five data sets are sufficient. Because of the randomness inherent in the algorithm, these data sets will differ somewhat from one another. Consequently, when some standard data analysis procedure (e.g., *multiple linear regression*) is applied to each set of data, the results will differ slightly from one analysis to another. At this point we will derive the final set of estimates by averaging over these estimates following a set of rules provided by Rubin (1987).

In conclusion, the *multiple imputation* method imputes several values (*N*) for each missing value (from the predictive distribution of the missing data), to represent the uncertainty about which values to impute. The *N* versions of completed data sets are analyzed by standard complete data methods and the results are combined using simple rules to yield single combined estimates (e.g., regression coefficients), standard errors, *p*-values, that formally incorporate missing data uncertainty. The pooling of the results of the analyses performed on the multiply imputed data sets implies that the resulting point estimates are averaged over the *N* completed sample points, and the resulting standard errors and *p*-values are adjusted according to the variance of the corresponding *N* completed sample point estimates. Thus, the variance of these estimates is divided into two components, the average within imputation variance and the between imputation variance. The total variance is then a weighted sum of these two variance components. In this way, the 'between imputation variance' provides a measure of the extra inferential uncertainty due to missing data, which is not reflected in *single imputation*. In addition, although nearly all MI analyses to date have assumed that the missing data are MAR, a few MNAR applications have been published (e.g., Glynn et al. (1993); Verbeke and Molenberghs (2000)). Nothing in the theory of MI requires us to keep the MAR assumption, and new methods for generating MIs under MNAR models will certainly become available in the future.

5. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING APPROACHES

At the time *multiple imputation* was just proposed, it was difficult to implement this method in large data sets, due to the amount of computer memory needed to store the different, multiply imputed data sets and the time required to run the analysis. In recent years, increased computer power and decreased cost have encouraged more research into the automated edit and imputation techniques. Advances in computer software and increased memory have made the use of MI more practical. At the same time, some artificial intelligence and machine learning (AI&ML) techniques have arisen in the area of missing data treatment (either in a direct way or in cooperation with other methods), which push the missing data research forward to a new stage.

5.1 Decision trees imputation

Decision trees are one of the classical artificial intelligence techniques. Relative to the regression analysis, which assumes a specific relationship between variables that may not hold for all data sets, decision trees provide one another way related to classification. For a variable with missing values, replacement values are estimated by treating this variable as a target, and using the remaining variables (and/or other variables) as predictors in a decision tree. Specifically, *decision trees imputation* uses a decision-tree based learning algorithm such as ID3 [Quinlan, 1993] to build a decision-tree classifier using the rows with no missing values, and the variable that has the missing value as the class variable. The tree is then evaluated on the row with the missing value to predict the missing value. For efficiency purposes, the decision tree construction can be 'lazy' [Friedman et al., 1996], in which only the needed path of the tree is constructed. For more information on decision trees imputation, we refer to Lakshminarayan et al. (1999), Farhangfar et al. (2004), and Saar-Tsechansky and Provost (2007).

5.2 Nearest-neighbour imputation

Nearest-neighbour imputation, also called *distance function matching*, is a donor method where the donor is selected by minimising a specified 'distance' [Kalton, 1983; Lessler and Kalsbeek, 1992; Rancourt, 1999; Chen and Shao, 2000]. This method involves defining a suitable distance measure, where the distance is a function of the auxiliary variables. The observed unit with the smallest distance to the non-respondent unit is identified and its value is substituted for the missing item according to the variable of concern. The easiest way is to consider just one continuous auxiliary variable X_1 and to compute the distance D from all respondents to the unit with the missing item, i.e., $D_{ji} = \left| x_{j1} - x_{i1} \right|$, where j denotes the unit with the missing item in Y. The missing item is replaced by the value y_{i*}, where the respondent $i*$ is the donor for non-respondent j if $D_{ji*} = \min_i \left| x_{j1} - x_{i1} \right|$.

An advantage of *nearest-neighbour imputation* is that actually observed values are used for imputation. Another advantage may be that if the cases are ordered for example geographically, it introduces geographical effects. However, it should be noted that the outcome could depend on the chosen order of the file. Some values might be used several times for imputation if more than one missing value occurs in a row, others may not be used at all. The variance of the estimates under *nearest-neighbour imputation* may be inflated if certain donors are used much more frequently than others. One solution is to penalize or restrict the multiple usage of donors for imputation to a certain number of times. For example, the distance function can be defined as $D_{ji*} = \min_i \left\{ \left| x_{j1} - x_{i1} \right| (1 + \eta t_i) \right\}$, where $\eta \in R^+$ is the assigned penalty for each usage, t_i is the number of times the respondent *i* has already been used as a donor [Kalton, 1983]. Another often used method for measuring the proximity between records for mixed data types is the Gower's general similarity coefficient (see Gower, 1971).

