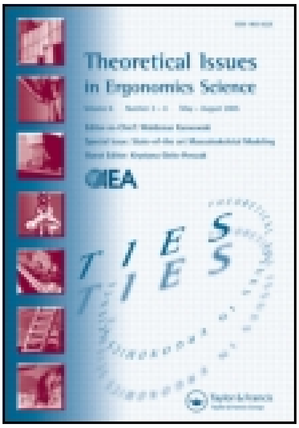

Theoretical Issues in Ergonomics Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ttie20

A survey of methodologies for the treatment of missing values within datasets: limitations and benefits

W. Young [a], G. Weckman [a] & W. Holland [a]

[a] Industrial and Systems Engineering, Ohio University, 270 Stocker Center, Athens, OH 45710, USA

Published online: 15 Jun 2010.

To cite this article: W. Young , G. Weckman & W. Holland (2011) A survey of methodologies for the treatment of missing values within datasets: limitations and benefits, Theoretical Issues in Ergonomics Science, 12:1, 15-43, DOI: 10.1080/14639220903470205

To link to this article: http://dx.doi.org/10.1080/14639220903470205

Theoretical Issues in Ergonomics Science
Vol. 12, No. 1, January–February 2011, 15–43

A survey of methodologies for the treatment of missing values within datasets: limitations and benefits

W. Young*, G. Weckman and W. Holland

Industrial and Systems Engineering, Ohio University, 270 Stocker Center, Athens, OH 45710, USA

(*Received 29 January 2009; final version received 23 October 2009*)

Knowledge discovery in ergonomics is complicated by the presence of missing data, because most methodologies do not tolerate incomplete sample instances. Data-miners cannot always remove sample instances when they occur. Imputation methods are needed to 'fill in' estimated values for the missing instances in order to construct a complete dataset. Even with emerging methodologies, the ergonomics field seems to rely on outdated imputation techniques. This survey presents an overview of a variety of imputation methods found in current academic research, which is not limited to ergonomic studies. The objective is to strengthen the communities' understanding of imputation methodologies and briefly highlight their benefits and limitations. This survey suggests that the multiple imputation method is the current state-of-the-art missing value technique. This method has proven to be robust to many of the shortcomings that plague other methods and should be considered the primary choice for missing value problems found in ergonomic studies.

Keywords: data mining; imputation methods; machine learning; multiple imputation; missing values

1. Introduction

The goal of this survey is to expose the ergonomics field to the gamut of techniques that are used in academic research to replace missing values that appear in datasets with imputed values. This survey is vital to advance the field of ergonomics, because a significant amount of current research utilises outdated imputation techniques even though better options exist. Inadequately addressing missing values in ergonomics studies can lead to unjustifiably supporting or not supporting research hypotheses. Missing data are commonly present in ergonomic research datasets, which often rely on questionnaires or surveys to collect data. These datasets can become contaminated with missing values because respondents can choose not to answer one or more of the questions. Due to the nature of ergonomics data collection, it is possible that novel research has been prevented from being published due to missing values. In some cases, missing values can be prevented by intelligent design (Abel *et al.* 2005). However, in most settings, the means to prevent the occurrence of missing data is infeasible. In a study on the physical ergonomics of working

*Corresponding author. Email: william.a.young.1@ohio.edu

ISSN 1464–536X online
© 2011 Taylor & Francis
DOI: 10.1080/14639220903470205
http://www.informaworld.com

with a video-display terminal, the missing data were simply eliminated from the dataset (Bayeh and Smith 1999). Elimination of the parts of the dataset scattered with missing data is an option, but basal patterns could be lost, which can cause biased results. With a suitably large dataset, the deletion method could be tolerable; however, data are often difficult to obtain and losing the valuable information that exists in partially completed surveys is not desired. Current methods that have been used in ergonomic studies vary from selective deletion (Chen *et al*. 2004), to filling in the missing value with a zero (Han *et al*. 2000) or the attribute average (Haynes and Williams 2008), to filling in using a value generated by primitive machine learning or statistical methods (Batista and Monard 2003). Because of the human element involved with ergonomic studies, datasets and the relationships that researchers try to discover are often non-linear in nature (Waters *et al*. 2006). For example, to improve appraisals for the costs of work-related accidents, it has been said that non-linear attributes and predictors are needed (Sun *et al*. 2006). For this reason, ergonomic researchers must not rely on older primitive methods that are incapable of maintaining the dataset's integrity. Since contemporary imputation methods have not been widely used in ergonomics research, the reviews of the imputation methods presented in this article mostly come from other fields of study. For this reason, it is not currently possible to indicate a preferred imputation method that consistently yields as the best option. Thus, to progress the field of ergonomics, this paper will review the benefits and limitations of imputation methods found in general research and briefly review why some methods should be chosen.

There are several reasons why missing values can occur. At times, missing values appear in datasets because the values simply do not exist. For example, one anthropometric investigation indicated that physical feature dimensions for women had never been collected in a particular country (van Schoor and Konz 1996). For this situation, researchers knew the reason why the values were missing. Since the origin of the lack of these data was known, the researchers were able to estimate these missing values based on ratios that were obtained from the male counterparts in order to perform the data-mining method. In some situations, expert domain knowledge can provide insight into the reason why the values are missing. However, researchers do not usually possess this intimate knowledge and rely on other methods to discover patterns for reasons why missing values occur. Ergonomic research that utilises surveys is commonplace, for this research setting, rough set or association rules have been used further to explain why missing values occur when evaluating income statistics using a census dataset (Wang and Wang 2009). Understanding certain aspects as to why missing values are occurring can ultimately lead to better practices when collecting data.

Missing values commonly occur in ergonomics studies, especially datasets that are compiled from surveys. In these studies, the indicator itself is often responsible for the absent value. In a study investigating the associations between balance difficulties and depression, most of the missing data occurred in a question regarding depression (Baker *et al*. 2003). Even if anonymity is protected, people are still not comfortable answering certain questions that can appear on surveys, such as indicators about one's race or economic status. Thus, it is understandable that the survey participants could feel uncomfortable in supplying this information. Aside from the rare cases of controlled data collection, missing data will continue to be an issue of data analysis in ergonomic research. Thus, this paper will focus on providing a brief review of the methodologies and a discussion on how these methods have led to the development of the state-of-the-art imputation methodology.

2. Knowledge discovery

Knowledge discovery in databases (KDD) is the process of extracting non-obvious trends or patterns in datasets, which is an emerging study for ergonomist (Karwowski 2003). KDD is not only limited to theoretical research, but it is also widely used in many areas of science and reasoning disciplines, where practitioners require quantifiable evidence to aid in decision making. KDD can also assist in understanding complex systems, where the behaviour of the system components is not well known or understood. For example, ergonomists have built traditional multivariate models to determine musculoskeletal disorders in longitudinal data (Swanson and Sauter 2006) and even probabilistic networks to new relationships of human performance in visual search tasks (Sarac *et al*. 2007). There are five distinct phases in the process: (1) understand the domain of the dataset; (2) collect and pre-process data; (3) extract correlation structure; (4) interpret and evaluate discovered knowledge; (5) use results in a practical manner (Mannila 1996). The third phase is also referred to as the data-mining phase. Data-mining tasks can be classified into two groups: descriptive and predictive. Descriptive mining methods attempt to generalise properties about the dataset, whereas predictive mining attempts to create associations about the dataset. A problem in the data-mining phase occurs when there are missing values in the dataset.

The standard form of presenting data consists of a data matrix. The data matrix consists of attribute names in the first row of columns, followed by attribute values in subsequent rows in the attribute's respective column. This standard form is used to illustrate three types of problems encountered with missing values, and is shown in Figure 1. The first instance, Figure 1a, shows the occurrence of missing values along several attributes. The second instance, Figure 1b, shows the condition of missing values along several instances. The third instance, Figure 1c, demonstrates the situation when missing values occur randomly.

Statisticians (Little and Rubin 1987) have extended the instances described above as missing complete at random (MCAR), missing at random (MAR) and not missing at random (NMAR). MCAR occurs when values are missing unrelated to its value or the value of other variables. Thus, the probability of a missing record to occur within a dataset is equal for every instance. MAR occurs when the probability is statistically equal that a missing value is dependent on other variables in the dataset. Finally, NMAR occurs when the values are missing because of the attribute itself, such as an embarrassing or imposing question on a survey.

Missing values might hide or mask vital underlying knowledge that can be obtained from the KDD process. Thus, simply removing sample sets of data with missing values is not always an option especially when datasets are limited in sample size. Few data-mining

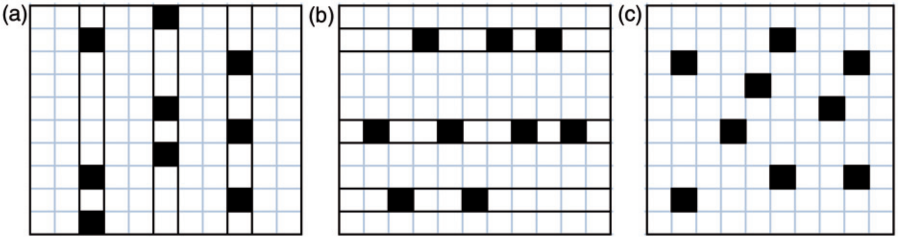

Figure 1. Missing data instances. (a) Along columns; (b) along rows; (c) randomly.

methods exist that can incorporate missing data directly into the data-mining phase (Stenbakken 2004). However, the number of existing data-mining techniques that incorporate full or complete datasets is enormous. Thus, either new methodologies must be created that incorporate missing values in the data-mining process or methods must be determined to 'fill in' missing values for existing data-mining methods. These 'filled-in' values are called imputes, where the process of 'filling in' is called imputation. The methodologies presented in this paper show that the available information in the datasets is sufficient for reconstructing the missing values.

