"Science Stays True Here"
Journal of Mathematics and Statistical Science, Volume 2016, 624-643 | Science Signpost Publishing

Multiple Imputation for a Continuous Variable

Seppo Laaksonen

University of Helsinki, Finland.

Abstract

Multiple imputation (MI) is invented by Rubin in 1970's. He recommends to create imputations through a Bayesian process. Most software's are respectively following Bayesian principles. Since MI is not much used in official statistics, on the contrary to single imputation, any Bayesian framework is not there necessarily appropriate. Nevertheless, MI can be considered to be useful for several reasons. Björnstad suggests another approach, non-Bayesian imputation. It thus does not require any Bayesian rules for imputing the missing values several times. This paper first presents an approach to imputation that is not much applied. It strictly follows a two-stage strategy, so that the imputation model is first estimated and then completed by the imputation task. These tasks are called model-donor and real-donor that are clearer than those used traditionally. The approach is concretized with apattern of examples in which the variable being imputed is income. Bayesian and non-Bayesian methods are compared. The main result is that a Bayesian framework is not necessarily superior, at least if the applications are done by software's such as SAS and SPSS. If a user himself can create imputations using an appropriate non-Bayesian framework, the performance is often better.

Keywords: Imputation model, Imputation variance, Model-donor imputation, Poor vs rich fit, Real-donor imputation, SAS, SPSS.

1. Introduction

Imputation is for replacing missing or other incorrect values (later in the text the term 'missing' covers those other problematic cases as well) with plausible ones. If this procedure has been done once, it is single imputation (SI). SI is a usual tool in statistical offices or other public survey institutes, in

Corresponding author: Seppo Laaksonen, University of Helsinki, Finland. Box 68, FIN-00014, +358442222759.

particular.However, SI can be performed several times as well. If this procedure is repeated a number of times and 'coordinated' well, the outcome is 'multiple imputation' (MI). What such a good coordination means, it is a special question? Rubin in his books (1987, 2004) says that each imputation should be 'proper' (Rubin 2004, 118-119; note that since that we refer to the 2004 book that is slightly revised from the older one). Rubin also gives some rules for proper imputation but they are not necessarily easy to follow, or their implementation is not automatic. A big question here is how to repeat the imputation process well, that is, what is an appropriate Monte Carlo technique in order to get $L>1$ simulated versions for missing values?

Rubin (1996, 476, 2004, 75 & 77) also says that a theoretically fundamental form of MI is repeated imputation. Repeated imputations are draws from the posterior predictive distribution under a specific model that is a particular Bayesian model both for the data and the missing-data mechanism.

Several proper MI implementations are given in Rubin's books and in software packages (e.g. SAS and SPSS) using this book.He thus recommends that imputations should be created through a Bayesian process as follows: (i) specify a parametric model for the complete data, (ii) apply a prior distribution to the unknown model parameters, and (iii) simulate L independent draws from the conditional distribution of the missing data given the observed data by Bayes' Theorem.

These Rubin's theoretical principles are one starting point of this paper. A good point is that MI is not difficult to apply since most types of estimates can be computed in a usual way (e.g. averages, quantiles, standard deviations and regression coefficients). The Rubin's framework also serves the formulas both for point estimates and for interval estimates. The point estimates are simply averages of L repeated complete-data estimates, and thus very logical. His interval estimates are not indisputably accepted. Björnstad (2007) gives a modified version for the second component of Rubin's formula. This leads to a larger confidence interval, as a function of the rate of imputed values. This is logical since Rubin's formula is without any explicit term of the imputation amount but his Bayesian rules might implicitly include the same; this is however difficult to recognize.

Björnstad (2007, 433) also invents a new term, non-Bayesian MI, since his imputation is not following a Bayesian process. This term 'non-Bayesian' is not used in ordinary imputation literature; it cannot be found 6 years alter from a book by Carpenter and Kenward (2013) that much follows Rubin's framework but they use the term 'frequentist'.We still use the term 'non-Bayesian,' since we cannot say whether it is equal to 'frequentist.'

Björnstad motivates his approach also from the practical points of view saying that in national

statistical institutes (NSI's) the methods used for imputing for nonresponse very seldom if ever satisfy the requirement of being "proper." Moreover, Muñoz and Rueda (2009) say that several statistical agencies seem to prefer single imputation, mainly due to operational difficulties in maintaining multiple complete data sets, especially in large-scale surveys. We agree with these views. Since a non-Bayesian approach also leads to single imputation, that is commonly used in NSI's if anything has been imputed, a conclusion could be that MI cannot be applied using a non-Bayesian framework. We do not agree with this argument. Consequently, we have over years (Laaksonen 2000, Laaksonen 2003, Laaksonen 2016) applied non-Bayesian tools both for single and multiple imputation, although most often for single imputation. This paper first summaries our approach to imputation.

This approach first makes attempts to impute the missing values once. That is, the focus is first in single imputation. Correspondingly, the main target in imputations is to succeed in such estimates that are most important in each case. Since it is hard to impute correctly individual values, it is more relevant to try to get least unbiased estimates for some key estimates. Since we here concentrate on a continuous variable, that is, income, two types of estimates are of a special importance. One is income average and the other is income distribution, respectively. Income distribution can be measured by various indicators such as quantiles or Gini coefficient, but the coefficient of variation is here considered to be simple enough to indicate well income differences between people.

