Brazilian Journal of Probability and Statistics
2016, Vol. 30, No. 4, 521–539
DOI: 10.1214/15-BJPS292

Multiple imputation of unordered categorical missing data: A comparison of the multivariate normal imputation and multiple imputation by chained equations

Innocent Karangwa, Danelle Kotze and Renette Blignaut

University of the Western Cape

Abstract. Missing data are common in survey data sets. Enrolled subjects do not often have data recorded for all variables of interest. The inappropriate handling of them may negatively affect the inferences drawn. Therefore, special attention is needed when analysing incomplete data. The multivariate normal imputation (MVNI) and the multiple imputation by chained equations (MICE) have emerged as the best techniques to deal with missing data. The former assumes a normal distribution of the variables in the imputation model and the latter fills in missing values taking into account the distributional form of the variables to be imputed. This study examines the performance of these methods when data are missing at random on unordered categorical variables treated as predictors in the regression models. First, a survey data set with no missing values is used to generate a data set with missing at random observations on unordered categorical variables. Then, the two methods are separately used to impute the missing values of the generated data set. Their performance is compared in terms of bias and standard errors of the estimates from the regression models that determine the association between the woman's contraceptive methods use status and her marital status, controlling for the region of origin. The baseline data used is the 2007 Demographic and Health Survey (DHS) data set from the Democratic Republic of Congo. The findings indicate that although the MVNI relies on the statistical parametric theory, it produces more accurate estimates than MICE for nonordered categorical variables.

1 Introduction

Missing data are common in survey research. Enrolled subjects do not often have data recorded for all variables of interest. This is due, for instance, to data entry errors or ineligibility or refusal by the respondents to answer some items from a survey. As a result, missing values are created in data sets, and if they are not modelled properly, it can lead to incorrect inferences. A common way of handling missing values is to discard them from the analysis, a technique that is provided by default in many statistical packages such as SPSS, STATA, and SAS amongst others. This approach is referred to as case deletion or complete case analysis

Key words and phrases. Missing data, missing at random, multiple imputation, multivariate normal imputation, multiple imputation by chained equations, categorical data.

Received September 2014; accepted April 2015.

and can lead to low power of the statistical test and biased parameter estimates when the proportion of missing values is high and data are missing in a systematic manner or at random (Graham, 2009).

To reduce these problems, various methods of rescuing missing data have been developed (Schafer and Graham, 2002; Tsikriktsis, 2005; Graham, 2009). Items with no observations at all are directly discarded from the analysis because they do not provide any particular information about the data. However, if data are partially missing on variables of interest, the latter should not be discarded as they still contain some information that can be used to draw useful inferences.

Estimating a model without doing any kind of processing when data are missing is difficult. For example, if a linear regression has to be run, say Y as a function of X_1 and X_2, but some of the values of X_1 and X_2 are missing, it is still possible to fit regression coefficients to the independent variables. One way of doing this is to get rid of the missing information and use the available data, which is sometimes problematic as previously stated. But when the researcher is forced to use the data set with missing data without discarding cases, it has to be done in a way that minimizes the damage to the inferences to be drawn.

The first thing to do is to identify the missingness mechanisms in the data set or the reasons why data are missing. These include the missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) mechanisms. Data are MCAR if the probability that a particular value is missing is not related to the value itself or any other observed values in the data set. When the probability that a particular value is missing depends on observed values in the data set, the missing mechanism is referred to as MAR data. These two mechanisms are termed ignorable, because conditional on the observed data set, one can draw valid inferences. If missingness is related to unobserved values in the data set, the missing mechanism is called nonignorable. In this case, even conditioning on observed data does not lead to valid inferences. Data sets with such missing mechanism is known as not missing at random or NMAR (Schafer and Graham, 2002; Graham, 2009).

If not fixed, all these missingness mechanisms may lead to serious consequences. Discarding cases with missing data from the analysis, for instance, leads to the lack of efficiency or greater variability in the obtained results. Not modelling MAR and NMAR data lead to bias and efficiency problems. When modelling MCAR and MAR data to look like nonmissing data, observed data are used to impute missing values. As a result, bias and efficiency problems are reduced.

A number of methods have been developed to model MAR and MCAR data. These include single-based imputation methods such as the mean imputation, regression imputation, interpolation (for panel data), multiple imputation based methods such as the multivariate normal imputation (MVNI) and the multiple imputation by chained equations or MICE (Raghunathan et al., 2001; Van Buuren, 2007).

These last two multiple imputation-based methods are increasingly being used and have been made popular in almost all the main statistical software packages such as SAS, STATA, etc. These methods are considered the best as they account for the statistical uncertainty in the imputations, which is not the case when single-based imputation methods are used (Lee and Carlin, 2010).

Despite the popularity of these methods, there is still no clear guidance on which method to choose between the two when the multiple imputation needs to be done on continuous, binary and categorical (polytomous with more than two categories) variables containing missing values.

The MVNI was initially designed to handle missing data of continuous and normally distributed variables, but it was later used to impute missing values of categorical data which do not assume normality (Allison, 2001). On the other hand, the MICE, also known as imputation by fully conditional specification (Van Buuren and Knook, 1999), conditional model (Carpenter and Kenward, 2012) or sequential regression multiple imputation (Raghunathan et al., 2001; Van Buuren, 2007), fills in missing values sequentially, taking into account the distributional form of the variables to be imputed. Details about this method are also given in Van Buuren (2007), Royston and White (2011) and Kropko et al. (2014) amongst others.