Three distinctions are made from methodologies dealing with missing data. These are classified as disregarded, pre-replacement and embedded methods. The disregarded and pre-replacement methods are performed during the data pre-processing phase, which is before data-mining techniques are employed. Disregarded methods typically remove sample instances containing missing values. Pre-replacement methods, which are the most popular form of missing value methodologies, replace missing values with a value that is often statistically determined from the dataset. Embedded methods attempt to combine phases 2 and 3 of the KDD process. This method relies on the data-mining technique to treat the missing values for its internal algorithm.

Datasets vary in the types of either independent or dependent values, which can take the form of continuous or classification values. Datasets can also vary in the number of samples, distribution of data and the distribution of missing data. Although many methodologies exist for the missing-data problem, very few comprehensive lists are found in literature that summarise the benefits and limitations. The objective of this literature is to compile a list of previously researched missing value methods. The methods are summarised in the following sections, where an overview of the theory is provided. The conclusions of the methods are based on available research literature.

3. Missing data techniques

The following sections investigate disregarded, pre-replacement and embedded missing value methods. Specific research findings are presented after each method is explained.

3.1. *Disregarded methods*

Disregarded missing value methods are the easiest to employ to create a complete dataset for data mining. There are two conventional deletion techniques: list-wise and pair-wise. These two methods are commonly implemented in statistical software packages because of their simplicity and very low computational costs.

List-wise deletion removes any sample instance that contains a missing value in any of the independent or dependent variables. This method is the most commonly used missing value method (Acock 2005). However, list-wise deletion techniques are simply not always a viable option for many datasets with limited data. One industrial research found that none of the sample data collected could be useable for data mining when list-wise deletion techniques were implemented (Lakshminarayan *et al.* 1999). Because of list-wise deletion, none of the information collected could be used for data mining. List-wise methods can cause the most bias in the KDD process when sample sizes are small (Liu *et al.* 1997). Generally, extreme biases will occur by performing list-wise deletion when variables are MAR because the remaining datasets under-represent the population of study in question (von Hippel 2004). If datasets have large sample sizes and missing values are MAR or

MCAR, list-wise deletions offer an unbiased complete dataset (Allison 2000). In a study researching lower back injuries due to physical workload, researchers used list-wise techniques to discover several statistically significant factors in their research (Krause *et al.* 2004). Initially, their dataset was very large, with 1974 samples, and less than 10% of the data samples was removed due to missing values. Their result further suggests that this method is acceptable when the sample sizes are high and the missing value rate is low. However, further research is needed to determine the sample size that is needed to employ these techniques without biasing the inference obtained.

Pair-wise deletion is a variation of list-wise deletion. This technique uses incomplete records during the analysis or data-mining phase. In comparison to list-wise deletion, more accurate and less biased descriptive sample statistics can be computed for some of the attributes collected. This occurs in special cases where a larger sample size of values can be used to calculate statistics, when they would otherwise have been removed in list-wise deletion. Implementing pair-wise deletion often causes difficulty when calculating degrees of freedom, since different parts of the model have different sample sizes. In this case, selecting samples with the largest or fewest observations reduces statistical power. One advantage that pair-wise offers over list-wise is that all of the possible information is used within a dataset without having to determine an imputation statistic or value for the missing. However, the detail of information that can be gathered using this technique is very limited and, thus, other methods are desired. For example, Campion (1988) utilised pair-wise deletion to discover descriptive statistics such as correlations and reliabilities for motivational, mechanistic, biological and perception work design outcomes. This study was reconstructed years later (Edwards *et al.* 2000), when multiple-linear regression imputation, which is a pre-replacement statistical imputation method, was used to increase the sample sizes for their investigation. As a result, the recreation of the study significantly increased the confidence of their descriptive statistics over the original methodology.

Table 1 shows a summary of disregarded methods with N and C representing numerical or categorical. In addition, the computational cost is labelled as very low, low, medium, high or very high.

3.2. *Pre-replacement non-statistical methods*

The following section is dedicated to pre-replacement methods that do not compute imputation values based on statistical properties of the dataset. There are three conventional methods: last observation carried forward (LOCF); hot/cold deck; random assignment.

The LOCF method fills in an attribute's missing value with the previous sample's value. This method is often used in longitudinal studies where values do not vary greatly over time or where values for a certain subject contain values before and after the missing value has occurred. LOCF has been successfully used when an attribute that does not contain missing values can be arranged in a logical manner and has little variation (Carpenter *et al.* 2004). In medical longitudinal studies, this method has been claimed to be asymptotically valid based on a single ANOVA test, when the variable being imputed does not have substantial variation (Shao and Zhong 2003). However, researchers have cautioned against the use of LOCF in longitudinal studies (Cook *et al.* 2004). In their study regarding the smoking behaviours of young adults, many of the samples contained missing values. In total, nearly 28% of the samples contained missing values due to missed or skipped patient assessments or drop-outs. After LOCF was used, regression coefficients

Table 1. Disregarded missing value methodologies.

Method Name	Primary Benefit	Secondary Benefit	Primary Limitation	Secondary Limitation	Missing Variables	Computational Cost
List-wise	Unbiased with large datasets	Easy to implement	Extreme bias with small datasets	Does not use all available information	N&C	VL
Pair-wise	Uses all available information	Allows data-mining with incomplete dataset	Limited knowledge is obtained	Degrees of freedom are hard to calculate	N&C	VL

Note: N = numerical; C = categorical; VL = very low.

were determined to be extremely biased, where type I error rates were significantly increased.

The term 'hot deck' was derived from computer punch card terminology. The analogy to imputation is that an imputed value comes from other 'cards' in the deck being processed by the computer. In this case, cards were physically 'hot' after the computer processed the cards in the deck. In contrast, cold-deck imputations are similar but are used much less in practice. In this type of imputation, the imputation data come from a donor of other previously collected databases or other decks of cards that were not currently being processed ('cold'). There are several variations of hot-deck methods found in the literature (Hu *et al.* 2007). The basic idea of a hot deck uses complete samples to determine an imputation value. Some hot-deck variants assume that all imputes must have equal probability distributions as the non-missing counterpart (Fuller and Kwang 2001). Other methods implement a weighting system to choose imputes from donors (Kim and Fuller 2004). Hot deck has successfully been used in large datasets; however, when the same methods were used in smaller datasets, the quality of estimates became poor and wildly increased sample variance (Kaiser 1983). One desirable property of hot deck is that imputation values come from an observed value, which makes it an attractive method for instances with discrete values. Hot deck uses data-matching techniques, which are often not strongly based on theory (Ford 1983). Problems of poor matching inputs occur with higher degrees of missing values (Reams and McCollum 1999). For example, one study investigating intensive care patients found extreme biases when using hot-deck imputation in comparison with other statistical imputation methods (Perez *et al.* 2002). These researchers validated their claims by using receiver operating characteristic (ROC) curve and promoted the use of multiple imputation (MI) for future clinical studies, which will be discussed in a later section. Thus, with inconsistent results and a method that is not strongly based on theory, LOCF is overlooked for other methods.

Random assignment involves sampling all of the existing values for a missing attribute. In this method, an impute value is randomly drawn from the missing attribute's distribution. This method is considered non-statistical, because the impute value is not precisely calculated, but rather randomly picked based on a determined distribution. Utilising the random assignment method can decrease the level of precision of data-mining methods by creating infeasible or inconsistent sample instances. Thus, if a dataset has significant missing values, this method can reduce or drastically distort the associations with other attributes in the dataset, since it does not consider other attributes when drawing imputes. Increasing the variance can be viewed as either a limitation or a benefit. Some missing value methods, which are discussed later, artificially reduce the variance, which is considered severe limitation. The problem with the random assignment method is that the degree of variance that is added is often very large (Ziarko and Yao 2001).

Table 2 shows a summary of pre-replacement non-statistical methods with N and C representing numerical or categorical. In addition, the computational cost is labelled as very low, low, medium, high or very high.

3.3. *Pre-replacement statistical methods*

With small sample sizes, where the disregarded method is heavily discouraged, methods exist that attempt to fill in or replace missing values in order to create a complete dataset for the data-mining phase. The following section is dedicated to methods that attempt to replace the disregarded methods based on a statistic calculated by the remaining

Table 2. Pre-replacement non-statistical missing value methodologies.