Rubin's approach can be implemented in various ways. We do not develop any own implementation but take advantage of the two existing implementations. These are derived from two general software packages, SAS and SPSS. We assume that their MI procedures follow a Bayesian process since there are such references in their manuals. Wethus use the term 'Bayesian MI' for application of SAS and SPSS. Respectively, our own imputation framework is called 'Non-Bayesian MI.'

There are naturally, several methods both for non-Bayesian and Bayesian imputations. We present here some key methods from both sides. Next section gives the core formulas for MI including Björnstad's version for the variance. Our own SI approaches are in more details explained in Section 3, the MI approaches in Section 4, respectively. Some special methods from software packages are explained in Section 5. One of these is predictive mean matching that is available both in SAS and SPSS, but our particular non-Bayesian methods have the same target to produce plausible (observed) values only.

Section 6 briefly describes our test data in which 27 per cent of incomes of persons are missing and thus required to impute. Good auxiliary data are available but not extremely good since the fit for the dependent variable is not high. The following two sections show our results via a number of imputation methods, first

for point estimates in Section 7 and then for interval estimates in Section 8. The final section gives first a brief conclusion from the main results with poorly fitting models but also gives some comparisons from the same data set into which a special artificial good covariate is added, and thus a strongly fitting model is estimated.

2. Point and Interval Estimates of Multiple Imputation

The point estimates of multiply imputed complete data sets are given similarly both by Rubin and Björnstad or all others. The parameter being estimated may thus be any statistic of interest. In this study, as explained in Section 1, we have chosen the two parameters. One is the overall income average and the second is the coefficient of variation (CV) of income, respectively.

In order to make the formulas simple, we denote the estimate by Q that is either the average or the CV. Such an estimate is calculated from a complete data set after each single imputation. Thus the estimate from a single imputed complete data set is Q_l and the respective variance is B_l, both calculated taken into account the sampling design. In our empirical data the sample is based on simple random sampling, and both estimates can be calculated using the simple formulas. Rubin even in 1996 (p. 479) says that 3 to 5 repeated imputations works well if the fraction of missing information is typical in careful surveys. In the Dacseis project of the EU (e.g. Laaksonen et al 2004) we were not fully happy with a small number imputations, in some cases even more than 10 imputations would have been needed. This number is mentioned also by Rubin (e.g. 2004, 227) but not generally recommended. For this study, we always calculate L=10 imputations that number seldom is exceeded in examples we have seen.

The MI point estimate is thus simply the average of the L imputations

$$(1)\quad Q_{MI} = \frac{\sum_i Q_l}{L}$$

Respectively, the variance estimate is

$$(2)\quad B_{MI-within} = \frac{\sum_l B_l}{L}$$ in which B_l is a SI variance.

There are two alternatives to calculate the MI variance of the L complete data sets. The first term of the variance, called within-imputation variability (variance), is in both cases equal that is formula (2). But the second term, the between-imputation-variability, is larger in Björnstad's version.

$$(3)\quad B_{MI} = B_{MI-within} + (k + \frac{1}{L})\frac{1}{L-1}\sum_l (Q_l - Q_{MI})^2$$

The difference is in the term $k = 1/(1-f)$ in which f is the fraction of missing values or the non-response rate. This increases while the fraction increases. Rubin's formula does not depend explicitly on the amount of the imputed values since $k = 1$. Rubin's Bayesian approach possibly takes this into account implicitly but it is hard to see. Björnstad developed his formula for some most common sampling designs including simple random sampling. Thus it is allowed to use this formula in this study although we give the results using Rubin's formula as well. Of course, it is not clear what is most correct in each case. It could be considered that Rubin's formula works with Bayesian MI and Björnstad's formula with non-Bayesian MI, but we will not give any definite conclusion to this even after the empirical tests.

3. Non-Bayesian Single Imputation for a Continuous Variable

Before imputation, it should be decided which variables are needed to impute. The first criterion for this decision is that some advantage is expected to get due to imputation. Such an advantage may be measured both with point estimates and with interval estimates. Usually, if the bias in a point estimate is smaller after imputation, it is a good think. On the other hand, the interval estimate should be reasonably small at the same time. Without imputation, both estimates thus are too biased.

In order to succeed in imputation, good auxiliary data or covariates in Rubin's terminology are needed. In the case of lacking covariates, simple methods based on observed values only can be applied. But if there are covariates both for the respondents and for the non-respondents, 'real' imputation methods can be used. In this case, the imputation framework (cf. Laaksonen 2016) includes the two core stages:

(i) Construction and implementing of the imputation model

(ii) Imputation itself or imputation task.

These two terms are also used by Rubin (2004) but these are integrated well together in our framework. Hence we use the term *IMAI* (Integrated modeling approach to imputation) in the examples of this paper.

Imputation model

An imputation model can be implemented using a smart knowledge of the imputation team or it can be estimated from the same data set or from a similar data set from an earlier survey or a parallel survey of another population. If the model is estimated from the same data set, it is expected that this replacer behaves more surely well in imputations (e.g. Chambers et al 2001). Hence we estimate the parameters of the imputation model from the same data set.