Several studies have compared these two techniques in terms of parameter estimation and standard errors and have indicated that these two methods produce approximately the same results when data are missing on continuous and normally distributed data (Raghunathan et al., 2001; Karangwa and Kotze, 2013; Kropko et al., 2014). The multivariate normal imputation outperformed the multiple imputation by chained equations when data were missing on ordinal data (Lee and Carlin, 2010; Finch, 2010) and on binary variables (Lee and Carlin, 2010). As suggested by these two authors, an empirical study is still needed to determine the performance of these two methods when data are missing on nonordered or nominal categorical variables. Kropko et al. (2014) attempted to compare the performance of these methods when data were missing at random on continuous, binary, ordinal and unordered categorical variables that were used as outcome variables in the regression models. Their findings indicated that MICE performed better than MVNI in terms of regression coefficients' accuracy.

This study considers the suggestion by Finch (2010) and Lee and Carlin (2010) to examine the performance of these methods when data are missing at random on nonordered categorical variables treated as predictors in the regression models. Simulated data sets with missing at random (MAR) observations on nominal variables with more than two categories are used to assess the performance of these methods.

The remainder of the paper is as follows. In Section 2, some single based imputation methods are reviewed. In Section 3, the theory behind the multiple imputation method is revised with a particular emphasis on the MVNI and MICE. In Section 4, the methodology to be used is highlighted. In Section 5, the results and discussion are provided. The last section (Section 6) provides the conclusion and recommendations for further research.

2 Single-based imputation methods

2.1 Mean imputation

The mean imputation replaces missing values with the observed mean of the available data on the variable containing missing data. With this technique, the efficiency problem is solved but standard errors of the estimates are underestimated (Schafer, 1997; Carpenter and Kenward, 2013). In addition, the estimate of the mean is treated as true whereas it is not the case. However, imputing missing values using the mean of the observed data is a good guess, better than not doing anything at all if there are no other options or the researcher does not have any knowledge about other missing data methods.

2.2 Regression imputation

Regression imputation consists of using some selected prediction of a missing value on a variable of interest. For instance, to predict a missing value for the variable say, X_1, use this variable as a function of other variables, say, X_2 and X_3, in a model that could even include the dependent variable, say Y. As an illustration, suppose that the initial model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \qquad \varepsilon \sim N\left(0, \sigma_\varepsilon^2\right). \tag{1}$$

To get the best guess of X_1, the following prediction is proposed:

$$X_1 = \hat{\phi}_0 + \hat{\phi}_1 X_2 + \hat{\phi}_2 X_3 + \hat{\phi}_3 Y + \varepsilon_1, \qquad \varepsilon_1 \sim N\left(0, \sigma_{\varepsilon_1}^2\right). \tag{2}$$

Just like the mean imputation, the uncertainty is not incorporated very well because the estimates are random variables. Therefore, there is uncertainty in $\hat{\phi}$ that should be incorporated in the model of Y. That is, if the estimated X_1 is substituted in the model of Y, the uncertainty on how the $\hat{\phi}$ coefficients were obtained should fit into uncertainty in β. However, the problem of this method is that it yields small standard errors (Carpenter and Kenward, 2013).

2.3 Imputation using interpolation

In panel data, interpolation is used to impute missing values (Norazian et al., 2008). For instance, suppose that a variable X is measured at times $t = 1, 2, 3$ (X_1, X_2 and X_3) and some of the values are missing at time $= 2$ (X_2). With this method, the quantity $X_2 = \frac{X_1 + X_3}{2}$ is computed and then substituted into the missing values. As highlighted by Norazian et al. (2008), this technique creates bias and large confidence intervals.

3 Multiple imputation

Single-based imputation methods mentioned earlier constitute an improvement over the case deletion method, but they do not account for uncertainty in the imputations as imputed values are treated as true rather than estimates of the missing values. This leads to the underestimation of the variance of the estimates and the distortion of relationships among variables (Stuart et al., 2009).

Currently, many researchers view the multiple imputation as a better way of doing imputation (Schafer and Graham, 2002; Azur et al., 2011). The goal of this method is to impute missing values in such a way that the uncertainty in the imputed values is accounted for. That is, imputed values are estimates rather than known values of missing observations, thus leading to the appropriate standard errors of the estimates.

This method uses a selected model such as the regression model to predict missing values based on observed data. Instead of picking one value for the missing value, many values are chosen and the uncertainty is represented in the variance covariance matrix (VCV) of β estimates used to predict missing values. As an example, suppose that a regression model of Y on X_1 and X_2 is run but the variable X_1 contains missing values. The following imputation model is specified:

$$X_1 = \phi_1 + \phi_2 X_2 + \phi_3 Y + \varepsilon_2, \qquad \varepsilon_2 \sim N\big(0, \sigma^2_{\varepsilon_2}\big). \tag{3}$$

In this case, there is a VCV matrix of the ϕ_i estimates that incorporates and measures uncertainty in extent to which Y and X_2 can be used to plug in the values of X_1. This can be done by just picking many copies of ϕ from its asymptotic distribution (e.g., a multivariate normal distribution for this regression model), and use the estimates of ϕ and the VCV(Σ) to fill in the mean and VCV of the distribution $\Phi(\hat{\phi}, \hat{\Sigma})$.

