Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

EXAMPLES OF MULTIPLE IMPUTATION

Nicholas T. Longford[1]

ABSTRACT

This paper discusses three examples of multiple imputation for incomplete data or information in large-scale surveys. Each example illustrates how a single-imputation procedure is improved by reflecting in the procedure the uncertainty about the missing values. The purpose of the paper is not to introduce any new methods or to prove any theoretical results, but to ease the practitioner's way to the routine application of multiple imputation in large-scale survey data analysis.

KEY WORDS: Auxiliary information, measurement error, missing information, multiple imputation, plausible values.

1. INTRODUCTION

Missing values are an unavoidable nuisance feature of most large-scale survey data. Collecting all the data as envisaged by a plan is in most surveys an unachievable ideal because the subjects cannot be regimented to obey the data collection protocol and its expectations, such as the person's availability, and possession and voluntary disclosure of the requested information. This should not discourage any efforts at reducing non-response and other sources of data incompleteness. However, ignoring the problem of incompleteness at the analysis stage, justifying it by claiming that all the efforts to collect the planned data completely have been expended, is not appropriate either.

To avoid lengthy preliminaries, we assume that the (survey) data to be analysed are collected by a specific sampling design and there is a well specified analysis agenda — a list of analyses, each to yield a pair comprising estimate $\hat{\theta}$ and its estimated sampling variance, $\hat{s}^2 = \hat{\text{var}}(\hat{\theta})$, or their multivariate version, vector $\hat{\boldsymbol{\theta}}$ and matrix $\hat{\text{var}}(\hat{\boldsymbol{\theta}})$. The snag is that these sample quantities would be straightforward to evaluate only if the data set collected by the survey were complete or had the same format as a complete data set — usually a rectangular array of n subjects and p completely recorded variables.

We refer to each $\hat{\theta}$ as a *complete-data estimator*; we assume that $\hat{\theta}$ would be unbiased and efficient if the data were complete. Further, we assume that $\hat{s}^2$ would be unbiased for $\text{var}(\hat{\theta})$ if the data were complete. A more rigorous notation, even if not standard, may avoid some lengthy and awkward expressions. We denote by $\mathbf{X}^*$ the complete data — the data set that was planned to have been collected, and by $\mathbf{X}$ the collected (recorded, or *incomplete*) data set. With each estimator $\hat{\theta}$ we associate the data set on which it is applied. Thus, given limited resources for data collection (conduct of the survey), obtaining the complete-data estimate $\hat{\theta}(\mathbf{X}^*)$ is the ideal. Its 'substitute', $\hat{\theta}(\mathbf{X})$, cannot be evaluated, unless we define suitable rules for the operations involving missing values. Having defined such rules, we should re-assess the efficiency of $\hat{\theta}(\mathbf{X})$.

Of the two natural ways of 'fixing up' the data so that it could be analysed by the methods (algorithms, computer programs, or the like) designed for complete data, data *reduction* and data *completion*, we entertain only the latter. In data reduction, the sample is reduced to the subjects who have complete records, resulting in a data set $\mathbf{X}_{\_}$. Discarding data from the subjects whose records are almost complete (e. g., only a small fraction of the p data items

[1]De Montfort University, James Went Bldg 2–8, The Gateway, Leicester, England, LE1 9BH (ntl@dmu.ac.uk).

is missing) represents a waste of information (collected at a non-trivial cost). Also, the sub-sample of the complete respondents may not be representative of the surveyed population, even if the entire sample is. In brief, $\hat{\theta}(\mathbf{X}_{-})$ may be quite inefficient (and biased) even when $\hat{\theta}(\mathbf{X}^{*})$ would have been efficient.

By completing the data set, by imputing a suitable value for each missing item, we generate a data set $\mathbf{X}_{+}$ that appears to contain more information than was in fact collected. Then $\hat{s}^2(\mathbf{X}_{+})$ will underestimate the sampling variance of $\hat{\theta}(\mathbf{X}_{+})$. Further, depending on the method of imputation, $\hat{\theta}(\mathbf{X}_{+})$ may be biased. The difficulties with analysing $\mathbf{X}_{+}$ arise because we have failed to inform $\hat{\theta}$ of the different status of the recorded and imputed values. Imputed values are guesses of the responses, and should be associated with uncertainty (variation) additional to the vagaries of the sampling process.