There are the two alternatives as a dependent variable in an imputation model. It is either (a) the variable being imputed or (b) the binary missingness indicator of the variable being imputed. The same auxiliary variables can be used in both models. Naturally, the estimations that are needed in the next step are derived from the different data sets, from the respondents for the model (a) and from both the respondents and the non-respondents for the model (b). The covariates need to be completely observed to compute the predicted values for the stage (ii).

Imputation task

The imputed values themselves can also be determined by the two options: (i) they are calculated using the imputation model or (ii) they are borrowed from the units with the observed values using the imputation model as well. The previous option is called 'model-donor' imputation, and the second is 'real-donor' imputation, respectively. The latter one is often called 'hot deck' but this term is not clear in all cases. Terms for the previous ones are often such that the model and the task are confused. For example, model imputation or regression imputation is not clear since these are referring to imputation model but the second step, imputation task, is not specified.

If a real-donor method is applied, an appropriate criterion and a valid technology to select a donor is needed. The natural criterion is to select an as a similar real-donor (observed values) as possible. This may be based on a kind of nearness metrics. If a clear criterion exists, it is good to select the nearest or another from the neighborhood. If any valid criterion does not exist, a random selection from the neighborhood can be used. This thus means that all units with observations are as close to each other within the neighborhood that can be called 'an imputation cell,'

This nearness metrics based on the predicted values is easier than such that are based on other information and technically on the Mahalanobis distance; for example, that is used in the Solas software (Statistical Solutions 2016). In this case, the user himself has to decide which auxiliary variables to use and how to scale them. Our approach is simpler since the metrics comes from the estimated imputation model.

In our approach, the predicted values of either the model (a) or the model (b) are used as the nearness metrics. The binary model (b) is used much earlier but the model (a) less often. Since it is not easy to see by an outsider what this means, we give one graph for illustrating it. Figure 1 shows how the predicted values vary at individual level when calculated either from the regression model or from the logit model. This scatter shows that those values differ much but they are fairly well correlated (0.32). Nevertheless, it

is expected that the imputed values differ to some extent. The figure also shows that some predicted values of the regression are negative whereas those are within the 'probability' interval (0, 1) in the logit regression.

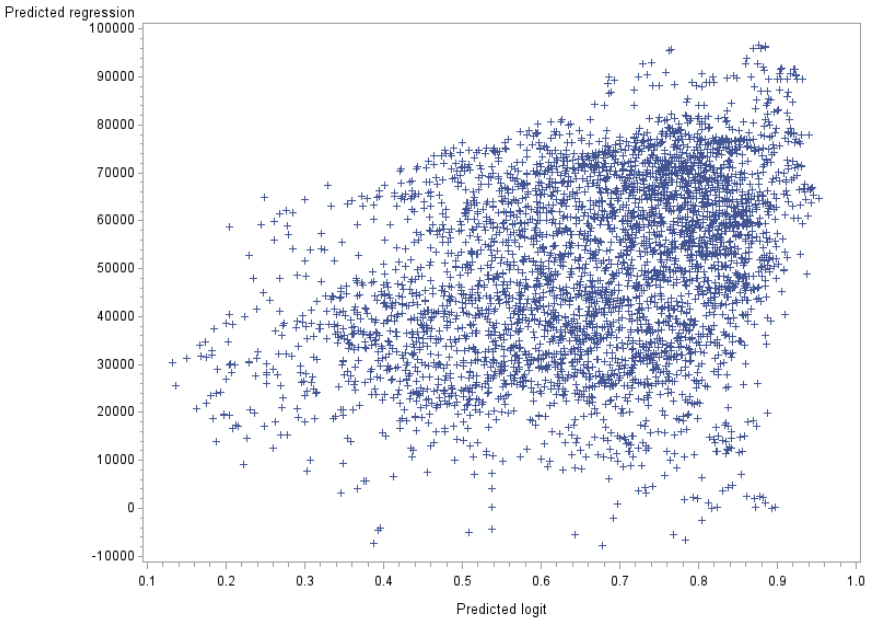

Figure1. The scatter plot between the predicted values of the logit imputation model (x-axis) and the linear regression imputation model (y-axis) for the 5315 non-respondents. Correlation = 0.32.

Imputation model and imputation task in this study

The variable being imputed is yearly income. For the imputation model this is here specified in the twomanners:

- Income as such.
- Binary indicator of the missingness of income.

It would be possible to build the imputation model for log-income but this case is not presented in this paper.

For the first dependent variable, ordinary multivariate regression model is usually applied, the target being to get an as good fit as possible, the criterion being the *R*-square of the model. In the second model, we here test three link functions: logit, probit and complementary log-log that are available in SAS which software package is used in our applications.

Each imputation model is estimated using a number of categorical variables that are described in Section 6. The auxiliary variables used in these models are equal to those found in the regression model;

this due to fairly compare the results.

The single model-donor imputations are simply predicted values of the linear regression.

The single real-donor imputations are made in the two rules, either using the predicted values of alinear regression models, or the predicted values (propensity scores) of a binary regression model. These predicted values are first sorted so that it is possible to search for a nearest (or near) neighbor of each missing-value unit.