Suppose that the main model is the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_3, \qquad \varepsilon_3 \sim N\big(0, \sigma^2_{\varepsilon_3}\big). \tag{4}$$

The values of X_1 are imputed using the imputation model in (3) and m copies of $\hat{\phi}$ are drawn from the asymptotic distribution of $\hat{\phi}$. Now m copies of the data that gives m copies of the β estimates are created when those m data sets are plugged back into the original model of Y. Therefore, m estimates of β for each data set are obtained, and from there, the final estimate of $\hat{\beta}$ is calculated. In other words, all the estimates of β are combined by taking the mean of the m estimates of β. The variance V_β of the new (combined) estimate of β is a function of the within (W) data set variance, S^2_m, which is an ordinary least square (OLS) estimate of the normal σ^2, and B is the between data set variance, which is a variance due to uncertainty in the imputation of X_1. The above mentioned quantities can be technically presented as: $\hat{\beta} = \sum_{m=1}^{M} \hat{\beta}_m$, $V_\beta = W + (1 + \frac{1}{m})B$ where $W = \frac{1}{m}\sum_{m=1}^{M} S^2_m$ and $B = \frac{1}{m-1}\sum_{m=1}^{M}(\hat{\beta}_m - \hat{\beta})^2$. The factor $(1 + \frac{1}{m})B$ corresponds to

the inflation in the standard errors (SEs) of $\hat{\beta}$ which is done in order to correct for imputation (Rubin, 1987).

These quantities are not computed manually; many statistical software packages do that. The multivariate normal imputation or MVNI and multiple imputation by chained equations or MICE are among the best ways of combining these estimates or implementing these procedures. A brief description of these methods is given in the next two sections of this paper.

3.1 Description of multivariate normal imputation

As previously stated, the multivariate normal imputation or MVNI assumes that all the variables in the imputation model are normally distributed. The Markov chain Monte Carlo procedure is used to obtain imputed values from the estimated multivariate distribution, allowing appropriately for uncertainty in the estimated model parameters, which is a requirement for proper imputation (Rubin, 1987).

Assuming multivariate normally distributed data, at the tth iteration one needs to draw missing values $Y_{\text{mis}}^{(t+1)}$ from $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$ which is the distribution of missing data given the observed data Y_{obs} and the model parameters $\theta^{(t)}$ (such as regression coefficients and covariance matrix) of the previous iteration, and then draw new model parameters $\theta^{(t+1)}$ from $p(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{t+1}, \theta^{(t)})$; the posterior distribution of the unknown parameters given the observed data, the estimated missing values and previously estimated model parameters. The resulting sequence forms a Markov chain $\{Y_{\text{mis}}^{1}, \theta^{1}; Y_{\text{mis}}^{2}, \theta^{2}; \ldots; Y_{\text{mis}}^{t+1}, \theta^{t+1}\}$, which must converge to the conditional distribution $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ and is used to impute missing values (Jackman, 2000; Horton and Lipsitz, 2001).

This method works properly under the MAR assumption and can handle both continuous and categorical missing data although the latter do not assume normality (see Allison, 2001; Graham, 2009). Given, for instance, a binary or a two-level categorical variable coded as 1 and 0, the proportion of responses with 1s will be the same as the mean of that variable. Therefore, unbiased estimates for the variables are obtained even if multiple imputation-based models that assume normality are used. When a two-level categorical variable is used as a covariate or independent variable in regression analysis, the imputed values should be used without rounding. If this variable is to be used in the analysis as a discrete binary variable, then rounding should be done to the nearest value (0 or 1) as suggested by Bernaards, Belin and Schafer (2007). For categorical variables with more than two levels, these need to be dummy-coded first and $k - 1$ (where k is the number of categories) dummy variables are included in the imputation model (Allison, 2001). For example, if a variable such as marital status with six categories (never married, married, divorced, widow, living together and not living together) contains missing values and therefore needs to be imputed, it has to be dichotomized to obtain dummies for never married, married, divorced, widow and living together respectively. The imputation is done with only these five variables and filled-in values are used

to produce final coding, while the sixth category (not living together) is treated as a reference category.

3.2 Description of multiple imputation by chained equations

As any other imputation method, the MICE technique discards observations with no information at all. This makes sense because if there is no information provided on the variables to be used, regression coefficients, for instance, cannot be fitted. However, when information is partially missing on these variables, the procedure works as follows: (1) For all missing observations in the data set, missing values are filled in with random draws from the observed values first or a simple imputation such as the mean imputation is done for every missing value in a data set (Azur et al., 2011). (2) By moving through the columns of variables, a single variable imputation is performed using a method such as regression imputation. The obtained new guess is temporally used to fill in the missing value of the variable on which the regression was performed. Note that as we go along, previous guesses are used in the regressions of other variables to be imputed until the whole data set is imputed. (3) The new fitted values are used as replacements to the original inputs in stage (1). (4) The process is repeated until a certain number of cycles have been completed or until convergence is attained (or until the distribution of parameters governing the imputations becomes stable). By repeating steps 1–4 above m times, m imputed data sets are generated and analyzed using simple rules (Little and Rubin, 2002). According to (Raghunathan et al., 2001), ten cycles are generally performed. However, as suggested by (Azur et al., 2011), research is needed to determine the best possible number of cycles required to impute data under different conditions. On the other hand, the number of m imputed data sets depends on the size of the data set and the amount of missing information in the data set. When generating imputations, a linear regression is used for continuous data, a logistic regression is applied for binary variables, multinomial logistic and Poisson regressions are utilised for polytomous and count variables respectively, etc. In fact, the choice of a regression model depends on the nature of the variable to be imputed.