5.3 Clustering imputation

As one of the most popular techniques in data mining, the *clustering method* can also be used for missing data imputation. Given a set of objects, the overall objective of clustering is to divide the data set into groups based on similarity of objects and to minimize the intra-cluster dissimilarity. In *K*-means clustering, the intra-cluster dissimilarity is measured by the summation of distance from centroid which represents the mean value of the objects in a cluster. A number of different distance functions, e.g., Euclidean distance, Cosine-based distance, can be adopted.

The *K*-means clustering based imputation can be divided into three steps. First, randomly select *K* complete data objects as *K* centroids. Rather than random selection, an alternative is to choose the first centroid as the object that is most central to the data set, and then pick other (*K*-1) centroids one by one in such a way that each one is most dissimilar to all the objects that have already been selected. This makes the initial *K* centroids evenly distributed. Second, iteratively modify the partition to reduce the sum of the distances for each object from the centroid of the cluster to which the object belongs. The process terminates once the summation of distances is less than a user-specified threshold θ. The last step is to fill in all the non-reference variables for each incomplete object based on the cluster information. Data objects that belong to the same cluster are taken as nearest neighbours of each other. The missing data are replaced using the Inverse Distance Weighted (IDW) approach based on the available data values from nearest neighbours.

5.4 Neural networks

Artificial *neural networks* (ANNs) are computational architectures that combine simple units in an arrangement that can then exhibit complex behavior. Over the years, ANNs have proven to be a very powerful tool capable of extracting reliable information and patterns from complicated data even in the absence of models describing the data, flexible in modeling many types of nonlinear relationships, and suitable to account for unit non-response as well. The most commonly used type of ANNs applied for missing data imputation are called feedforward, where input terminals receive values of explanatory variables *X*, whereas the output provides the imputed variable *Y*. Multilayer feedforward networks consist of one or more hidden layers. The role of the hidden layer of neurons is to intervene between the external input and the network output. Inputs and outputs are connected through neurons that transform the sum of all received input values to an output value, according to connection weights and an activation function. The connection weights represent the strength of the connection between the neurons. The network weights (parameters) are randomly initialized and are then changed in an iterative process to reflect the relationships between the inputs and outputs. Many linear or nonlinear functions are suitable candidates for an activation function. ANNs do not require a model, which is advantageous in large data set imputation. However, this technique remains a kind of "black-box" in many critical application domains. More discussion about ANNs imputation can be found in Wang (2005); Lingras et al. (2008); Amer (2009); Silva-Ramírez et al. (2010).

5.5 Fuzzy logic systems

Fuzzy sets were introduced by Zadeh in 1965 to manipulate data and information possessing non-statistical uncertainties. *Fuzzy logic systems* are specifically designed to represent mathematical uncertainty and vagueness and to provide formalized tools for dealing with the imprecision intrinsic to many problems. They provide an inference morphology that enables approximate human reasoning capabilities to be applied to knowledge-based systems. The theory of fuzzy logic provides a mathematical strength to capture the uncertainties associated with human cognitive processes, such as thinking and reasoning. In recent years, using fuzzy logic for missing data research has been widely studied, and it is usually incorporated with other techniques. Some of the good applications are fuzzy cluster imputation [Li et al., 2006], fuzzy majority imputation [Peláez et al., 2007], fuzzy preference relation [Herrera-Viedma et al., 2007], and so on. For example, fuzzy logic can be incorporated into the *K*-means clustering constituting a fuzzy *K*-means clustering, which provides a better tool when the clusters are not well separated, as is sometimes the case in missing data imputation. Moreover, the original *K*-means clustering may be trapped in a local minimum if the initial points are not selected properly. However, continuous membership values in fuzzy clustering make the resulting algorithms less susceptible to get stuck in a local minimum. For more research work on fuzzy imputation we refer to Pospiech-Kurkowska (2008).