Note that all these predicted values are calculated both for the respondents and the non-respondents even though that the true values are available for the respondents. The same predicted values are equal to the imputed values of the model-donor imputation but here the complete predicted values serve as nearest neighbor indicators for real-donor imputation. Note that a nearest neighbor may be close to the unit being imputed or quite far. In our tests, most neighbors were fairly close but a most distant neighbor was about 40^{th}closest. It is expected that the imputations are less good in such cases.

4. Non-Bayesian Multiple Imputation for a Continuous Variable

MI thus requires develop a tool for a Monte Carlo type of simulations to SI. It means that each MI imputation should be close to initial SI but certain randomness is added. Thus the MI point estimate (average of single MI estimates) should be close to the respective SI estimate.

Our MI procedure is solved in the same way for the regression model. Rubin (2004, 159) describes this model approach but his MI application (Section 5) is no more similar to this. Carpenter and Kenword (2013, 38) present an intuitive outline of the MI procedure that resembles our approach in the case of one auxiliary variable but it is difficult to see what their imputation specification finally is.

Our strategy for applying a multivariate linear regression for model-donor imputation first estimates the model. Next we compute the predicted values and then add to each of them a normally distributed random term with the zero mean and with the standard deviation of the estimated root mean square error (RMSE). The RMSE is possible to get as an output of the SAS regression procedure that gives opportunity to go forward without manual work as well. That is, this non-Bayesian method is not difficult to implement. Since some stochastic predicted values may be very large, either positive or negative, we have truncated this standard deviation within the range (-1, +1). This was observed as a good robustness tool already in Laaksonen (2003). The procedure removes some annoying outliers from the imputed data into an acceptable level. This is one appropriate technic in practice. In our experiments we used this robustness correction before the imputation task. In the results sections, the term ‘Robust’ means just this adjustment.

Nevertheless, plausible values are not ensured to obtain when robusting the model predictions. Some bad values, negative imputed incomes in this case, can still be obtained. We removed such values into zero for all methods. This is a simple solution but it is fair for method comparisons, and we call the results with this option by 'Constraints.' SPSS offers its own tool for constraints and we just use it, but the SAS MI constraints are similar as ours.

This 'error-term' based specification for MI in the regression models is logical. For each repeated imputation a seed number is changed. The point estimates (average and income difference) are expected to be close to pure model-donor imputation, whereas the variance and the standard error respectively are expected to follow the uncertainty of the imputations in a correct way since it depends on the model fit. The better fit thus leads to a smaller(imputation) variance.

The non-Bayesian MI procedure can be solved in the different ways if the imputation model is a binary regression. It is possible to add some random noise into the predicted values of the model, but what could be a criterion for that, it is not clear. Hence we do not present any such application but maybe later when we have found a good theory behind this application. A second option, preliminarily tested, is to add a random covariate into the imputation model. Since we do not know what type of such a term is appropriate, we have to continue this examination as well. Both these options look promising.

Instead our application for each model with three link functions is simply to categorize the predicted values into imputation cells. In our applications, 14 imputation cells are tested. We know that this is a subjective selection to some extent but the respective Bayesian method (Propensity score method in Section 5) is subjective as well. Within each cell a real-donor is selected randomly. Such a donor is found easily for most cells, but the three cells were problematic due to many missing values. This thus means that some estimates may cause the bias in estimates. This is a common problem in real-donor methods if the missingness is substantial. Of course, good covariates help in solving this problem.

5. Bayesian Multiple Imputation is Some SAS and SPSS Modules

SAS includes more MI modules than SPSS. Two modules look similar, that is, linear regression and predictive mean match imputation. Both software's have a specification of Markov Chain Monte Carlo (MCMC) as well. SPSS offers this specification both for linear regression and predictive mean matching. Its manual (2016) says that 'the fully conditional specification (MCMC) is suitable for data with an arbitrary pattern of missing values.' MCMC specification is in SAS for the regression imputation. Allison (2005) even considers that the most popular method for multiple imputation of missing data is the MCMC

algorithm that is the default method in SAS MI. We do not here describe the MCMC methods in details, but apply this option.Respectively, we firstpresent the basic points of the two methods, linear regression and predictive mean matching. In the end of this section we describe the SAS module of propensity score method that includes similar features as our real-donor method with binary regression imputation model.

SAS 9.3 documentation uses the term "Regression Method for Monotone Missing Data." It says that the regression method is the default imputation method for continuous variables in a data set with a monotone missing pattern.

In the regression method, a model is fitted for a continuous variable with the covariates constructed from a set of effects. Based on the fitted regression model, a new regression model is simulated from the posterior predictive distribution of the parameters and is used to impute the missing values for each variable (Rubin 2004, 166–167). That is, for a continuous variable with missing values, anordinary least squares multivariate regression model is specified and estimated, similarly as in Sections 4 and 5 for a non-Bayesian case. But the Bayesian MI procedure is not equal to our non-Bayesian IMAI method.

The SAS document says that the following three coresteps are used to generate imputed values for each imputation:

1. New regression parameters are drawn from the posterior predictive distribution of the parameters. That is, they are simulated. Our strategy does not take care of any posterior distribution but the regression estimates are directly derived from the model results.