Method Name	Primary Benefit	Secondary Benefit	Primary Limitation	Secondary Limitation	Missing Variables	Computational Cost
Last Observation Carried Forward	Valid with variables that do not change significantly over time	Easy to implement	Does not maintain associations between attributes	Can cause infeasible sample instances	N&C	VL
Sequential Hot-Deck	Variance is not artificially reduced	Imputes come from observed values	Search not based on strong theory	Can cause infeasible sample instances	N&C	VL
Random Hot-Deck	Variance is not artificially reduced	Does not require distribution assumptions	Search not based on strong theory	Can cause infeasible sample instances	N&C	VL
Nearest Complete Hot-Deck	Variance is not artificially reduced	Non-parametric technique	All possible combinations need to be computed	Poor imputes with high degree of missing values	N&C	VL
Random Assignment	Variance is not artificially reduced	Imputes come from observed values	Can cause infeasible sample instances	Does not maintain associations between attributes	N&C	VL

Note: N = numerical; C = categorical; VL = very low.

non-missing values for a given attribute, such as mean-mode, cluster-based mean-mode and multiple linear regression.

The mean-mode method assigns numerical missing values with the attribute's mean and replaces a categorical missing value with the most frequently occurring value (Han and Kamber 2001). This method could be slightly modified by replacing missing values with sample medians or even quartiles instead of means or modes. This method is the most common method used in research (Batista and Monard 2003), which is not a disregard method. However, many statisticians discount this method entirely because of the serious ramifications of artificially reducing the dataset's standard deviation. Even when the data is MCAR, the standard deviation is lowered by imputing numerous instances that lie on the central tendency. Thus, the importance of an attribute's value that lies on the tails of a distribution can be diminished. As a result, rule-based data-mining techniques suffer greatly in their contingent predicting power when the mean-mode method is used (Matteo 2004). The mean-mode method can be a reasonable method for missing data when the chance of missing is very low. For example, researchers investigating work-related musculoskeletal disorders opted for the mean-mode approach because their collected samples had a very low likelihood of missing values occurring, which was less than 5% (Campo *et al*. 2008). This decision was supported by independent research, which suggests that employing other imputation strategies would have resulted in similar results when the chance of missing is small (Tabachnick and Fidell 2007). However, even with these successes, researchers claim that the mean-mode method is the worst possible option, regardless of the amount missing, because of it artificial minimises the dataset's standard deviation (Acock 2005).

Cluster-based methods attempt to group similar data instances together prior to imputing a missing value. Once the sets are grouped, either means or modes are computed for each cluster and imputed for the missing value. Three variants of the mean-mode algorithm have been proposed (Fujikawa and Ho 2002). These methods include the natural cluster-based mean-mode (NCBMM), attribute rank cluster-based mean-mode (RCBMM) and *k*-means cluster-based mean-mode algorithms (KMCMM). The NCBMM uses class attributes to divide the samples into 'natural' clusters. Thus, once samples are subdivided, cluster mean and modes are used for imputation. The RCBMM algorithm investigates a single attribute one at a time until all attributes contain no missing values. The distance from the categorical attribute containing a missing value is calculated from all other categorical attributes in the dataset. Each attribute is ranked based on the distance found, where the attribute that has the smallest distance is then used for clustering. A cluster mode is determined and imputed into the missing value instances. KMCMM (Fujikawa and Ho 2002) can impute values for numerical attributes and can be used independent of whether there is a dependent attribute in the dataset. The procedure used for the *k*-means clustering procedure randomly selects *k* number of objects. Thus, the cluster mean is based on a single point estimate. The algorithm processes the remaining samples by assigning them to a cluster *k* by determining which is the most, based around the cluster mean. The cluster mean is recalculated and the process is iterated until all samples are clustered. One research study collected 15 very large datasets and compared the three cluster-based mean-mode methods to six other missing value methods (Fujikawa and Ho 2002). In this research, it was found that the NCBMM and RCBMM achieved consistently about the same rate of error. More importantly, when applicable, all three cluster-based mean-mode methods outperformed other more complex missing value methods, such as linear regression, auto-associative neural networks (AA-NNs) and decision-tree imputation, in a much less computationally expensive manner. However, one can speculate that the cluster-based methods might artificially reduce a dataset's variation under non-ideal circumstances. For example, if NCBMM is used and the dependent categorical value is heavily imbalanced, an artificial reduction in variation might occur. The same issue can arise for RCBMM, when the highest-ranking attribute has limited categorical values, or if k is small when using KMCMM. However, even with the early success with large datasets, cluster-based methods have not been used greatly in academic research. Thus, it is hard to assess the capability of these methods without further research.

Unlike the previous methods, linear and multiple linear regression models attempt to use the information from other attributes for the imputation. A dataset can be partitioned to derive a linear regression equation to determine the missing value regardless if the variable is a predictor or a response (Pyle 1999). The linear regression process minimises a sum of squared error from the observed and the predicted. Thus, after the linear estimation has been formed, it can be used to estimate the missing value based on the relationship between the two correlated values. As a result, the probability of creating infeasible samples is reduced.

Linear regression inherently depends on the association of a linear relationship between the missing attribute and another correlated attribute. A problem with this technique is that if the relationships among the missing and correlated values are not linear, a poor value can be imputed into the missing case. Multiple linear regression somewhat overcomes the strict linear relationship enforced by the single attribute to another. However, if a dataset has values that are MCAR, the estimates might be derived from a small sample size, since it becomes increasingly difficult to find complete samples. Another drawback with either linear or multiple linear regression is that the estimated imputes are generally better behaved than the values would be if they were not missing. This happens because the missing values are predicted from other attributes. Therefore, the imputes more strongly correlate with other attributes than they would be if they were not missing (Wayman 2003). Adding a stronger correlation to other attributes is not desired for many data-mining applications. For example, step-wise regression imputation strategy was used in a simulated case–control study when a large portion of the data was missing from 1000 participants (Viallefont *et al.* 2001). The results greatly overstated the strength of predictors for cervical cancer indicators, where only 49% of the predictors that were significant were deemed to be true risk factors by experts.

Table 3 shows a summary of pre-replacement statistical methods with N and C representing numerical or categorical. In addition, the computational cost is labelled as very low, low, medium, high or very high.

3.4. *Pre-replacement machine learning methods*

The following section is dedicated to methods that are considered more powerful than previously discussed missing value methods because they tend to utilise random number assignments and iterative methods in order to achieve a certain goal of the computer method. The machine learning pre-replacement methods include replacement under same standard deviation, k-nearest neighbour, expectation maximisation (EM), artificial neural network (ANN), AA-NN and MI. These methods are also more computationally intensive that the previous methods discussed and are rarely seen in ergonomic-related research. Further discussions or relevant findings in the ergonomics studies are limited due to the sparse use of these methods; however, they are included in this review to advance the

Table 3. Pre-replacement statistical missing value methodologies.

Method Name	Primary Benefit	Secondary Benefit	Primary Limitation	Secondary Limitation	Missing Variables	Computational Cost
Mean-Mode	Valid with very low occurs of missing values	Easy to implement	Variance is artificially reduced	Does not maintain associations between attributes	N&C	L
NCBMM	Option for extremely large datasets	Easy to implement	Dataset is split on dependent categorical	Cluster size is vital to not reducing variance artificially	N&C	L
RCBMM	Option for extremely large datasets	Easy to implement	All possible combinations need to be computed	Cluster size is vital to not reducing variance artificially	C	M
KMCMM	Option for extremely large datasets	Uses all available information	Cluster size is vital to not reducing variance artificially	Poor imputes with high degree of missing values	N	M
Linear Regression	Uses limited information for impute	Infeasible sample instances is reduced	Model produces imputes that are too well behaved	Relies on linear relationship	N	M
Logistic Regression	Uses limited information for impute	Infeasible sample instances is reduced	Model produces imputes that are too well behaved	Limited to molding two classifications	C	M
Multiple-Linear Regression	Uses all information for impute	Infeasible sample instances are reduced	Model produces imputes that are too well behaved	Model based on small sample sizes if numerous missing values	N	M

Note: NCBMM = natural cluster-based mean-mode; RCBMM = rank cluster-based mean-mode; KMCMM = *k* means cluster mean-mode; N = numerical; C = categorical; L = low; M = medium.

field's knowledge of missing value methodologies. Thus, slightly more detail is provided on specified details of the methodology throughout the remainder of this review.

3.4.1. *Statistically inspired methods*

Replacement under same standard deviation is a pre-replacement method that assigns values randomly to missing values in such a way that imputes will not disturb an attribute's standard deviation (Pyle 1999). This process is iterative, where each missing value is assigned with a value and then validated such that the imputation did not disrupt the attribute's standard deviation. When such a value is found, the next missing value is investigated. This method attempts to control the attribute's variance; however, it is not well explored in literature. One goal of this method is to avoid artificially reducing the standard error of an attribute. However, the method does not maintain associations among other variables. Thus, applying this method might distort rules that can be mined since the variance between variables is also not controlled. This method also relies on the assumption that missing values would not alter the theoretically complete dataset's variation. This assumption might not be valid, especially in surveys where data could be NMAR.