5.6 Rough sets

Rough sets theory originated by Pawlak (1991) is a formal mathematical theory modeling knowledge about the domain of interest in terms of a collection of equivalence relations. It is mainly an automated transformation of data into knowledge without any preliminary or additional information about the data like probability in probability theory, or grade of membership in fuzzy sets theory. It is based on the concept of an upper and a lower approximation of a set, approximation space, reduct and core, etc. The concepts in rough sets theory are used to define the necessity of features. The measures of necessity are calculated by the lower and upper approximations. These measures are employed as heuristics to guide the feature selection process. Nowadays, many rough sets based approaches have been successfully applied in the field of knowledge discovery, and missing data treatment is one of its important applications. In this respect, one direction is to use rough sets as a data mining technique in addition to the decision tree method. Specifically, it aims to reduce knowledge by eliminating only the non-essential information in order to classify or to make decisions, which is based on its discernibility function. Another direction is related to the third strategy to deal with missing data---*using as it is*, i.e., data sets are not preliminarily converted into complete data sets. In doing so, missing values are classified in three groups: 'lost values', 'irrelevant (or attribute concept) values', and 'do not care (or non relevant) values'. In each case, a decision to complete or to expand or to not include the data is taken. For more information, we refer to Kryszkiewicz (1998); Grzymala-Busse (2008); Wang and Wang (2009).

6. AVAILABLE SOFTWARE

One of the original problems for dealing with missing data, especially with those complicated techniques such as EM and MI, was the lack of statistical software. That is no longer a problem. The statistical literature is filled with papers on the algorithm and a number of programs exist to do the calculations. Below some of the most popular software for the missing data problem is listed [MathWorks Inc. 1999; Scheffer, 2000; Raghunathan et al., 2002; Yuan, 2004; SPSS Inc., 2007].

- SPSS: Missing Values Analysis (MVA) Module, an optional module for SPSS; Available for all value analysis, listwise deletion, pairwise deletion, regression imputation, and EM imputation. From SPSS 17.0 on, a Multiple Imputation Module is included for multiple imputation.
- SAS: %SingleImpute, as a SAS macro, provides an easy and quick approach for dealing with missing data. PROC MI and PROC MIANALYZE, which were introduced as experimental software in Releases 8.1 and 8.2, are production software in Version 9.0. It is a quite complete (and can be complex) implementation of a variety of model based imputation procedures for creating multiple imputations for incomplete multivariate data and for analyzing results from multiply imputed data sets.
- IVEware (Imputation and Variance Estimation Software): It is a free package from the University of Michigan ISR and can be downloaded from the internet (http://www.isr.umich.edu/src/smp/ive/). It is a program that runs in the SAS environment. The program can be used for both imputation and variance estimation in survey data (clustered, stratified, and weighted data). It uses a 'sequence of regression models' to yield the imputations. The regression models can differ depending on the level of measurement of the variable. Bounds can be imposed and cases can be identified in which imputations will not be created for a given variable.
- R: It is a programming language and software environment for statistical computing and graphics. The R language has become a de facto standard among statisticians for the development of statistical software, and is widely used for missing data imputation.
- SOLAS: This is from Statistical Solutions, Ireland. Available are group means, last value carried forward (for longitudinal data), hot deck imputation, and multiple imputation.
- S-PLUS: It supports the Norm, Cat, Mix and Pan libraries, which use the MCMC multiple imputation. Also under the S-PLUS platform is MICE (multiple imputation by chained equations).
- BMDP: It has routines available to impute data, using both single and multiple imputation.
- BUGS: MCMC Multiple Imputation is a natural extension of Bayesian analysis.
- MATLAB: It contains the Neural Network Toolbox and the Fuzzy Logic Toolbox, which extend the MATLAB technical computing environment with tools for designing systems based on neural network and fuzzy logic. However, they are not particularly designed for missing data imputation.

Some other available software implementations of missing data treatment can be found in the following websites:

- A web devoted to missing values with available software: http://missingdata.lshtm.ac.uk/index.php.
- A data mining tool --- KEEL: http://keel.es, with source code for individual imputation techniques.

- Schafer's Multiple Imputation Webpage with Windows binaries: http://www.stat.psu.edu/~jls/misoftwa.html.
- Regularized Expectation-Maximization Algorithm homepage with MATLAB code: http://www.gps.caltech.edu/~tapio/imputation/.