2. Using a chi-square random variate the regression coefficients are drawn so that the respective random terms are added to the initial coefficients. This is a big difference to our approach since we use the regression error term in simulations. Rubin's method thus is problematic in our view since it introduces the randomness for each regression parameter. This is hard to understand since it changes the model specification unless one covariate only is used. A problem thus is that if regression coefficients vary randomly to some extent, the whole multivariate model does not correspond to any real situation since the covariates depend on each other to some extent (collinearity). It means that the predicted values are damaged more or less although the variability is ensured. In our approach this problem is definitely avoided since one randomness term only is used.

3. The missing values are then replaced by the regression prediction function that includes randomness for each regression coefficient and also a normally distributed random term. This last point is close to our method. It is however difficult to see what really happens during this procedure.

Predictive mean matching imputation

The predictive mean matching method is an imputation method available for continuous variables. It is similar to the regression method except that for each missing value, it imputes a value randomly from a set of observed values whose predicted values are closest to the predicted values for the missing value from the simulated regression model (Little 1988, Schenker and Taylor 1996). Rubin (2004, 168) also uses the term 'predictive mean hot deck imputation' that is one application of this method.

The predictive mean matching method requires the number of closest observations to be specified. A smaller number tends to increase the correlation among the multiple imputations for the missing observation and results in a higher variability of point estimates in repeated sampling. On the other hand, a larger number tends to lessen the effect from the imputation model and results in biased estimators.

This method is similar with our real-donor method in one respect. It gives only plausible values unlike linear regression methods cannot ensure it. The difference is in the imputation model that is basically the same as in Bayesian MI regression imputation method. Our imputation model specification for MI is simpler, that is, the sum of the predicted values and the normally distributed random term. This thus is a real-donor method in which the nearness indicator is estimated from the regression model. But its competitive indicator could be from the binary regression model or the propensity score model that follows Bayesian rules in SAS as explained below.

Propensity Score Method for Monotone Missing Data (SAS Manual 2015)

The propensity score method is another imputation method available for continuous variables when the data set has a monotone missing pattern. A propensity score is generally defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum and Rubin 1983). In the propensity score method, for a variable with missing values, a propensity score is generated for each observation to estimate the probability that the observation is missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap (ABB) imputation (e.g. Rubin 2004, 124, 136) is applied to each group. It divides the observations into a fixed number of groups (typically assumed to be five) based on these propensity scores (See the technical details from the SAS manual and more theory about Rubin's book9.

Our non-Bayesian method thus does not use Bayesian bootstrap. We take the predicted values of the binary regression model and use these for constructing imputation cells (groups) that is a fixed number, 14 in our test applications. The imputation model in SAS is the logistic regression that is one option in our applications but a probit and complementary log-log link is experienced as well.

It is important to point out that we also use the linear regression model for real-donor imputation. This option we have not found in SPSS or SAS but it has been used earlier (e.g. Laaksonen 2003). Sections 7 to 9 present comparable results both from Bayesian (SAS and SPSS) and non-Bayesian imputations for a fairly demanding data set that is explained in next section.

6. Test Data in More Details

The micro dataset of this study is close to real-life situation in an anonymous European country. Its initial version (Danish Labour Force Survey) has been already used in imputation tests of the Euredit project (Wagstaff 2003). For this study, a random sample of 19774 persons (10%) was drawn from the initial one. Some cleaning was made at the same time. For instance, the persons with zero income were deleted since this is not any plausible value in real life. The incomes were also exchangedinto Euros. The missingness mechanism was remained the same, since it was observed to be good in the Euredit experiments that tested both traditional and so-called new imputation methods. The main aim of these experiments was to compare methods in their biases using a number of evaluation criteria (Chambers 2003). The bias was possible to test at individual level since the true values were available after imputations. Euredit only tested its methods for single imputation. Also the variance estimates were left.

The number of missing values or the imputation size is 5315 (27%) that is fairly realistic. The data set consists of a quite good number of covariates which all except age are categorical. The age was however categorized for tests of this study. The full list with the number of categories that is used in all imputation models is as follows: gender (2), five-year age group (11), marriage (2), civil status (2), education level (4), region (12), Internet at home or not (2), socio-economic status (4), unemployed or not (2), children or not (2).

All these categorical variables are statistically significant for income, best ones being gender, education and socio-economic status. Nevertheless, the fit is fairly low since any individual level strong variable is not available. The R-square of the income model is 39 per cent.

The response indicator $R = 1$ if a value is observed, and = 0 if not observed (thus to be imputed). This indicator is needed for SAS MI propensity score method that this module creates automatically from the missing values. For our real-donor methods with binary regression models it should be created by a user, before the model estimation.

Sections 7 and 8 summarize the results of the different imputation methods. All test results are presented for the imputed units only, thus for 5315 persons. These results distinguish better the differences

in various methods, although the final estimates are made for the complete data set, that is, for 19774 persons.

7. Results for Point Estimates by Single and Multiple Imputation Methods

This section concentrates only on point estimates that are two in our test: mean income and the CV of income. This allows compare both pure SI methods and MI methods since the average of 10 imputations is used as a point estimate.