Another approach that is considered to be the best is the MICE Bayesian data augmentation (see Rubin, 1987). It can be compared to a Markov chain that works as follow: (1) start with imputation values (obtained from the mean imputation for example) and update each imputation based on the state of the rest of the imputed values. For instance, consider variables Y, X_1, X_2 and X_3 with initial values chosen randomly. To impute X_3 the values of Y, X_1 and X_2 are used to generate imputation values. To impute X_2, the imputed values of X_3 are used together with the values of Y and X_1 and so on. This is in fact how the Markov chain is updated using the Gibbs sampling method (Geman and Geman, 1984; Gelfand and Smith, 1990). In other words, the previous states of the Markov chain are utilised plus any update that was already made about this particular iteration to create a new link in the chain for the variable of interest.

As this method is compared to a Markov chain, the aim is to build Markov chains as part of the Bayesian estimates to draw samples from the posterior distribution in order to derive some inferences about that posterior. To build a missing data model that fits the Bayesian approach, missing values are treated as other parameters to estimate by drawing them from their posterior distribution. The model is as follows:

$$f(\beta, Y_{\text{mis}}|Y_{\text{obs}}) \propto f(Y_{\text{obs}}|\beta, Y_{\text{mis}})\, f(\beta, Y_{\text{mis}}), \tag{5}$$

where β denotes the model parameters, Y_{mis} and Y_{obs} are the missing part and observed data respectively. Equation (5) says that $f(\beta, Y_{\text{mis}})$ is a function of the data at hand rather than $f(\beta)$. It is written as $f(\beta, Y_{\text{mis}}|Y_{\text{obs}})$ and is proportional to the distribution of observed data conditional on model parameters β and the missing values, Y_{mis}, times some prior about β and Y_{mis}. This proportional relationship between the right and the left of the equation (5) constitutes a key to the Bayes law (Rubin, 1987). In this case, what is done (which is an augmentation process) is sampling not only the model parameters β, but also the missing values, Y_{mis}, out of the posterior distribution using the Markov chain Monte Carlo procedure. By doing so, many samples of Y_{mis} and β are obtained. The draws or samples of Y_{mis} serve as imputations or filled in missing values.

Despite the growing popularity of MICE, it lacks theoretical justification (Raghunathan et al., 2001). One concern is the probable incompatibility among the conditional models; that is, the possibility that there is no joint distribution with the conditionals of the assumed forms (He et al., 2010). However, as suggested by Brand (1999) and Schafer and Graham (2002), this should not be a big problem in applied settings. A number of researchers continuously use this technique as they believe that it is the right method to handle any missing data given its flexibility and capability to be used in a broad range of settings (Azur et al., 2011; Hughes et al., 2014; Twisk et al., 2013; Lee and Carlin, 2010).

MICE works under the assumption that data are missing at random (MAR) and unbiased results can only be obtained when this assumption is met.

4 Methodology

4.1 Description of the data sets used

The 2007 Demographic and Health Survey (DHS) data set conducted in the Democratic Republic of Congo (DRC) was used for the analysis. It consists of a household and women's questionnaire, where a sample of women of reproductive age (15–49 years old) were interviewed regardless of their marital status, in each sampled household. Data were collected on fertility and family planning in addition to socio-demographic and economic data. The sample considered in the analysis included only women with the characteristics mentioned earlier, who were not

pregnant at the time the interview was conducted and who were sexually active. Respondents were asked about their knowledge and use of contraceptive methods, etc. Information on whether they have ever used any contraceptive method was first obtained and then the types of contraceptive methods used were asked. Contraceptive methods used included the modern (i.e., pill, injections and other), traditional (i.e., abstinence and other) and folkloric (i.e., herbal plant and other) methods. The dependent variable considered for the analysis was the women's contraceptive use status measured as any contraceptive method used by including all women who reported using modern, traditional and folkloric methods coded as 1 and 0 to represent women who have never used any contraceptive method. The objective was to determine the impact of marital status on contraceptive methods status use, controlling for the region where the respondent resides.

To assess the performance of the multiple imputation using the MICE and joint model or MVNI techniques, a completely observed data set or data set with no missing data was used as a baseline. Based on this data set, a data set with values missing at random (MAR) was created in such a way that missingness (missing or present) was related to variables of interest in the data set. This assumption is more practical than MCAR, as missing values must be connected to some of the variables in the data sets. Therefore, to obtain the data set with missing values according to this assumption, values were deleted such that the bias could be corrected with the observed data at hand. As the study baseline example is to determine the impact of marital status on the contraceptive methods use status of the women of reproductive age, controlling for the region of origin, values (about 30%) were deleted at random on variables marital status and region for women who were not using any contraceptive method. This allowed missingness to be associated with the contraceptive method use status, which is a requirement of the MAR assumption.

4.2 Analysis method

4.2.1 *Imputation of missing values*. Existing multiple imputation-based methods assume that data are MAR, but as suggested by Rubin (1987), MNAR and MCAR missing mechanisms can also be assumed if the objective is to compare the performance of the multiple imputation-based methods. This study compares the performance of the multivariate normal model (MVNI) and the multiple imputation by chained equations (MICE) in terms of bias and standard errors' estimates when data are MAR.

The multiple imputation method normally replaces each missing value by an array of $m > 1$ pseudo random values generated by a computer algorithm (Rubin, 1987) included in many statistical software packages such as SPSS, STATA, SAS, etc. As highlighted by White, Royston and Wood (2011), many studies on multiple imputations say that 3 to 5 imputations are enough. However, according to Wood et al. (2005), larger numbers of imputations are required if the objective

is to compare imputation methods. To obtain sufficient accuracy while comparing these methods, 100 imputations were used, which resulted in 100 different imputed simulated versions of complete data sets. Each imputed data set was analysed separately using standard statistical techniques and the point estimates and standard errors were recorded and then combined (averaged) to produce single estimates that account for uncertainty due to missing data (Rubin, 1987). To avoid bias in the analysis model, all the variables in the regression analysis model were included in the imputation model as suggested by (White, Royston and Wood, 2011).