7. Conclusions

This report discussed a range of techniques for dealing with missing data. Some of the earlier solutions, such as *listwise* and *pairwise deletion*, *mean imputation*, the *indicator method* and *hot deck imputation*, are slowly tending to fall by the wayside because they lead to bias in parameter estimation especially when the data are not *missing completely at random*. Some other methods, such as *reweighting* and *regression imputation*, though having their own drawbacks more or less, are still widely adopted in many real data analyses due to their simplicity and effectiveness. However, the most important techniques, now that the necessary software is available, are the *maximum likelihood estimation* and *multiple imputation*. They both rely on iterative solutions in which the parameter estimates lead to imputed values, which in turn change the parameter estimates, and so on. *Multiple imputation* is an interesting approach because it uses randomized techniques to do its imputation, and then relies on multiple imputed data sets for the analysis. More recently, with the development of computer science and technology, some artificial intelligence and machine learning techniques, such as *decision trees imputation*, *nearest-neighbour imputation*, *clustering imputation, neural networks*, *fuzzy logic systems*, and *rough sets*, have arisen in the area of missing data research. They are still rapidly developing and will probably be the direction of choice for the next few years until something even better comes along.

In the field of road safety data analysis, missing values can exist in socio-economic costs associated with road trauma; final outcomes (e.g., the number of injuries); exposure measures (e.g., distances travelled); safety performance indicators (e.g., correct use of child restraint systems); policy performance indicators (e.g., different categories of enforcement effort); and background information (e.g., attitudes towards risk); etc. In order to handle these possible missing data, desired techniques should be:

- Structurally sound with accurate results
- Reliable with repeatable and robust results
- With an analytical method to estimate the uncertainty of the prediction
- User-friendly (simple, easily trained and easily retrained)
- Transparent

Therefore, the possible techniques mentioned above could be adopted based on the data sets information, such as the nature of data, the quality of data, and the source of data.

Moreover, apart from the approaches discussed in this report, some other missing data techniques are also worthwhile for exploration in the future, such as *singular value decomposition (SVD)* [Troyanskaya et al., 2001], *Bayesian missing value estimation* [Oba et al, 2003], *local least squares imputation (LLSI)* [Kim et al., 2005], *support vector machines (SVMs)* [Pelckmans et al., 2005], and so on. Moreover, as mentioned at the end of Section 1.2, most of the methods available now apply only when the data are at least *missing at random*. Without the MAR assumption, one must explicitly specify a distribution for the missingness in addition to the model for the complete data. In this respect, *selection models* and *pattern-mixture models* [Schafer and Graham, 2002] may be the two fundamentally different ways to realize it. Furthermore, in the ideal situation, one could say that the best treatment for missing data is not to have any, because any imputation method only provides an estimated value, and its facticity is always doubtable. Consequently, the third strategy of treatment of missing data, i.e., *using as it is*, may be another option. Recent techniques have also come far in narrowing the gap between the ideal and the practical. For instance, Kabak and Ruan (2010) proposed a cumulative belief degree based approach to solve missing data problems in nuclear safeguards evaluation, where the missing values are managed within this strategy. Li (2006) built a rule-based classification model to tolerate missing values. Instead of imputing missing values, it is proposed to make a system 'immunize' from missing values to a certain degree.

8. REFERENCES

Acock, A.C., (2005). Working with missing values, Journal of Marriage and the Family, Vol. 67, pp. 1012-1028.

Allison, P.D., (2001). Missing data, Thousand Oaks, CA: Sage Publications.

Amer, S., (2009). Neural network imputation in complex survey design, International Journal of Electrical and Electronics Engineering. Vol. 3, No. 1, pp. 52-57.

Chen, J. and Shao, J., (2000): Nearest Neighbour Imputation for Survey Data, Journal of Official Statistics, Vol. 16, No. 2, pp. 113-131.

Dempster, A.P., Laird, N.M. and Rubin, D.B., (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), Journal of the Royal Statistical Society, Series B, Vol. 39, pp. 1-38.

Dempster A.P. and Rubin D.B., (1983). Introduction. In: Madow, W.G., Olkin I. and Rubin D.B. (Eds.), Incomplete Data in Sample Surveys (vol. 2): Theory and Bibliography. New York: Academic Press, pp.3-10.

Dunning, T. and Freedman, D.A., (2008). Modeling Selection Effects. In: Outhwaite, W. and Turner, S. (Eds.), Handbook of Social Science Methodology. London: Sage, pp. 225-231.

Farhangfar, A., Kurgan, L. and Pedrycz, W., (2004). Experimental analysis of methods for handling missing values in databases. In: Intelligent Computing: Theory and Application II.

Farhangfar, A., Kurgan, L. and Pedrycz, W., (2007). Novel framework for imputation of missing values in databases, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, Vol. 37, No. 5, pp. 692-709.