The *k*-nearest neighbour is an extension of the nearest neighbour procedure. Similar to the nearest neighbour method, *k*-nearest neighbour can be used to predict qualitative and quantitative missing values, as well as a strategy to handle multiple missing values within a sample instance. *k*-Nearest neighbour is a simple form of machine learning, where objects are classified by a majority among *k* neighbours, where *k* is a small positive integer. For this implementation, vectors in multidimensional feature space represent sample instances. The locations of the vectors are used in such a way that a space is assigned to a class when it is the majority class in the partition. Several methods exist to determine the distance (d_{ij}) between sample sets, which include Euclidean, Manhattan, Maholanobis, correlation, cosine and squared-chord, (Yenduri 2005) with the Euclidean measure being the most popular (Fujikawa 2001). Suppose the nearest neighbour method is used on a dataset that has p attributes and k missing values. If k is very small in a sample instance, finding a distance that is small in magnitude is more likely to occur and thus a better value can be assigned to the missing attribute. In general, the nearest neighbour method becomes a less viable option as k becomes large; in addition, if k is too small, the standard deviation can artificially be reduced. For this reason, many experiments should be performed in order to determine k as well as the distance metric. Due to this, evolutionary algorithms are often used to determine a value for k intelligently (Frédéric and Serge 2007). The conclusion of one extensive research investigating low and high amounts of missing datasets (Batista and Monard 2003) found that *k*-nearest neighbour performed consistently better than embedded methods such as C4.5 (Quinlan 1986) and CN2 (Clark and Niblett 1989), which will be discussed later.

3.4.2. *Expectation maximisation*

Previous missing data methodologies did nothing to rectify the problem of artificially reducing the dataset's standard deviation. This should be avoided, since missing values methodologies should not add information to the dataset (Schafer 1997). EM is the first of this survey that addresses this problem. EM uses available information within the dataset to preserve the relationship in the entire sample (Dempster *et al.* 1977). EM is used to estimate the parameters of the probability density function of an incomplete sample by maximising the likelihood of the available data. The procedure assumes that a dataset forms to a given distribution and determines the parameters of the distribution that an observed sample would most likely have been produced. Imputes are assigned until successive iterations are sufficiently similar. This process generally converges quickly to a covariance matrix that is very similar to the prior attempt. For this reason, the characteristics of the variables and the variation of the dataset itself are preserved.

The EM algorithm is iterative and alternates between two steps called the E-step (expectation) and M-step (maximisation). The E-step calculates the conditional expectation of the complete data likelihood given the observed data and estimates. Given complete data likelihood, the M-step finds the parameter estimates to maximise the complete dataset from the E-step. EM is very difficult to implement, requires an assumed statistical model for each analysis and requires bootstrapping in order to determine standard errors (Wayman 2003). However, several software packages are available as either freeware or commercially based to implement the EM imputation algorithm.

EM is an iterative algorithm that is computationally intensive before converging to a solution. Inherently, this method also relies heavily on assuming the correct distribution and often works well for values that are MCAR. Unfortunately, knowing the probability density function before the data-mining procedure is often very difficult. However, if the missing values in the dataset are MCAR, EM can produce unbiased results. In a longitudinal study investigating the repercussion of body weight, researchers noted no meaningful differences were obtained by using the computationally more expensive EM algorithm to the list-wise deletion method (Keel *et al.* 2007).

3.4.3. *Artificial neural network inspired methods*

ANNs, which are similar to multiple linear or non-linear regressions, provide another method, but with more accurate imputations (Gupta and Lam 1996). ANNs are mathematical models, which are inspired by biological neural networks. The mathematical goal of ANNs is to map input space by changing weights of interconnected elements in the network in order to map the output space. ANNs are practically used as non-linear statistical data modelling tools and are adaptive in nature since the internal structure of the model changes based on internal or external information that flows through the network during the learning phase. ANNs also have the benefit of being robust to noisy data. ANNs have been shown to recover missing values very efficiently in temporal datasets (Gorban *et al.* 2002). ANNs capture non-linear trends, which accounts for variable interaction more so than other linear-based models. The ability of the ANN to capture these trends makes imputation more accurate than even complex interpolation procedures (Fariñas and Pedreira 2002).

ANNs have been well documented (Haykin 1999). There are two main learning paradigms used for creating neural networks, which include supervised and unsupervised learning. ANN methods that use supervised learning, such as multi-layered perceptron (MLP), are often said to use a 'teacher' in the training process, because both the inputs and outputs are used to update the internal weights of the ANN. A popular unsupervised learning technique is self-organising maps (SOMs). This technique differs because it does not use the output of a dataset to modify the internal weights of the ANN structure. This technique is referred to as training 'without a teacher', where only the inputs are used to update the model.

MLPs can be created to learn missing value relationships, after data partition with complete records is performed. Thus, similar to multiple-linear regression, the non-missing values are used to predict the attributes with missing values. Once individual networks have been trained for each attribute that is missing, the sample instances containing missing values can be used to predict the missing attributes. This form of a MLP for the missing value problem has been referred to as the pre-processing perceptron (Westin 2002). Although there were few examples that were found using ANNs in ergonomic studies involving imputation, one particular study found them useful in designing complex surveys (Amer 2006). In this novel study, ANNs were compared with another imputation strategy, MI, which will be described later in this section, to incorporate sampling weights from an ANN to account for the uncertainty of the imputation value as well as the survey design. This study found ANNs to require fewer resources than the MI approach and accounted for more of the variation in the complex survey that was being analysed.

SOMs are an unsupervised learning ANN that are often used to produce low-dimensional representations from high-dimensional input space. SOMs are useful for visualising datasets and are often called Kohonen maps (Kohonen 1995). The goal of training is motivated by how visual, auditory and other sensory information is handled by the brain, which consists of reducing high-dimensional space into a lower dimension. With a SOM, training causes different parts of the ANN to respond similarly to certain input patterns. A SOM is associated with two steps: training and mapping. The training of a SOM has a competitive strategy using vector quantisation. SOMs are calibrated to perform classification based on observations without previous class assignments. For this ANN structure, the clustering process takes into account neighbouring relations between nodes of the ANN map. The weights are adjusted in such a way that the SOM lattice is adjusted towards the input vector (Haykin 1999). This process is repeated so that the network ends up associating groups or patterns in the input data to output nodes. The second step is called mapping, where input vectors are classified by the winning neuron or the neuron in which the weight vector of the neuron is closest to the input vector. When an example is induced into the map, the distance, which is often Euclidean, to all of the weight vectors is computed. The neuron with the weight vector that is most similar to the input example is called the best matching unit. Thus, the classification process is performed with regard to the proximity of relations between classes.

SOMs can be used to impute classification attributes after a map is trained using non-missing cases. Following training, all sample instances without missing values are assigned with a best matching unit. A majority rule (or mode) is then used to determine each unit's classification, which is based on the classification type of the original sample instances. Thus, when a sample instance is given to the SOM with incomplete records, a best matching unit is determined and a classification impute is assigned to the missing value. An experiment that utilised MLPs and SOMs on a very large dataset with missing values concluded that each of the methods performed similarly, where an accuracy interval of ±3% was achieved (Westin 2002). Other studies resulted in similar conclusions, where SOM outperformed other ANN methodologies by trivial margins (Piela 2003). ANNs generally require a sufficient amount of data for training to make better generalisations. This is often a problem in survey data, where sizeable datasets are often not obtained. ANNs are a viable option over other standard statistical modelling techniques, but are far from maturity in automated systems (Fessant and Midenet 1999). This method is similar to multiple linear regression, but does not have the limitation of relying on linear relationships since neural networks, with more than a single hidden layer, are non-linear. Since ANNs do not require any distribution assumptions for the model, they are more robust than other methods that require an assumed distribution. For these reasons, an ANN imputation method is preferred over other linear or multiple-linear methods.

An AA-NN is another form of missing value methodology. AA-NN differs from prior neural network methods. Previously, the neural network was trained to learn how a set of independent attributes maps a dependent variable. AA-NNs are used to learn efficient encoding of input space. The goal of an AA-NN is to compress the input representation. There are three layers in an AA-NN. The first and last layers of the AA-NN have the same number of processing elements. The second layer contains significantly less processing elements than the other two. The goal of an AA-NN is to encode input space at the second layer, so that the original input can be reproduced in the final layer with lower dimensionality. In a sense, the independent attributes are the same as the dependent ones. It should be noted that if linear units are used in the network, it is very similar to principle component analysis (Hinton and Salakhutdinov 2006). To impute for a missing value, optimising software is needed, which slightly increases the method's complexity. However, many practitioners in the ergonomics field are quite used to optimisations and their applications, which may make the AA-NN approach attractive.

Research has been conducted (Nelwamondo *et al.* 2007) that compares the abilities of AA-NN and EM missing value methods using datasets of an industrial power plant, an industrial winding process and HIV seroprevalence survey data. The results of this extensive testing show that the EM algorithm is better suited in situations where there are few correlations between the input variables as seen in the survey dataset. However, in this study, the AA-NN outperformed the EM algorithm when there were non-linear relationships between the input variables. AA-NNs seem to be robust to distribution assumptions and can handle categorical or numeric missing values simultaneously; however, limited research was found on the methodology.

3.4.4. *Multiple imputation methods*

Methods that have been discussed previous rely on single imputation. Single imputation methods are only applicable when the proportions of missing values are small, which is suggested to be less than 5% (Schafer 1997). These procedures do not correctly represent the uncertainty of the imputed values to the dataset. Thus, they often underestimate standard errors and overestimate the level of precision that can be obtained from the data. This bias can distort confidence intervals and statistical tests. Even complex single imputation methods are said to cause more problems than they solve, which is unacceptable (Graham *et al.* 1997). Addressing incomplete data inadequately leads to biased analysis and incorrect inferences.