As this study example is to examine the impact of the marital status of a woman of reproductive age on contraceptive method used controlling for her region of origin, only the dependent variable (contraceptive method use status) was included in the imputation model. The variable marital status is a 6-level categorical variable (never married, married, living together, widowed, divorced and not living together) with missing at random values on it, whereas the variable region contains 11 levels (Kinshasa, Bas Kongo, Bandundu, Equateur, Orientale, Nord Kivu, Maniema, Sud Kivu, Katanga, Kasai Occidental and Kasai Oriental). To impute these variables using MVNI, the Allison (2001) approach was used. That is, they were first dichotomized before being imputed. Therefore, five dummy variables (married, living together, widowed, divorced and not living together) were created for marital status and included in the imputation model, treating the first category (never married) as a reference category. Ten dummies (Bas Kongo, Bandundu, Equateur, Orientale, Nord Kivu, Maniema, Sud Kivu, Katanga, Kasai Occidental and Kasai Oriental) were also formed and included in the imputation model treating the first category (Kinshasa) as a reference. As suggested by several authors, variables with binary outcomes can be imputed using parametric-based imputation methods such as MVNI. These include Catellier et al. (2005) and Efron (1994). Therefore, after dichotomisation, the variables considered in this study were imputed taking into account the suggestions of these authors. The rounding of the dichotomized variables was not needed as these variables were used as independent variables in regression analysis. Dummy variables were included in the regression model of interest to estimate parameters of interest.

The MVNI was performed using Stata's implementation of Schafer's NORM program (Galati and Carlin, 2008) while the imputation using MICE was carried out using a multinomial logistic regression to fill in missing values of nominal variables (marital status and region variables in our case). The MICE command in Stata was used to perform imputation with MICE (see Van Buuren and Knook, 1999).

4.2.2 *Model development and computation of the performance measures*. The regression model with the baseline data set or data set with no missing values was first estimated to get the values of the regression coefficients and their corresponding standard errors. Then regression models with the imputed data sets using MVNI and MICE were estimated and the results (in terms of slopes and standard

errors) were recorded. To judge the performance of the multiple imputation methods of interest (MVNI and MICE), the bias was first computed based on the values of the estimated coefficients for each data set. Then the MVNI and MICE techniques were compared in terms of bias and standard errors to assess their performance when data are missing at random (MAR) on unordered or nominal variables with more than two levels or categories.

4.2.3 *Imputation models' diagnostics*. The convergence check is normally done through the Markov Chain Monte Carlo (MCMC) sequence, which must converge to the desired distribution. This process is always accompanied by the investigation of the serial dependence among the MCMC draws to obtain independent imputations. In fact, at each iteration, say T^t, for instance, the imputation model is first estimated using the observed data and the imputed data from the previous iteration, say T^{t-1} and so on. New imputed values are then drawn from their distributions. Consequently, each iteration is correlated with the previous imputation. The first iteration is normally known to be atypical or different from other iterations and because iterations are correlated, it can make other following iterations atypical too. To avoid this problem, the algorithm of the multiple imputation methods goes through the 10 first iterations and save only the results of the 10th iteration. The first 9 iterations are referred to as the burn-in, say b, and to attain this period, the number of imputations should be increased (Schafer, 1997).

Convergence is always examined using the series of parameter estimates rather than the series of imputations. It is frequently examined visually from the trace plots, which are plots of estimated parameters (summaries of the distribution such as means, standard deviations, quantiles, etc.) against iteration numbers, and autocorrelation plots of the estimated parameters. Long-term trends in trace plots and a high serial dependence in autocorrelation plots indicate a slow convergence to stationarity.

When the number of parameters in the imputation model is large, it may not be possible to examine the convergence of all the individual series. One way to solve this problem is to find a function of the parameters that would be the slowest to converge to stationarity. If the series of this function converges, then it is a good sign that other functions will converge too, especially, individual parameter series. As suggested by Schafer (1997), the worst linear function (WLF) is a good choice of that function. It is a scalar function of parameter estimates which is worse in the sense that its function values converge most slowly among parameters in the MCMC method. For linear functions of the estimated parameters, a worst linear function of parameters has the highest rate of missing information. Schafer (1997) was able to show that when the observed data posterior distribution is approximately normal, this function becomes one of the slowest to attain stationarity. The number of iterations needed for the chain to reach stationarity, say b, and the number of iterations between imputations needed to get independent values of the

chain, say k, can be also determined by checking the convergence of WLF. Visually, long-term trends in the estimates of the WLF are an indication that the chain has attained convergence, whereas autocorrelations in the WLF gives an idea of how many iterations can be used between imputations to ensure their approximate independence.

The number of iterations needed for MVNI and MICE to converge to the stationary distribution depends on, among other things, the proportion of missing information and initial values. The higher the percentage of missing information and the farther the initial values are from the type of the posterior distribution of missing data, the slower the convergence, and thus the larger the number of iterations required. The existing literature suggests that in many real-world applications, a number of burn-in iterations sufficient for convergence lies between 5 and 20 iterations (Van Buuren, 2007). In this study, both trace plots and the worst linear function were used to assess convergence as the number of parameters in the imputation was large (16 in total).