Farhangfar, A., Kurgan, L. and Dy, J., (2008). Impact of imputation of missing values on classification error for discrete data, Pattern Recognition, Vol. 41, pp. 3692-3705.

Friedman, J.H., Kohavi, R. and Yun, Y., (1996). Lazy decision trees. In: 13th Artificial Intelligence and the 8th Innovative Applications of Artificial Intelligence, pp. 717-724.

Glynn, R.J., Laird, N.M. and Rubin, D.B., (1993). Multiple imputation in mixture models for nonignorable non-response with followups. Journal of American Statistical Association, Vol. 88, pp. 984-993.

Gower, J., (1971). A general coefficient of similarity and some of its properties, Biometrics, Vol. 27, No. 4, pp. 857-871.

Grzymala-Busse, J.W., (2008). Three approaches to missing attribute values: A rough set perspective. Studies in Computational Intelligence, Vol. 118, pp. 139-152.

Hansen, M.H., Hurwitz, W. and Madow, W., (1953). Sample Survey Methods and Theory, New York: Wiley.

Herrera-Viedma, E., Chiclana, F., Herrera, F. and Alanso, S., (2007). Group decision-making model with incomplete fuzzy preference relations based on additive consistency, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 37, No. 1, pp. 176-189.

Howell, D.C., (2008). The Analysis of Missing Data. In: Outhwaite, W. and Turner, S. (Eds.), Handbook of Social Science Methodology. London: Sage.

Jiang, N. and Gruenwald, L., (2008). Estimating Missing Data in Data Streams. In: Kotagiri, R., Krishna, P. R., Mohania, M. and Nantajeewarawat E. (Eds.), Advances in Databases: Concepts, Systems and Applications, Vol. 4443, pp. 981–987.

Jones, M.P., (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression, Journal of the American Statistical Association, Vol. 91, pp. 222-230.

Kabak, Ö. and Ruan, D., (2010). A cumulative belief degree-based approach for missing values in nuclear safeguards evaluation, IEEE Transactions on Knowledge and Data Engineering (http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.60), 2010.

Kalton, G., (1983). Compensating for Missing Survey Data, Michigan.

Kim, H., Golub, G.H. and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: Local least squares imputation, Bioinformatics, Vol. 21, No. 2, pp. 187-198.

Kryszkiewicz, M., (1998). Rough set approach to incomplete information systems, Information Science, Vol. 112, pp. 39-49.

Lakshminarayan, K., Harp, S. A. and Samad, T., (1999). Imputation of missing data in industrial databases. Applied Intelligence, Vol. 11, No. 3, pp. 259–275.

Lessler, J.T. and Kalsbeek W.D., (1992). Nonsampling Error in Surveys, New York: Chichester.

Li, D., Deogun, J., Spaulding, W. and Shuart, B., (2004). Towards missing data imputation: A study of fuzzy K-means clustering method, Lecture Notes in Computer Science, Vol. 3066, pp. 573-579.

Li, J., (2006). Robust rule-based prediction, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 8, pp. 1043-1054.

Lingras, P., Zhong, M. and Sharma, S., (2008). Evolutionary Regression and Neural Imputations of Missing Values. In: Prasad, B. (Ed.), Soft Computing Applications in Industry, Germany: Springer-Verlag, pp. 151-163.

Little, R.J.A., (1998). A test of missing completely at random for multivariate data with missing values, Journal of the American Statistical Association, Vol. 83, pp. 1198-1202.

Little, R.J.A. and Rubin D.B., (1987). Statistical Analysis with Missing Data. New York: John Wiley & Sons.

MathWorks Inc., (1999). Fuzzy Logic Toolbox User’s Guide, Version 2.

McLachlan, G. J. and Peel, D., (2000). Finite Mixture Models. New York: Wiley.

Meyers, M.K. and Garfinkel, I., (1999). Social indicators and the study of inequality, Economic Policy Review, Vol. 5, No. 3, pp. 149-163.

Molnar, F., Hutton, B. and Fergusson, D., (2008). Does analysis using “last observation carried forward” introduce bias in dementia research? Canadian Medical Association, Vol. 179, No. 8, pp. 751-753.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara K. and Ishii, S., (2003). A Bayesian missing value estimation method for gene expression profile data, Bioinformatics, Vol. 19, No. 16, pp. 2088-2096.