MI is a process where each missing value is replaced by multiple different values that are drawn from an implicit model derived from all non-missing values. MI produces multiple sets of complete datasets (all having varying degrees of randomness) to account for uncertainties. The method not only attempts to restore the natural variability in the missing values, but it also incorporates the uncertainty from estimates formed to predict the missing value. Maintaining associations of the dataset is preserved by creating imputed values that are from attributes correlated with the missing value attribute, while variability between imputed datasets is accounted for by observing the error formed by the missing value estimation.

Varieties of MI methods have been explored, which can handle both continuous and discrete variables. Methods have been proposed, such as propensity (Rosenbaum and Rubin 1983), predictive mean matching (Little 1988), ANNs (Richman *et al.* 2007), regression trees (He 2006) for continuous variables and discriminate analysis, logistic regression (Williams *et al.* 2005) and classification trees (Feldesman 2002) for discrete variables. Markov Chain Monte Carlo (Schunk 2007) is an alternative approach for complicated non-responses. MI can be performed on many different types of datasets, which in itself is an appealing property. Software is now becoming readily available for MI, which should assist researchers with varying degrees of statistical knowledge and increase the usage of imputation strategies since they are often time-consuming to implement.

MI strategies are more efficient than EM, because it does not require lengthy simulation and normality assumptions (Rubin 1996). Unlike other single imputation methods, MI has been shown to produce unbiased estimates for parameters, which reflects the uncertainty associated with estimating missing values (Schafer and Graham 2002). Accounting for this uncertainty is what separates this method from other ‘unaccepted’ methods, which were described earlier. MI, which often uses multiple linear regression, produces better missing value imputations than even more complex non-linear single imputation modes. It also has shown to be an appropriate method when sample sizes are low, even with high rates of missing values (Wayman 2003).

The ill-effects of using MI are minimal when the data are MCAR or MAR and thus provides a robust solution to many missing value problems (Graham *et al.* 1997). However, even with the advancements made with MI, several researchers are using other less than optimal strategies. Some have speculated that there is a lack of utilisation of MIs because of unfamiliarity with the technique (Wayman 2003). Recently, however, there has been a rapid increase of literature being published and commercialised software implementing MIs. MI methods are flexible and produce superior imputations for a wide range of datasets. MI performs well when there is a high degree of non-responses in survey data where the patterns of missing values are complex (Rubin 1996).

There is no standard number of imputed datasets to construct before aggregation. However, an estimation of the expected efficiencies that can be obtained through MIs is shown in Equation (1) (Rubin 1987). In this equation, m represents the number of imputations and γ represents the rate of missing information for the quantity being estimated. Thus, the information that can be obtained due to the missing value quickly diminishes after just a few imputations.

$$\text{Imputation Efficiency} = \left(1 + \frac{\gamma}{m}\right)^{-1} \tag{1}$$

Table 4 shows examples of expected efficiency with various values of m and γ. This table shows that there is very little advantage of having 10 or more imputations unless the error rate is extremely high.

A simulation was performed independently to verify the expected efficiency when missing values occurred at MAR and the probability of missing values within a sample instance was 30%. The results of expected efficiency were validated through the

Table 4. Multiple imputation efficiencies.

M	$\gamma = 0.1$	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.9$
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Note: M = number of imputed datasets.

simulation, where 94% efficiency was obtained after five multiply imputed datasets were aggregated (Schafer 1997). When the same experiment was repeated with 50% chance of missing values within a sample instance, 10 imputations were needed to achieve the same efficiencies. Likewise, datasets with MAR missing values produce a standard error that is just 2% larger than an infinite number of imputations even when the missing data probability is 40% (Allison 2002). From these two independent studies, it appears that in most missing value instances, only three to 10 imputations are needed to achieve a sufficiently high quality of imputation estimates.

For multiple imputations, diagnostic measures are formed to indicate how strongly the quantity being estimated is influenced by missing values. This estimation is defined in Equation (2), where r represents the relative increase in variance due to a non-response (Rubin 1987). For more information regarding the theory and implementation of MI, readers should refer to Harel and Zhou's (2007) review of MI.

$$\gamma = \frac{\left(\frac{r+2}{df+3}\right)}{r+1} \tag{2}$$

where, $r = \frac{(1+m^{-1})}{\bar{U}}B$.

Table 5 shows a summary of pre-replacement machine learning methods with N and C representing numerical or categorical. In addition, the computational cost is labelled as very low, low, medium, high or very high.

3.5. *Embedded machine learning methods*

The following section is dedicated to methods that handle missing data internally. These methods include C4.5, CandRT, robust association rules, extended ANNs and enhanced ANNs.

3.5.1. *Decision tree-based methods*

Decision trees can be helpful to ergonomists to better visualise and comprehend classifications stemming from surveys (Liu and Salvendy 2007). C4.5, a tree-based methodology, is a well-known approach used when dealing with incomplete supervised classification problems. It has proven to be a very effective solution over other techniques (Grzymala-Busse and Hu 2000, Ziarko and Yao 2001). C4.5 is an extension of ID3, which is another tree induction method for classification problems. There are two enhancements made to the original ID3. The first limitation occurs when two or more cases with identical values for attributes belong to different classes. The second is a procedure known as pruning to reduce what statisticians call over-fitting. The C4.5 algorithm can only predict non-continuous, or classification based, dependent variables. However, C4.5 is not capable of handling a situation where a missing value occurs on dependent attributes. In addition, C4.5 only accounts for missing attributes that are continuous values. It uses the knowledge of the computed probabilities to create a complete data table. Therefore, other methods in data mining cannot be employed directly after completion of the algorithm. Some researchers have stated that C4.5's internal method of handling missing values is only appropriate when they are insensitive because the values are too small or too large to be measured. Perhaps the most important benefit of decision tree induction is transparency, which makes decision trees simple to understand.

Table 5. Pre-replacement machine learning missing value methodologies.

Method Name	Primary Benefit	Secondary Benefit	Primary Limitation	Secondary Limitation	Missing Variables	Computational Cost
Replacement under Same Deviation	Avoids artificially reducing variation	Uses available information to determine distribution	Does not maintain associations between attributes	Assumes variation is the same as observed completed samples	N	M
Nearest Neighbour	Uses all information for impute	Imputes come from observed values	Importance of categorical values is reduced in distance calculation	Poor imputes if sample has many missing values	N&C	M
k-Nearest Neighbour	Does not generally disrupt distribution of data	Imputes come from observed values	Cluster size is vital for variance inference	Importance of categorical values is reduced in distance calculation	N&C	H
EM	Completed dataset in one step	Adapts to any assumed distribution	Requires an assumed distribution	Is inferior to non-linear models	N	VH
MLP	Maintains complex non-linear relationships	Lower uncertainty for impute	Requires large training sets of data	Far from an automated system	N&C	VH

SVM	Maintains complex non-linear relationships	Lower uncertainty for impute	Requires large training sets of data	Far from an automated system	N&C	VH
SOM	Maintains complex non-linear relationships	Robust to small datasets	Interment knowledge of ANN is required	Far from an automated system	C	VH
Auto-associative Neural Network	Maintains complex non-linear relationships	Do not need dependent variables	Requires optimisation software	Requires large training sets of data	N&C	H
Reversed Engineered ANN	Maintains complex non-linear relationships	No distribution assumptions	Requires optimisation software	Problems with determining insensitive values	N&C	VH
Multiple Imputation	Unbiased imputes	Robust to quality of model	Requires MAR assumption	No standardised number of imputed datasets	N&C	H

Note: EM = expectation maximisation; MLP = multi-layered perceptron; SVM = support vector machine; SOM = self-organising map; ANN = artificial neural network; MAR = missing at random; N = numerical; C = categorical; M = medium; H = high; VH = very high.

CandRT has two methods for dealing with missing values internally. If there are dependent missing values, they are ignored in the induction. They are also ignored if all of the attributes are missing. However, if an independent value is missing, the surrogate split is used to determine to which child node a case should belong. If ties occur in this situation, the majority rule is used (Breiman *et al.* 1984). Decision trees can be used as another form of imputation of missing values. The goal of this type of imputation is to assign missing values based on a specified decision tree. In this method, the original dataset is divided into two sets, where one contains samples with non-missing values and the other with missing value(s). A decision tree is constructed from the dataset containing non-missing values for an attribute in the dataset with missing values. Once the tree is constructed, an estimation of missing values can be assigned, which is based on the position of the tree. The task of constructing a tree is handled by a recursive-partitioning algorithm, which, at each non-terminal node, determines a value for an attribute, which best discriminates against, or branches, the remaining filtered classes.

One downfall of decision trees is that they are sensitive to the number of samples used to make splits in the nodes; therefore, created trees are often unstable (Timofeev 2004). CandRT also only provides a binary split, which could pose unnecessary complexity in the tree structure. For this reason, decision trees, such as CHAID, are performed because they allow more than just binary splits per node (Kass 1980). The main drawback from this method lies within data partitioning. If datasets are MCAR, it can become troublesome to find sufficient samples to create a decision tree for imputation purposes. Several trees might need to be created due to the randomness in which missing values occur within the dataset. However, since embedded tree-based methods handle missing values internally, researchers do not often reflect on how the missing values could be biasing the results of their study.