5 Results and discussion

The primary goal of this paper was to investigate whether MVNI and MICE produce similar results when data are missing at random on nominal variables with more than two levels or categories. Previous studies have already compared these methods under different circumstances and the results were reported. Schafer (1997), Raghunathan et al. (2001) and Karangwa and Kotze (2013), for instance, have found that these methods produce similar results when applied to continuous and normally distributed data containing missing at random values. Finch (2010) and Lee and Carlin (2010) showed that MVNI yield better estimates than MICE when data are missing at random on ordinal variables. The performance of these methods was also investigated when observations were partially missing on binary variables and their findings indicated that MVNI still outperformed MICE in terms of parameter estimation and standard errors when data were missing at random (Lee and Carlin, 2010; Demirtas et al., 2008). Kropko et al. (2014) investigated the performance of these two methods when data were missing at random on continuous, binary, ordered and unordered categorical data with more than two categories that were treated as outcome variables in the regression models. Their findings revealed that MICE resulted in estimates that were less biased than MVNI for each type of variable.

As suggested by Finch (2010) and Lee and Carlin (2010), further research was still needed to determine the performance of these methods when data were missing at random on unordered or nominal variables. This study assessed the performance of these methods when data were missing at random on nominal variables used as covariates in the regression models. Under MVNI, the suggestion by Allison (2001) was taken into consideration. That is, the dichotomization of

Table 1 *Estimates of bias and standard errors (SE) obtained when data are missing at random on variables marital status and region*

Variable	MVNI	MICE
Marital status		
Married	−0.013 (0.133)	0.007 (0.137)
Living together	−0.038 (0.139)	−0.018 (0.143)
Widowed	0.126 (0.180)	−0.014 (0.196)
Divorced	−0.072 (0.176)	−0.039 (0.180)
Not living together	−0.026 (0.149)	−0.011 (0.153)
Region		
Bas Kongo	0.032 (0.062)	0.015 (0.066)
Bandundu	0.049 (0.062)	0.018 (0.064)
Equateur	0.055 (0.064)	0.045 (0.065)
Orientale	0.020 (0.075)	0.018 (0.074)
Nord Kivu	0.025 (0.072)	0.020 (0.073)
Maniema	0.038 (0.067)	0.033 (0.068)
Sud Kivu	−0.010 (0.077)	−0.011 (0.079)
Katanga	0.019 (0.065)	0.008 (0.066)
Kasai Occidental	0.025 (0.077)	−0.008 (0.071)
Kasai Oriental	−0.070 (0.073)	−0.029 (0.076)

the variables of interest before their imputation. When imputing with MICE, the multinomial logistic regression was used as the variables to be imputed were polytomous.

The results regarding the performance of these methods are highlighted in this section. In Table 1, the bias and standard errors obtained after the imputation of missing values of the variables of interest, marital status and region, using MVNI and MICE techniques are presented. In Figures 1 and 2, these two measures are plotted. The plot of bias shows that the MVNI method produces less biased estimates (estimates of bias are close to zero) of slopes than MICE (see Figure 1).

The plot of standard errors of the data set with no missing values (*SE_BD*) and standard errors of the imputed data sets with MVNI (*SE_MVNI*) and MICE (*SE_MICE*) is shown in Figure 2. As indicated, MVNI yields standard errors that are closer to the standard errors from the model estimated with the data set without missing data compared to MICE.

To ensure that the imputations converged to the desired distribution, the convergence check was done. The estimates of the WLF were plotted against the iteration numbers first and then versus the lag numbers for both MVNI and MICE methods. The results are shown in Figures 3 and 4 for the MVNI approach, and in Figures 5 and 6 for the MICE technique. As indicated, the plots of the estimates of WLF against the iteration numbers show no visible trend, thus indicating that convergence is assured with the number of iterations used (1000 iterations). On the other

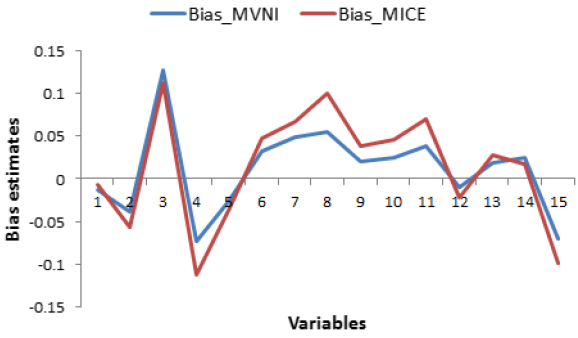

Figure 1 *Plot of bias after imputation of missing at random values using MVNI and MICE on the variables marital status and region with the numbers* 1–15 *referring to the estimated coefficients of variables defined in Table* 1.

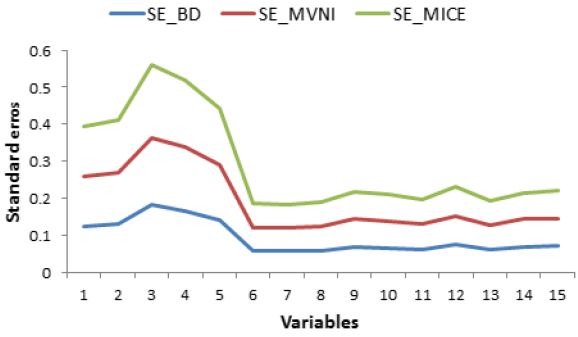

Figure 2 *Plot of standard errors of the slopes obtained after imputation of missing at random values using MVNI and MICE on the variables marital status and region with the numbers* 1–15 *referring to the estimated coefficients of variables defined in Table* 1.

hand, the plots of WLF's estimates against the lag numbers show the autocorrelations that die off quickly, which suggest that even a smaller number (of iterations) than what was used, such as 10 iterations between imputations, can be used to obtain independent samples.