Note that the true mean income is 46606 and the income of the respondents 52857, respectively. The last average thus would be the imputed valuewith ‘mean imputation.’ This thus is a benchmarking value but another could be the estimate ‘RRD’ that is ‘random real-donor method’ or ‘random hot decking method’using the conventional terminology. Here a real donor has been borrowed at random from the list of the observed values. We thus expect that all results should be better than these benchmarking values. Otherwise, that imputation method should not be used.

We go next to details of these methods first presenting the performance of the single imputation methods.

The results that use linear regression are in Table 1. As deduced, this model-donor strategy may give impossible values since they are calculated straightforwardly. Of course, it is not known exactly which values are impossible but such are negative values, definitely. Both Bayesian and non-Bayesian methods lead to some such values. There are tools to correct for these values in all cases. SPSS offers an option for constraints of the variable being imputed. The software gives opportunity to put there a minimum and a maximum as well as the rounding option. We only used the minimum that we required to be zero. This option can be made for our method IMAI and SAS for example so that the negative values are removed to zero. In Table 1, the term ‘Constraints’ means this fairly subjective solution for SAS and IMAI but there is the special ‘constraints’ option for SPSS.

Table 1. Singlemodel-donor imputation results when the imputation model is linear regression.IMAI = Integrated Modelling Approach to Imputation.

Method	Average	Rate of negative values, %	CV, %
True	46606		65.1
Random real-donor	52922		63.2
SAS MI	50656	7.4	66.4
SAS MI Constraints	51625		61.5
SAS MCMC	48556	7.1	68.3
SAS MCMC Constraints	49582		63.3
SPSS MI	45959	8.2	71.9
SPSS Constraints	50297		56.5
SPSS MCMC	45701	8.0	68.3
SPSS MCMC Robust	50282		56.5
IMAI MI Robust No Constraints	46142	5.1	71.8
IMAI MI Robust Constraints	47380		65.1
IMAI MI No Robust No Constraints	46576	8.4	57.5
IMAI SI	46272	0.2	43.4

A big problem in interpreting results is a fairly big number of negative values that cannot be accepted in real-life. Without that several averages are fairly close to the true value including all IMAI methods as well as SPSS without constraints and SPSS MCMC without constraints. However, using constraints and robustness correction, the results will worsen. The best methods are now IMAI SI and IMAI MI with robustness and constraints. Since IMAI SI is very bad in CV, the latter can be considered as a best method since it is very good in CV. A strange thing is that that SPSS with constraints changes both estimates heavily but not in a correct direction. SAS MI looks good in CV but this cannot be used without any robustness correction due to negative values. Although those two IMAI methods are acceptable for averages, the income differences are too far from the true values. At contrary, the best IMAI for CV (MI Robust) is not very good for average but not either bad. This is thus considered to be the best overall method when using model-donor methodology by linear regression.

Bayesian methods are surprisingly bad either for both point estimates. Special attention should be paid to these methods without constraints or robustness correction. They produce a fairly big number of

negative values. It is possible that there are special tools to avoid them but these cannot be easily found by an ordinary user. It looks that the Bayesian algorithms do not work well with a demanding data set like this.

The negative imputed values can be avoided using log-transformation. Unfortunately, we found other problems with such imputed values and hence we do not present them here.

The third group of the results gives always plausible values since they borrow the imputed values from the observed ones. Table 2 summaries the findings.

Table 2. Single real-donor imputation results. Notation: PMM = Predictive Mean Matching, PSCORE = Propensity Score, Regression = Imputation model for real-donor search.

Method	Average	CV, %
SAS PMM	51428	62.4
SAS PSCORE	48547	64.6
SPSS PMM	37265	80.7
IMAI LOGIT SI	45890	65.3
IMAI LOGIT MI	47345	65.5
IMAI PROBIT SI	46801	64.8
IMAI PROBIT MI	46152	65.8
IMAI CLL SI	45563	65.6
IMAI CLL MI	46960	64.6
IMAI Regression	49341	65.1

Single imputation IMAI methods work best being relatively acceptable for both estimates, best being Probit link; this is more or less due to a good luck. IMAI multiple imputation methods are also good for income differences but not as good for averages. SAS propensity score method is the acceptable Bayesian method but only for income differences. The same method gives a too large average but not as large as the IMAI regression methods. It is not easy to interpret why IMAI Logit MI is upward biased for the average. One reason is that there were not in some score groups enough real-donors to borrow and they had to take from too far. This was not as problematic when using other two link functions.

However, it is hard to understand why Bayesian methods are so far from the true average, either too large or too small. CV's are not either good, being sometimes highly unbiased. SPSS predictive mean

matching works very badly, in particular. SAS results are biased as well but not as dramatically.

8. Results for Interval Estimates by Multiple Imputation Methods

It is possible to estimate variances for estimates when some values are imputed so that any MI methods are not used (e.g. Lee et al 2002), thus for singly imputed data. The basic component here is the imputation variance that is needed to add into the ordinary sampling variance. This section however presents results of MI based variances, both Bayesian and non-Bayesian. The variances are calculated applying both Rubin's and Björnstad's formulas. Since the variances are not illustrative, the results are presented as standard errors of the mean and as their relative versions respectively. The mean here is the income average for the imputed values. It should be noted that the results are not automatically interpreted since the standard error depends on the mean and thus on its bias respectively. We cannot know which standard error istrue, but a kind of the minimum standard error can be estimated from the true values. This is 416 and its relative version is 0.9 per cent, respectively.