3.5.2. *Association rule learning*

Association rule learning is often used as a text data-mining technique, which can be used as a missing value treatment method. It should be noted that few text data-mining methods exist, which make association rules an attractive imputation methodology when other imputation methods that use categorical or continuous values cannot be used. Associated learning is used to discover patterns of multiple independent text elements to discover rules about the dependent text event. This analysis method has origins in purchasing transitions, where the analysis is called ‘market basket analysis’. The goal of association rule learning, as with other data-mining techniques, is to reduce complexity about a dataset and present the relationships found in an easily comprehensible form.

In the case of missing data, the goal is to satisfy a user specified minimum support and minimum confidence level for imputation after a rule has been created. One methodology that uses association rules to fill missing values is called the missing values completion (MVC) method (Ragel and Cremilleux 1999). This method incorporates the robust association rule algorithm (Ragel and Cremilleux 1998), which attempts to create subsets of data for the rules to fill in missing values. The method attempts to impute values with rules that have high confidence. If a complete dataset cannot be constructed after the first iteration, a lower confidence rule will be used in the next scan of the incomplete dataset until a complete dataset is constructed. MVC is robust, because a complete dataset can always be created from low support and low confidence rules and it is easy to understand (Crémilleux *et al.* 2005). The method has a benefit of imputing multiple missing text values and is efficient when dealing with large datasets. It has been shown that the MVC method can reduce the imputation error when compared to C4.5 by as much as half (Ragel and Cremilleux 1999).

Although association rules are primarily used for text mining, it has been used in Likert type surveys to explain the behaviour of why missing values are occurring (Wang and Wang 2009). This is a benefit, where few methodologies are presented in academic research that is concerned with better understanding the non-response.

3.5.3. *Extended and enhanced artificial neural network methods*

Extended networks give a systematic structure of training incomplete datasets, which also can be completed in the data pre-processing stage. The benefit of the extended ANN is that it is straightforward to implement and is less computationally intensive than other missing value methods such as EM, while also being more accurate. These methods allow for all of the data to be used in the ANN. However, research shows that expanded networks in general have poor generalising abilities in a small complex dataset (Vamplew 1996). Generalisation issues are compounded when multiple missing values occur within the dataset. Thus, networks that attempt to estimate missing values generally outperform substitution or extended networks (Vamplew *et al.* 1996).

An enhanced SOM, also known as the Kohonen competitive learning strategy, groups objects in such a way that the similarity between the objects allocated to the same cluster is maximised. The learning strategy used is often called 'winner-take-all', where objects are fed into the network multiple times during training until weights are appropriately adjusted in such a way that inputs are associated with a certain output node or, in this case, a cluster. SOM training can be altered to incorporate missing values directly in such a way that when an observation with missing values is presented to the network, the missing values are ignored when distances are calculated between input space and output nodes. The available dimensions are used to update the weights of the network. Because these maps are built with incomplete data, they are sometimes referred to as fuzzy-maps (Wang 2003). Interestingly, researchers have indicated that the performances of SOM benefit from being trained with incomplete samples (Samad and Harp 1992). In this case, it was noted that the benefit was greater with higher dimensional problems over lower dimensional ones.

A method called learning associations by self-organisation (LASSO) avoids sequentially imputing missing values with different models (Midenet and Grumbach 1994). Instead, this method creates a single model that is trained with both complete and incomplete sample instances for a single imputation model. A result of a study using LASSO found comparable imputation results to other imputation models, such as hot deck, and was only inferior to other methods such as MLP imputation (Fessant and Midenet 1999). One benefit of a compromise map is that it requires substantially less observations for network training than other machine learning methods such as MLPs. Not only can MLP be used for imputation, it can also be used for determining erroneous data, can capture attribute non-linearity and is a useful tool for data visualising.

Table 6 shows a summary of embedded missing value methods with N and C representing numerical or categorical. In addition, the computational cost is labelled as very low, low, medium, high or very high.

4. Conclusions

There are numerous methods of managing missing values in datasets for ergonomic studies. From a recent review of literature, it appears that the research community for

Table 6. Embedded missing value methodologies.

Method Name	Primary Benefit	Secondary Benefit	Primary Limitation	Secondary Limitation	Missing Variables	Computational Cost
C4.5	Uses all information for tree construction	Final transparent model	Does not offer a complete data table	Many trees need to be constructed if used for testing	C	VH
C&RT	Uses all information for tree construction	Final transparent model	Does not offer a complete data table	Many trees need to be constructed if used for testing	N&C	VH
Robust Association Rules	One of the few options for text imputations	Can be used to help explain why missing values are occurring	Imputes from low support can artificially reduce variance	No standard for poorly supported support rules	C	M
Flagged Ex-ANN	Uses all information for ANN construction	Maintains complex non-linear relationships	Poor generalisation in complex datasets	Requires large training sets of data	N&C	H
High-Low Ex-ANN	Uses all information for ANN construction	Maintains complex non-linear relationships	Poor generalisation in complex datasets	Requires large training sets of data	N&C	H
Shadow Ex-ANN	Uses all information for ANN construction	Maintains complex non-linear relationships	Poor generalisation in complex datasets	Requires large training sets of data	N&C	H
SOM En-ANN	A single model is required for imputation	Not a lot of data are needed	Often only used for data visualisation	Size of map must be assumed	C	H

Note: ANN = artificial neural network; Ex-ANN = extended ANN; En-ANN = enhanced ANN; N = numerical; C = categorical; VH = very high; M = medium; H = high.

ergonomic studies has not fully adopted methods that might benefit their research. Some of the methods that appear in ergonomic studies are considered statistically unacceptable (Schafer and Graham 2002). For cases where the amount of missing is great, single imputation methods can artificially reduce the variance within an attribute and can diminish the predicting power that it might possess if it were not missing. Single imputation methods can also impose false relationships within the dataset, which bias the inference that researchers are disseminating in their work because they do not account for the variation of the model itself. The consequences of using these methods are especially important for ergonomic researchers to consider. Missing values often appear in ergonomic studies because respondents drop out of longitudinal studies, equipment might be prone to failure or due to a participant's unwillingness to answer questions that appear in surveys. Because more advanced imputation methods are not prevalent in ergonomics research, this paper reviews methods described by other fields of science and engineering to increase the awareness of the method's benefits and limitations to the ergonomics community.

No universal method seems to be superior for a particular dataset type problem. Even if one methodology works well with one type of dataset, the results often cannot be repeated on similar datasets. This is due to the underlying distributions in the datasets, correlations of attributes, the amount of missing values and the sample size. Methodologies can create biases with imputed values if the correct underlying behaviour of the dataset is not known and applied. Understanding the relationships needed to create a superior imputation method is not a luxury when missing values are present. A researcher in the missing value problem stated that: 'the only really good solution to the missing data problem is not to have any' (Allison 2002). However, others believe that missing rates of less than 1% are generally considered trivial to deal with and 1–5% is manageable. As missing values increase to 5–15%, methods that are more sophisticated are required to handle the downfalls of single imputation methods. Once the rate of missing goes beyond 15%, some suggest that it is unlikely that any imputation methods can lead to any kind of meaningful interpretation (Pyle 1999). Due to these problems outlined, it is clear that more work is needed to advance all fields of scientific research (Harel and Zhou 2007).

Limited studies were found that discussed the statistical power value gained by using an imputation method. Thus, more research is needed to derive a set of rules that could be used to define which imputation method should be better suited for a given dataset problem. One constant conclusion resulting from missing-data research indicates that if data are MCAR and do not contain a substantial amount of missing values, which is often lower than 10% (Scheffer 2002), any suitable treatment method can be applied without introducing substantial bias to the data (Batista and Monard 2003). Although research shows that any method could be used with limited amounts of missing data, there are few examples found that evaluate the imputation efficiencies when dealing with large amounts of missing values. Although new diagnostic methods are being developed to test the quality of imputation (Abayomi *et al.* 2008), more research is needed to test when the amount of missing values is too great to expect meaningful information. This is because when the chance of missing increases, imputation methods cause biased results, because imputes are better behaved than if the values were not missing. Many missing value methods assume values are MAR, which is a non-testable assumption (Schafer 1997). However, the MAR assumption is viable when sufficient variables are included in the imputation model (Demirtas and Schafer 2003).

The current state-of-the-art method for replacing missing values is MI, which has offered substantial improvements over a problem that is considered an epidemic in social and physical sciences (Juster and Smith 1997). MI seems to be more robust than single imputation methods to higher degrees of missing values. Some suggest that MI is a viable option for up to 25% chance of missing (Scheffer 2002). The severity of the missing value problem can be seen in the recent development of the number of software titles that implement the MI method. These recent software enhancements include models for all types of missing values, such as numerical, binary, logistical or categorical. Currently, software is being developed that is robust enough to handle a mixture of missing values concurrently, rather than performing the analysis with separate application scripts. The development of software is vital for most researchers to employ more sophisticated imputation schemes, especially when the chance of missing is high. Surprisingly, new MI methods found in literature rely on statistical model procedures rather than machine-learning methods. Generally, machine-learning methods require longer computational times and more data, but achieve greater accuracy than statistical methods, which are easier to implement (Fujikawa 2001).