Considering all the above findings, one can deduce that when data are missing at random on unordered or nominal variables with more than two levels, MVNI would be a better approach to impute missing values, as it appears to be less biased than MICE and produces better standard errors. These results differ from findings by Kropko et al. (2014) possibly for the following reasons: (1) Unordered data with missing at random observations considered in the analysis were treated as predictors rather than outcome variables in the regression models. (2) Regression models used in the analysis considered only unordered or categorical variables predictors, as we believe that the performance of these two methods can be fully observed when there is no influence of any other kind of variable. (3) A large number of imputations (100 imputations) were used in the imputation algorithm as opposed to

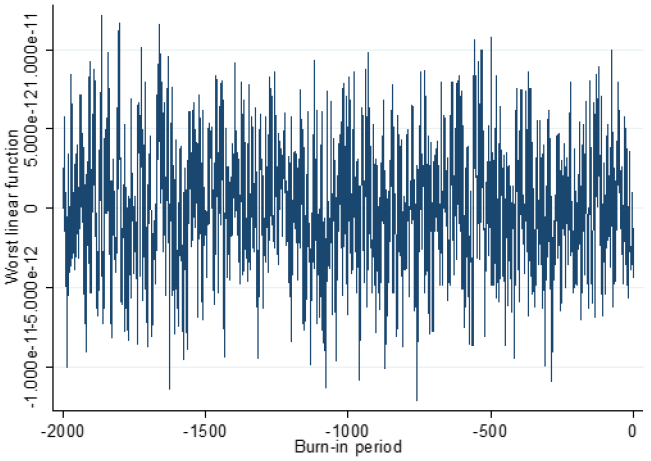

Figure 3 *Convergence of MCMC after MVNI when data are missing at random*: *plot of the estimates of WLF against the iteration numbers*.

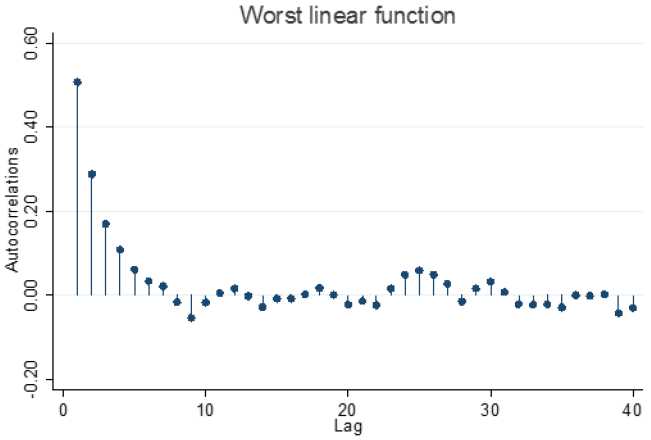

Figure 4 *Convergence of MCMC after MVNI when data are missing at random*: *plot of the estimates of WLF versus the lag numbers*.

Kropko et al. (2014) who used only 5. As suggested by Wood et al. (2005), when it comes to comparing missing data methods, the number of imputations should be increased as much as possible.

6 Conclusion and recommendations

When faced with missing values on nonordered or nominal categorical missing data, the aim of this study was to assess the performance of the MICE approach that takes into account the distributional form of the variables to be imputed and

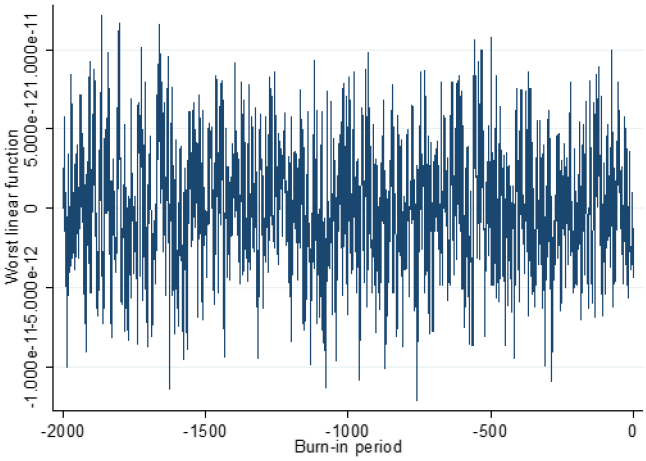

Figure 5 *Convergence of MCMC after MICE when data are missing at random*: *plot of the estimates of WLF against the iteration numbers*.

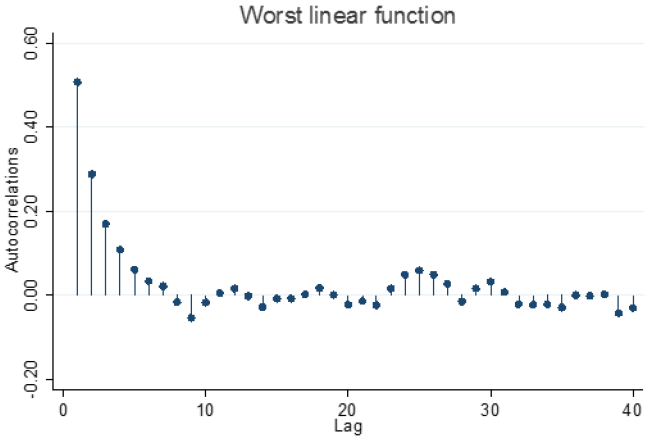

Figure 6 *Convergence of MCMC after MICE when data are missing at random*: *plot of the estimates of WLF versus the lag numbers*.