Organisation for Economic Cooperation and Development (OECD), (2008). Handbook on Constructing Composite Indicators: Methogology and User Guide. www.oecd.org/ publishing/corrigenda. Accessed Dec. 8, 2010

Oltman L.B. and Yahia, S.B., (2008). Yet Another Approach for Completing Missing Values. In: Yahia, S.B., Nguifo, E.M. and Belohlavek, R. (Eds.), Concept Lattices and Their Applications, Germany: Springer-Verlag, pp. 155-169.

Pawlak, Z., (1991). Rough sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston.

Peláez, J.I., Doña, J.M. and Gómez-Ruiz, J.A., (2007). Analysis of OWA operators in decision making for modelling the majority concept. Applied Mathematics and Computation, Vol. 186, pp. 1263-1275.

Pelckmans, K., De Brabanter, J., Suykens, J.A.K. and De Moor, B., (2005). Handling missing values in support vector machine classifiers, Neural Networks, Vol. 18, pp. 684-692.

Pickles, A., (2005). Missing Data, Problems and Solutions. In: Kimberly K.-L. (Ed.), Encyclopedia of Social Measurement. Amsterdam: Elsevier, pp. 689-694.

Pospiech-Kurkowska, S., (2008). Processing of missing data in a fuzzy system. In: Pietka E. and Kawa J. (Eds.), Information Technology in Biomedicine, Germany: Springer, pp. 453-460.

Quinlan, J.R., (1993). Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems, In: Induction of Decision Trees, Morgan Kaufmann Publishers Inc., USA, pp. 349-361.

Raghunathan, T.E., Solenberger, P. and Van Hoewyk, J., (2002). IVEware: Imputation and variance estimation software users guide. University of Michigan: Survey Research Center, Institute for Social Research.

Rancourt, E., (1999): Estimation with nearest neighbour imputation at statistics Canada, In: Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 131-138.

Robins, J.M., and Rotnitzky, A., (1995). Semi-parametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, Vol. 90, pp. 122-129.

Robins, J.M., Rotnitzky, A. and Scharfstein, D.O., (1998). Semi-parametric regression for repeated outcomes with non-ignorable non-response. Journal of the American Statistical Association, Vol. 93, pp. 1321-1339.

Robins, J.M., Rotnitzky, A. and Zhao, L.P., (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, Vol. 89, pp. 846-866.

Rodríguez, R., Martínez, L., Ruan, D. and Liu, J., (2010). Using collaborative filering for dealing with missing values in nuclear safeguards evaluation, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 18, No. 4, pp.431-449.

Rubin, D.B., (1976). Inference and missing data, Biometrika*,* Vol. 63, pp. 581-592.

Rubin, D.B., (1987). Multiple Imputation for Non-response in Surveys. New York: John Wiley & Sons.

Rubin, D.B., (1996). Multiple imputation after 18+ years, Journal of the American Statistical Association, Vol. 91, pp. 473-489.

Saar-Tsechansky, M. and Provost, F., (2007). Handling missing values when applying classification models, Journal of Machine Learning Research, Vol. 8, pp. 1625-1657.

Schafer, J.L., (1999). Multiple imputation: A primer, Statistical Methods in Medical Research, Vol. 8, pp. 3-15.

Schafer, J.L. and Graham, J.W., (2002). Missing data: Our view of the state of the art, Psychological Methods, Vol. 7, No. 2, pp. 147-177.

Schafer, J.L. and Olsen M.K., (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research. Vol. 33, pp. 545-571.

Scheffer, J., (2000). An Analysis of the Missing Data Methodology for Different Types of Data. Master Thesis, Massey University, Auckland.

Scheuren, F., (2005). Multiple imputation: How it began and continues, The American Statistician, Vol. 59, pp. 315-319.

Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M., Cubiles-de-la-Vega, M.-D., (2010). Missing value imputation on missing completely at random data using multilayer perceptrons, Neural Networks, in press.

SPSS Inc., (2007). SPSS Missing Values™ 17.0. http://www.spss.com. Accessed Dec. 8, 2010.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B., (2001). Missing value estimation methods for DNA microarrays, Bioinformatics, Vol. 17, pp. 520-525.

Verbeke, G. and Molenberghs, G., (2000). Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

Wang H. and Wang, S., (2009). Discovering patterns of missing data in survey databases: An application of rough sets, Expert Systems with Applications, Vol. 36, pp. 6256-6260.

Wang, S., (2005). Classification with incomplete survey data: a Hopfield neural network approach, Computers & Operations Research, Vol. 32, pp. 2583-2594.

Yuan, Y., (2004). Multiple imputation for missing data: Concepts and new development (Version 9.0), SAS Institute Inc.