A major transition has occurred in recent research regarding the missing value problem, where practitioners are utilising more statistically acceptable methods such as MI over primitive methods such as mean-mode or single imputation. From the review of literature, it is apparent that researchers rely on free software such as NORM, CAT, MIX and PAN (Schafer 1999) or from commercial software manufacturers such as SPSS for their missing value needs. Elaborate MI methods are generally not available for most practitioners unless software exists, due to the time and the statistical knowledge required for implementation.

Machine-learning methods, such as ANN, create an explicit model that is able to model higher dimensional space more accurately than standard statistical models. ANNs are able to generalise systems to a high degree of accuracy and are often called universal 'approximators'. Part of the intrigue with ANNs, or other non-linear models, is that prior knowledge of a system being modelled is not needed. Another benefit that ANNs have on datasets is that a pre-assumed distribution is not required, which would reduce errors when sampling posterior distributions to formulate MIs. For this reason, it seems likely that future MI methods will incorporate machine-learning schemes. Even though there has been an increased awareness of machine-learning schemes in non-academic settings, it is far from being accepted methodology. Thus, it seems unlikely that software developers will develop MI that use machine-learning schemes in the near future.

Currently, statistical-based MI is the most attractive option for missing value problems in ergonomics research because it can be used on a variety of dataset types. Although researchers suggest that more research is needed to improve MI methods (Harel and Zhou 2007), it offers several advantages over other methods being used in practice. One of the most important aspects that MI offers is that software is currently being developed on a commercialised scale. This development can aid researchers with varying degrees of statistical knowledge and overcome the cumbersome process of manually implementing the methods. Thus, until significant progress has been made regarding imputation methodology, MI should be heavily considered for datasets with large amounts of missing values.

In summary, this article has provided an overview of the theoretical issues and findings regarding the treatment of missing values in or outside of the ergonomics field. The hope of this comprehensive survey of methods is to motivate researchers to employ better practices in their research by providing a brief overview of the available options.

References

Abayomi, K., Gelman, A., and Levy, M., 2008. Diagnostics for multivariate imputations. *Applied Statistics*, 57 (3), 273–291.

Abel, A.L., Sardone, N.B., and Brock, S., 2005. Simulation in the college classroom: enhancing the survey research methods learning process. *Information Technology, Learning, and Performance Journal*, 23 (2), 39–46.

Acock, A., 2005. Working with missing values. *Journal of Marriage and Family*, 67, 1012–1028.

Allison, P., 2000. Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research*, 28, 301–309.

Allison, P., 2002. *Missing data*. Thousand Oaks, CA: Sage.

Amer, S., 2006. Neural network imputation in complex survey design. *International Journal of Computer Systems Science and Engineering*, 3 (1), 12–17.

Baker, S.G., Ko, C.-W., and Graubard, B.I., 2003. A sensitivity analysis for nonrandomly missing categorical data arising from a national health disability survey. *Biostatistics*, 4 (1), 41–56.

Batista, G. and Monard, M., 2003. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17, 519–533.

Bayeh, A.D. and Smith, M.J., 1999. Effect of physical ergonomics on VDT workers' health: A longitudinal intervention field study in a service organization. *International Journal of Human–Computer Interaction*, 11 (2), 109–135.

Breiman, L., *et al.*, 1984. *Classification and regression trees*. London: Chapman and Hall.

Campion, M., 1988. Interdisciplinary approaches to job design: a constructive replication with extensions. *Journal of Applied Psychology*, 73 (3), 467–481.

Campo, M., *et al.*, 2008. Work-related musculoskeletal disorders in physical therapists: a prospective cohort study with 1-year follow-up. *Physical Therapy*, 88 (5), 608–619.

Carpenter, J., *et al.*, 2004. Last observation carry-forward and last observation analysis. *Statistics in Medicine*, 23, 3241–3244.

Chen, C.-L., Kaber, D.B., and Dempsey, P.G., 2004. Using feedforward neural networks and forward selection of input variables for an ergonomics data classification problem. *Human Factors and Ergonomics in Manufacturing*, 14 (1), 31–49.

Clark, P. and Niblett, T., 1989. The CN2 induction algorithm. *Machine Learning Journal*, 3 (4), 261–283.

Cook, R., Zeng, L., and Yi, G., 2004. Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*, 60 (3), 820–828.

Crémilleux, B., Ragel, A., and Bosson, L., 2005. An interactive and understandable method to treat missing values: Application to a medical data set. *In*: K. Blackmore, T. Bossomaier, S. Foy and D. Thomson, eds. *Studies in computational intelligence*. Vol. 4, Heidelberg: Springer Berlin, 305–314.

Demirtas, H. and Schafer, J., 2003. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22, 2553–2575.

Dempster, A., Laird, N., and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 (1), 1–38.

Edwards, J., Scully, J., and Brtek, M., 2000. The nature and outcomes of work: a replication and extension of interdisciplinary work-design research. *Journal of Applied Psychology*, 85 (6), 860–868.

Fariñas, M. and Pedreira, C., 2002. Missing data interpolation by using local-global neural networks. *Engineering Intelligent Systems for Electrical Engineering and Comunications*, 10 (2), 85–91.

Feldesman, M., 2002. Classification trees as an alternative to linear discriminant analysis. *American Journal of Physical Anthropoloy*, 119, 257–275.

Fessant, F. and Midenet, S., 1999. *A knowledge-based parser: Nerual network based approaches*. Arcueil, France: Technologies for European Surveys of Travel Behaviour.

Ford, B.N., 1983. An overview of hot deck procedures. *In*: W.G. Meadow, I. Olkin and D.B. Rubin, eds. *Incomplete data in sample surveys, Vol II: Theory and annotated bibliography*. New York: Academic Press, 185–206.

Frédéric, R. and Serge, G., 2007. An efficient nearest neighbor classifier. *In*: C. Grosan, A. Abraham and H. Ishibuchi, eds. *Hybrid evolutionary algorithms*. Berlin: Springer Heidelberg, 127–145.

Fujikawa, Y., 2001. *Efficient algorithms for dealing with missing values in knowledge discovery*. Thesis (Masters). Japan Advanced Institute of Science and Technology, School of Knowledge Science.

Fujikawa, Y. and Ho, T.B., 2002. Cluster-based algorithms for filling missing values. *In*: *6th Pacific-Asia conference, knowledge discovery and data mining*, Taiwan, Lecture Notes in Artificial Intelligence, Vol. 2336, Springer, 549–554.

Fuller, W.A. and Kim, J.K., 2005. Hot deck imputation for the response model. *Survey Methodology*, 31, 139–149.

Gorban, A., *et al.*, 2002. Recovering data gaps through neural network methods. *International Journal of Geomagnetism and Aeronomy*, 3 (2), 191–197.

Graham, J.W., *et al.*, 1997. Analysis with missing data in prevention research. *In*: M.W.K. Bryant, ed. *The science of prevention: Methodological advances from alcohol and substance abuse research*. Washington, DC: American Psychological Association, 325–336.

Grzymala-Busse, J. and Hu, M., 2000. A comparison of several approaches to missing attribute values in data mining. *Rough Sets and Current Trends in Computing*, 340–347.

Gupta, A. and Lam, M., 1996. Estimating missing values using neural networks. *The Journal of the Operational Research Society*, 47 (2), 229–238.

Han, J. and Kamber, M., 2001. *Data mining: Concepts and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Han, S.H., *et al.*, 2000. Evaluation of product usability: development and validation of usability dimensions and design elements based on empirical models. *International Journal of Industrial Ergonomics*, 26, 477–488.

Harel, O. and Zhou, X.-H., 2007. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26, 3057–3077.

Haykin, S., 1999. *Neural networks: A comprehensive foundation*. Upper Saddle River, NJ: Prentice Hall.

Haynes, S. and Williams, K., 2008. Impact of seating posture on user comfort and typing performance for people with chronic low back pain. *International Journal of Industrial Ergonomics*, 38, 35–46.

He, Y., 2006. *Missing data imputation for tree-based models*. Los Angeles, CA: University of California, Department of Statistics.

Hinton, G. and Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786), 504–507.

Hu, M., *et al.*, 2007. *A review of imputation procedures across NCES programs*. Washington, DC: Department of Education, National Center for Education Statistics.

Juster, F. and Smith, J., 1997. Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association*, 92 (440), 1268–1278.

Kaiser, J., 1983. The effectiveness of hot-deck procedures in small samples. Meeting of the American Statistical Association. *Journal of American Statistical Association* (http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/45/5a/05.pdf).

Karwowski, W., 2003. Purely editorial: a theory, design and practice of ergonomics. *Theoretical Issues in Ergonomics Science*, 4 (3–4), 259–260.

Kass, G., 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119–127.

Keel, P., *et al.*, 2007. A 20-year longitudinal study of body weight, dieting, and eating disorder symptoms. *Journal of Abnormal Psychology*, 116 (2), 422–432.

Kim, J. and Fuller, W., 2004. Fractional hot deck imputation. *Biometrika*, 91 (3), 559–578.

Kohonen, T., 1995. *Self-organizing maps*. 1st ed. Vol. 30, Springer Series in Information Sciences. Heidelberg: Springer Verlag.