Table 3 presents the results of all MI methods for the point estimates of Tables 1 and2. The methods are sorted here by theirrelative standard errors (both Rubin's and Björnstad's). This facilitates to interpret interval estimates especially for comparing Bayesian vs Non-Bayesian estimates with each other. It is not however automatic since they are conditional to the average estimates. Hence the table includes also the averages. It is good to check how biased the estimate is when interpreting the interval estimates.

Table 3. Interval estimates based on the two alternative formulas of multiple imputations. The two bolded results are for benchmarking. If Mark=B then the method is Bayesian, and if Mark=NB then non-Bayesian

			Standard error		Relative standard error	
Mark	Method	Average	Rubin	Björnstad	Rubin	Björnstad
T	True values	46606	416		0.90	
B	SAS REG ROBUST Model-donor	51625	597	654	1.15	1.26
NB	Random Real-donor	52922	613	666	1.16	1.26
B	SPSS MCMC REG ROBUST Model-donor	50282	619	658	1.23	1.31
NB	IMAI Regression Model-donor Robust	47380	598	650	1.26	1.37
B	SAS PMM Real-donor	51428	660	714	1.28	1.39
NB	IMAI Regression Real-donor	49341	644	693	1.30	1.40
B	SAS REG Model-donor	50656	664	720	1.31	1.42

B	SAS MCMC ROBUST Model-donor	49582	669	712	1.35	1.44
NB	IMAI Regression Model-donor	46102	642	699	1.39	1.51
NB	IMAI Probit Real-donor	46732	676	722	1.45	1.55
B	SAS MCMC Model-donor	48556	705	752	1.45	1.55
B	SPSS PMM Real-donor	37265	581	627	1.56	1.68
B	SPSS REG Model-donor	45959	719	764	1.56	1.66
B	SPSS REG ROBUST real-donor	50297	832	861	1.65	1.71
B	SAS PSCORE Real-donor	48547	822	857	1.69	1.77
NB	IMAI Cll Real-donor	47071	800	838	1.70	1.78
NB	IMAI Logit Real-donor	46713	841	877	1.80	1.88
B	SPSS MCMC REG Model-donor	45701	871	908	1.91	1.99

A big problem in interpreting the results is that so a point estimate of most Bayesian methods is too biased. Hence it is not easy to see what the impact of the Bayesian algorithm is in each interval estimate.

The view looks similar when using Björnstad's formula but all MI standard errors are somewhat larger. The non-Bayesian averages are often close to the true value, the Bayesian ones more far. It cannot be definitely said which standard errors are best but their variation is higher in Bayesian methods. If using Björnstad's formula, the standard errors are higher but the difference does not look dramatic from the practical points of view. The bias in some point estimates is more problematic.

9. Conclusion

Bayesian methods using standard software packages like SAS and SPPS can be problematic from the point of view of point estimates in particular. This does not mean that the theory and the principles of the Bayesian MI methods are not appropriate. But non-Bayesian MI methods seem to be less risky to apply since their algorithms are not so much like 'black boxes.' Imputation in general needs a good knowledge of the phenomenon and the data behind it, and this is not possible to understand by any automatic software package. Thus a user has to test different strategies, that is, different imputation models and imputation tasks, and then finally to choose an appropriate method for a particular estimation case.

The above conclusion is derived from the experiments for a demanding continuous variable in which case, the imputation model doesnot fit well. Correspondingly, the performance differences are remarkable. But what happens if the imputation model fits well? Are the results more consistent, respectively? This is tested using a modification of the same data set so that a proxy covariate was added into all the imputation

models. This variable improved the fit substantially, so that the R-Square for the linear regression model is 95 per cent. Such a good fit should give rise for the good performance in all imputation methods.

Moreover, we continue to compare Bayesian and non-Bayesian methods with each other. A conclusion of the performance of the methods is in Table 4. The table does not include all the methods of the previous tables when the model fit is poor. The missing methods do not give any essential value added. From the point of view of point estimates, their biases only are shown. Their estimates can be calculated using the true values of Table 1. Rubin's formula is used for the interval estimates but Björnstad's formula does not make any difference for the conclusions.

Table 4. MI Results using the test data in which the imputation model fits extremely well. The methods are sorted by the bias of the average income estimate. NB = Non-Bayesian, B=Bayesian as in Table 3.