MVNI which assume a joint multivariate normal distribution of the variables in the imputation model. When imputing data using MVNI, the Allison (2001) approach was used. That is, the variables to be imputed were first dichotomised and then included directly in the imputation model without rounding, treating the first categories as references. The results obtained after imputation of data sets with MVNI and MICE techniques were compared to the results obtained using the data set with no missing data. The findings indicated that although MVNI was initially designed for parametric variables, it produces more accurate and less biased estimates than MICE when observations are missing at random. As with any research paper, this

study has some limitations. First, only unordered variables with missing at random data were considered in the analysis. In future, we aim to consider missing at random data for a mixture of variables types such as continuous, ordinal and nominal as predictors in the regression models. This study used a DHS data set, which is a complex survey with a complex sampling design and weighting procedure that need to be taken into consideration during the analysis. This issue is addressed by Reiter, Raghunathan and Kinney (2006), Schenker et al. (2006), He et al. (2010) and Molenberghs et al. (2015) amongst others. However, the results of this study are based on the regular database (without taking into account the randomization distribution due to the sample selection procedure) as we believe that this does not invalidate the findings that aimed to compare the multiple imputation methods of interest. Therefore, further research is still needed to determine the performance of MVNI and MICE when this issue is considered.

References

Allison, A. P. (2001). *Missing data*. Thousand Oaks, CA: Sage publications.

Azur, M. J., Stuart, E. A., Frangakis, C. and Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research* **20**, 40–49.

Bernaards, C. C., Belin, T. R. and Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine* **26**, 1368–1382. MR2345726

Brand, J. P. L. (1999). Development, implementation, and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. PhD Dissertation, Erasmus Univ., Rotterdam.

Carpenter, J. and Kenward, M. (2012). *Multiple imputation and its application*. Wiley: New York.

Carpenter, J. R. and Kenward, M. G. (2013). *Multiple Imputation and Its Application*. United Kingdom: John Wiley and Sons.

Catellier, D. J., Hannan, P. J., Murray, D. M., Addy, C. L., Conway, T. L., Yang, S. and Rice, J. C. (2005). Imputation of missing data when measuring physical activity by accelerometry. *Medicine and science in sports and exercise* **37**.

Demirtas, H., Freels, S. A. and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation* **78**, 69–84. MR2420089

Demirtas, H., Freels, S. A. and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation* **78**, 69–84. MR2420089

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association* **89**, 463–475. MR1294072

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science* **8**, 361–378.

Galati, J. C. and Carlin, J. B. (2008). *INORM: Stata Module to Perform Multiple Imputation Using Schafer's Method [software]*. Chestnut Hill, MA: Department of Economics, Boston College.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409. MR1141740

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology* **60**, 549–576.

He, Y., Zaslavsky, A. M., Landrum, M. B., Harrington, D. P. and Catalano, P. (2010). Multiple imputation in a large-scale complex survey: A practical guide. *Statistical Methods in Medical Research* **19**, 653–670. MR2744515

Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistical Association* **55**, 244–254. MR1963401

Hughes, R., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K. and Sterne, J. C. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology* **14**, 28.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science* **44**, 375–404.

Karangwa, I. and Kotze, D. (2013). Using the Markov chain Monte Carlo method to make inferences on items of data contaminated by missing values. *American Journal of Theoretical and Applied Statistics* **2**, 48–53.

Kropko, J., Goodrich, B., Gelman, A. and Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis*, 1–23.

Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* **171**, 624–632.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics. Wiley-Interscience*. Hoboken, NJ: Wiley Interscience. MR1925014

Molenberghs, G., Fitzmaurice, G. M., Kenward, G. M., Tsiatis, A. A. and Verbeke, G. (2015). *Handbook of Missing Data Methodology*. London: Chapman and Hall/CRC.

Norazian, M. N., Shukri, Y. A., Azam, R. N., Mohd, A. and Al Bakri, M. (2008). Estimation of Missing Data Using Interpolation Technique: Fitting on Weibull Distribution. In *Malaysian Technical Universities Conference on Engineering and Technology* (*Putra Palace, Perlis, Malaysia*, *8-10 March 2008*).

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberg, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85–95.

Reiter, J. P., Raghunathan, T. E. and Kinney, S. K. (2006). The importance of modelling the sampling design in multiple imputation for missing data. *Survey Methodology* **32**, 143.

Royston, P. and White, I. R. (2011). Multiple imputation by chained equations (MICE): Implementation in stata. *Journal of Statistical Software* **45**, 1–20.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley and Sons. MR0899519

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. *Monographs on Statistics and Applied Probability*, Vol. **72**. London: Chapman and Hall. MR1692799

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

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G. and Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association* **101**, 924–933. MR2324093

Stuart, E. A., Azur, M., Frangakis, C. and Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology* **169**, 1133–1139.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management* **24**, 53–62.

Twisk, J., De Boer, M., De Vente, W. and Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of Clinical Epidemiology* **66**, 1022–1028.

Van Buuren, S. and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681–694.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16**, 219–242. MR2371007

White, I. R., Royston, P. and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**, 377–399. MR2758870

Wood, A., White, I., Hillsdon, M. and Carpenter, J. (2005). Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *International Journal of Epidemiology* **34**, 89–99.

Department of Statistics and Population Studies
University of the Western Cape
Private Bag X17
Bellville 7535
Republic of South Africa
E-mail: ikarangwa@uwc.ac.za