Krause, N., *et al*., 2004. Physical workload, ergonomic problems, and incidence of low back injury: A 7.5-year prospective study of San Francisco transit operators. *American Journal of Industrial Medicine*, 46, 570–585.

Lakshminarayan, K., Harp, S., and Samad, T., 1999. Imputation of missing data in industrial databases. *Applied Intelligence*, 11, 259–275.

Little, R., 1988. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287–296.

Little, R.J.A. and Rubin, D.B., 1987. *Statistical analysis with missing data*. New York: John Wiley and Sons, Inc.

Liu, W., *et al*., 1997. Techniques fordealing with missing values in classification. *Journal of Intelligent Data Analysis*, 1280, 527–536.

Liu, Y. and Salvendy, G., 2007. Visualization support to better comprehend and improve decision tree classification modelling process: a survey and appraisal. *Theoretical Issues in Ergonomics Science*, 8 (1), 63–92.

Mannila, H., 1996. Data mining: machine learning, statistics, and databases. *In*: *8th international conference on scientific and statistical database management (SSDBM '96)*, 2.

Matteo, M., 2004. *Techniques for dealing with missing data in knowledge discovery tasks*. White Paper, Computer Science. Bologna: University of Bologna.

Midenet, S. and Grumbach, A., 1994. Learning associations by self-organization: the LASSO model. *NeuroComputing*, 6, 343–361.

Nelwamondo, F., Mohamed, S., and Marwala, T., 2007. *Missing data: A comparison of neural network and expectation maximisation techniques*. White Paper. University of the Witwatersrand: School of Electrical and Information Engineering.

Perez, A., *et al*., 2002. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. *Statistics in Medicine*, 21, 3885–3896.

Piela, P., 2003. Exploitation of neural methods for imputation. *Federal Committee on Statistical Methodology Research Conference*, 49–54 (http://www.fcsm.gov/03papers/Piela.pdf).

Pyle, D., 1999. *Data preparation for data mining*. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Quinlan, J., 1986. Industion of decision trees. *In*: T. Mitchell, ed. *Machine learning*. Vol. 1, New York: McGraw-Hill, 81–106.

Ragel, A. and Cremilleux, B., 1998. Treatment of missing values for association rules. *The second Pacific-Asia conference on knowledge discovery and data mining*, 1394. Melbourne, Australia. 258–270.

Ragel, A. and Cremilleux, B., 1999. MVC- A preprocessing method to deal with missing values. *Knowledge-Based Systems*, 12 (5), 285–291.

Reams, G. and McCollum, J., 1999. Evaluating multiple imputation models for the southern annual forest inventory. *In*: *Proceedings of the section on statistics and the environment*, 8–12 August, Baltimore, MD: American Statistical Association, 13–18.

Richman, M.B., Trafalis, T.B., and Adrianto, I., 2007. Multiple imputation through machine learning algorithms. *Fifth conference on artificial intelligence applications to environmental science*, 3.9. San Antonio, TX: American Meteorological Society.

Rosenbaum, P. and Rubin, D., 1983. Assessing sensitivity to an un-observed binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society*, 45, 212–218.

Rubin, D., 1987. *Multiple imputation for nonresponse in surveys*. New York: Wiley and Sons.

Rubin, D., 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91 (434), 473–489.

Samad, T. and Harp, S., 1992. Self organization with partial data. *Network*, 3, 205–212.

Sarac, A., Batia, R., and Drury, C., 2007. Extension of the visual search models of inspection. *Theoretical Issues in Ergonomics Science*, 8 (6), 531–556.

Schafer, J., 1997. *Analysis of incomplete multivariate data*. London: Chapman and Hall.

Schafer, J., 1999. *Software for multiple imputation* [online]. Available from: http://www.stat.psu.edu/~jls/misoftwa.html [Accessed 23 April 2008].

Schafer, J. and Graham, J., 2002. Missing data: Our view of the state of the art. *Psychological Methods*, 7 (2), 147–177.

Scheffer, J., 2002. Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, 3, 153–160.

Schunk, D., 2007. *A Markov Chain Monte Carlo multiple imputation procedure for dealing with item nonresponse in the German SAVE Survey* (http://www.mea.uni-mannheim.de/mea_neu/pages/files/nopage_pubs/2f0kofojt50kmddf_meadp_121-2007.pdf).

Shao, J. and Zhong, B., 2003. Last observation carry-forward and last observation analysis. *Statistics in Medicine*, 22, 2429–2441.

Stenbakken, G., 2004. Empirical modeling methods using partial data. *IEEE Transactions on Instrumentation and Measurement*, 53 (2), 271–276.

Sun, L., *et al.*, 2006. Estimating the uninsured costs of work-related accidents, part I: a systematic review. *Theoretical Issues in Ergonomics Science*, 7 (3), 227–245.

Swanson, N. and Sauter, S., 2006. A multivariate evaluation of an office ergonomic intervention using longitudinal data. *Theoretical Issues in Ergonomics Science*, 7 (1), 3–17.

Tabachnick, B. and Fidell, L., 2007. *Using multivariate statistics*. 5th ed. New York: Pearson.

Timofeev, R., 2004. *Classification and Regression Trees (CART) theory and applications*. Thesis (Masters). Humboldt University, Center of Applied Statistics and Economics, Berlin.

Vamplew, P., 1996. *Recognition of sign language using neural networks*. Thesis (PhD). Department of Computer Science, University of Tasmania.

Vamplew, P., *et al.*, 1996. Techniques for dealing with missing values in feedforward networks. *Proceedings of the Australian conference on neural networks*, Sydney. 250–254.

van Schoor, H. and Konz, S., 1996. Males/females: An anthropometric comparison for modelling missing data. *International Journal of Industrial Ergonomics*, 17, 437–440.

Viallefont, V., Raftery, A., and Richardson, S., 2001. Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine*, 20 (21), 3215–3230.

von Hippel, P., 2004. Biases in SPSS 12.0 missing value analysis. *American Statistician*, 58, 160–165.

Wang, H. and Wang, S., 2009. Discovering patterns of missing data in survey databases: An application. *Expert Systems with Applications*, 36, 6256–6260.

Wang, S., 2003. Application of self-organising maps for data mining with incomplete data sets. *Neural Computing & Applications*, 12 (1), 42–48.

Waters, T., *et al.*, 2006. Cumulative spinal loading exposure methods for manual material handling tasks. Part 1: is cumulative spinal loading associated with lower back disorders? *Theoretical Issues in Ergonomics Science*, 7 (2), 113–130.

Wayman, J., 2003. Multiple imputation for missing data: What is it and how can i use it? *Annual meeting of the American Educational Research Association*. Chicago, IL. 1–16.

Westin, L., 2002. *Missing data and the preprocessing perceptron*. Umeå University: Department of Computing Science.

Williams, D., *et al.*, 2005. Incomplete-data classification using logistic regression. *Proceedings of the 22nd international conference on machine learning,* 119. Bonn, Germany: ACM Press, 972–979.

Yenduri, S., 2005. *An empirical study of imputation techniques for software data sets*. Thesis (PhD). Louisiana State University and The Department of Computer Science Agricultural and Mechanical College.

Ziarko, W. and Yao, Y., eds., 2001. A comparison of several approaches to missing values in data mining, *In*: *Rough sets and current trends in computing*. London: Springer, 378–385.

About the authors

William Young is a doctoral candidate in the Integrated Engineering program at Ohio University. His dissertation is focused on developing a team-compatibility decision support system. To fund this project, Mr. Young received Ohio University's 2007 Student Enhancement Award, which promotes creative academic research. Mr. Young has worked on projects that were funded by the General Electric Aircraft Engines, National Science Foundation (GK-12 Fellow), and the Ohio Department of Labor. William received his bachelor's (BSEE) and master's (MSEE) degrees in Electrical Engineering at Ohio in 2002 and 2005, respectively. Young's research interest is focused on developed decision supports systems utilising statistical and machine learning methodologies. Specifically, these topics include: cost modelling, team-compatibility, virtual worlds for educational, sports modelling, ecological monitoring, and advanced knowledge extraction techniques.

Gary Weckman was a faculty member at Texas A&M University-Kingsville for six years before joining the Ohio University faculty in 2002 as an associate professor in Industrial and Systems Engineering. He has also practiced industrial engineering for over 12 years with firms such as General Electric Aircraft Engines, Kenner Products and The Trane Company. Dr. Weckman's primary research focus has been multidisciplinary applications utilising knowledge extraction techniques with artificial neural networks. He has used ANNs to model complex systems such as large-scale telecommunication network reliability, ecological relationships, stock market behaviour, and industrial process scheduling. In addition, his research includes industrial safety and health applications and is on the Advisory Board for the University of Cincinnati NIOSH Occupational Safety and Health Education and Research Center Pilot Research Project.

William Holland is currently an Operations Research Analyst in Alexandria, VA. While attending Ohio University, Holland received an MS degree in Industrial and Systems Engineering in 2009 as well as a BS in Applied Mathematics in 2005. In pursuit of his academic degrees, William studied RFID location systems, machine learning methods, sports modelling, green facility renovation and artificial neural networks. From these endeavours, he presented his research at annual conferences and research consortiums, which included the Alien RFID Solutions Center Expo, 2008 AIDCTI Conference at Ohio University, and the 2009 Industrial Engineering Research Conference in Miami, FL.