	Method	Std. Error	Relative Std. error	Bias for Average, %	Bias for CV, %
NB	IMAI Regression Model-donor	464	1.00	-0.8	2.8
NB	IMAI Regression Real-donor	473	1.01	-0.7	0.5
B	SAS PMM Real-donor	478	1.03	-0.6	0.6
NB	IMAI Regression Model-donor Robust	462	1.00	-0.5	1.8
B	SAS REG Model-donor	468	1.01	-0.5	0.9
B	SPSS REG Model-donor	449	1.00	-0.4	0.9
B	SAS REG Model-donor ROBUST	467	1.01	-0.4	0.8
B	SPSS REG Model-donor Robust	447	1.00	-0.4	0.8
B	SPSS PMM Real-donor	440	0.95	-0.3	0.9
T	True values	-	-	0.0	0.0
NB	IMAI Logit Real-donor	591	1.26	0.3	-0.2
B	SAS PSCORE Real-donor	605	1.28	1.4	-1.4
NB	Random Real-donor	613	1.16	13.6	-2.9

The main conclusion is clearly that the imputation performances do not vary very much if the model fit is like here, extremely good. The biases are in most methods small, both for averages and income differences. The biases for averages are fairly small, and in both sides of the true value. The largest one is for SAS propensity score method that is due to a small number of imputation cells. The corresponding IMAI method (Logit Real-donor) is good for both point estimates. Instead, the standard errors of these two methods are largest and almost at the same level as for the benchmarking method, random real-donor. It is

obvious that these standard errors are too high.

Several methods give too high income differences. This drawback was earlier observed in poor imputation models as well but much more dramatically.

This additional imputation test gives a confirmation for the results of the major part of this study. ABayesian imputation method implemented into SAS and SPSS is not better than a corresponding non-Bayesian method. It is even worse for point estimates but these two approaches yield about the same quality for interval estimates although it is difficult to compare explicitly their quality.

Bayesian imputation calculations are more complex than the IMAI non-Bayesian ones. From the users points of view the software packages are easier to apply, SPSS being easier than SAS. On the other hand, the software's are not flexible to apply and to develop when comparing to the IMAI methods in which case, both the imputation model and the imputation task can be implemented by a user and with other tools than used in this study. Since we wished to compare all methods fairly, we used the same models in each. It would be possible to improve the imputations, using for example more complex models that are available in SAS MI procedures.

References

Allison, B.D. (2005).Imputation of Categorical Variables with PROC MI. SUGI 30 Proceedings.
http://www2.sas.com/proceedings/sugi30/113-30.pdf

Björnstad, J. (2007). Non-Bayesian Multiple Imputation. Journal of Official Statistics, 433–452.

Carpenter, J. and Kenward, M. (2013): Multiple Imputation and its Application. Wiley & Sons.

Chambers, R. (2003). Evaluation Criteria for Statistical Editing and Imputation. Euredit project. Papers.
http://www.cs.york.ac.uk/euredit/

Chambers, R.L., Hoogland, J., Laaksonen, S., Mesa, D.M., Pannekoek, J., Piela, P., Tsai, P. and de Waal, T. (2001). The AUTIMP-project: Evaluation of Imputation Software. Research Paper 0122. Statistics Netherlands.

Enders, C. K. (2010) Applied Missing Data Analysis. New York: Guildford Press.

IBM SPSS Missing Values. http://www.spss.ch/fr/logiciels/21/pdf/ [read June 2016]

Laaksonen, S. (2000). Regression-Based Nearest Neighbor Hot Decking. Computational Statistics. 15,1, 65-71.

Laaksonen, S. (2003). Alternative Imputation Techniques for Complex Metric Variables. The Journal of Applied Statistics, 1009-1020.

Laaksonen, S. (2016). A new framework for multiple imputation and applications to a binary variable. Model Assisted Statistics and Applications, 11.3, IOSPress.
http://content.iospress.com/journals/model-assisted-statistics-and-applications/11/2

Laaksonen, S., Rässler, S. and Skinner, C. (2004). Documentation of Pseudo Code of Imputation Methods for the Simulation Study. Dacseis Project Research Papers under Workpackage 11.2 'Imputation and Nonresponse'. 51 pp. www.dacseis.de/deliverables.

Lago, L.P. and Clark, R.G. (2015). Imputation of Household Survey Data using Linear Mixed Models. Australian and New Zealand Journal of Statistics 57, 169-187. doi: 10.1111/anzs.12108

Little R.J.A. (1988). Missing-Data Adjustments in Large Surveys. Journal of Business & Economic Statistics, Vol. 6, No. 3, 287-296.

Muñoz, J.F. and Rueda, M.M. (2009). New imputation methods for missing dada using quantiles. Journal of Computational and Applied Mathematics 232, 305-317.

Rosenbaum, P. R. and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika, 70, 41-55.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley & Sons: New York.

Rubin, D. (2004). Multiple Imputation for Nonresponse in Surveys. Wiley Classics Library Edition.

Rubin, D. (1996). Multiple Imputation After 18+ Years. Journal of American Statistical Association, 473-489.

SAS 9.4Help and Documentation, Users' Guide for the MI Procedure. Details.

Schenker, N. and Taylor, J. M. G. (1996), "Partially Parametric Techniques for Multiple Imputation," Computational Statistics and Data Analysis, 22, 425-446.

Statistical Solutions (2016) Mahalanobis Distance Matching Method.

http://www.statsols.com/mahalanobis-distance-matching-method/ Read June.

Wagstaff, H. (2003). Appendix B: Data Sets and Perturbations. Euredit project. Volume 2.

http://www.cs.york.ac.uk/euredit/

Published: Volume 2016, Issue 10 / October 25, 2016