The route from the population to the (incomplete) data set leads through the sampling and non-response processes. Both processes reduce information; the first from the population to the complete data set, and the second from the complete data set to the incomplete data set. The sampling process is under our control (we define it by the sampling design), a key feature that enables us to find efficient estimators of the quantities of interest. The non-response process may have a similar description as a sampling process, but its details are not available. The incomplete data set contains no information about some of its features. We observe only the composition of the two processes, sampling (deliberate, designed in response to the limited resources available) and non-response (a nuisance, due to imperfect cooperation of the subjects).

To undo the damage (loss of information) caused by non-response, we should try to recover the inferences that would have been obtained with the complete data. This is the motivation for both single and multiple imputation. Single imputation (SI) attempts to recover the complete data by doing the best that can be done for each missing value. In multiple imputation (MI), recovery of the complete data set is secondary to the goals of

1. efficient estimation of θ, by estimating $\hat{\theta}(\mathbf{X}^{*})$, and
2. unbiased estimation of the sampling variance with respect to the composition of the sampling and non-response processes.

Reflecting the uncertainty about the missing values is a key feature of MI.

Hybrid methods that impute only for some of the missing items and reduce the completed data set to have the standard format, inherit the deficiencies of both reduction and SI methods.

1.2 The arithmetic of uncertainty

The deficiencies of any SI procedure can be illustrated on the following elementary example. Suppose a survey data set contains missing values in only one variable. The application of a SI method to a specific missing item *x* yields the value of 0. Suppose we are quite certain that the response would have been 0, although +1 and –1 are distinct possibilities. This might be expressed by associating the value of 0 with probability 0.8, and the values Γ1with 0.1 each. As the problem is presented, it may be hard to argue against imputing any value other than zero. However, once we consider the operations to which the value will be subjected, the deficiency of the choice becomes obvious. If only linear functions of this value are used for evaluating $\hat{\theta}$ and $\hat{s}^2$, the choice of 0 is appropriate. But if the value is used in a quadratic function, such as $\sum_i x_i^2$, the choice appears to be irrational — we would substitute the smallest possible value for x^2! The expectation of the distribution derived for x^2, equal to 0.2, is clearly the 'right' choice. But how should we to reconcile the paradox $0^2 = 0.2$?

The answer is at hand, in the identity $\mathrm{E}(x^2) = \{\mathrm{E}(x)\}^2 + \mathrm{var}(x)$ for any random variable *x*. Or, more generally, that $\mathrm{E}\{f(x)\} \neq f\{\mathrm{E}(x)\}$ for most non-linear functions f (e. g., all strictly convex and concave functions *f*) when *x* is not

degenerate. Thus, if we impute a value for a given item, it may be suitable for the use with some functions, but will not be suitable with (most) others.

If we 'fix up' the data by suitable imputations for one function, we will fail to fix it for another function. Since the estimators of a population quantity and of its sampling variance usually involve different functions of the data (e. g., linear for the estimator and quadratic for the sampling variance), we cannot fix up the data even for a single analysis. One of the quantities evaluated, usually $\hat{s}^2$, will need some adjustment. For survey data that are subjected to numerous analyses, of several types and varied complexity, adjusting each analysis is not a viable proposition, especially with analysts who have limited expertise in the issues involved and do not have software tools and other equipment to conduct non-standard analyses.

These arguments can be rephrased in the language of the EM algorithm (Dempster, Laird and Rubin, 1977), although it is formulated for maximum likelihood (ML); some estimators used in surveys cannot be easily expressed as ML estimators. In the E step of the EM algorithm, the contributions of the missing values to the complete-data log-likelihood are estimated by their conditional expectations given the data and current parameter estimates. The estimates of the missing values (as their conditional expectations), would yield a different estimate of the contribution, and lead to an incomplete-data estimator that is not efficient. Although the EM algorithm is applicable for problems with missing values, it is poorly suited for analysis of survey data with an extensive analysis agenda because a large number of conditional expectations have to be evaluated.

2. MULTIPLE IMPUTATION

In MI, a number of plausible (alternative) completions of $\mathbf{X}$ is generated by a random process that reflects the uncertainty about the missing values. The process comprises four steps:

1. fitting a model for the missing values;
2. generating (simulating) several (*M*) sets of plausible values from the model;
3. analysing the completed data sets;
4. summarising the completed-data results.

The assumptions of the method are that

a. the model for the missing values is correctly specified;
b. the plausible values are generated properly;
c. the complete-data analysis involves an efficient and unbiased estimator $\hat{\theta}(\mathbf{X}^*)$ with its sampling variance with respect to the sampling process, $s^2(\mathbf{X}^*) = \text{var}\{\hat{\theta}(\mathbf{X}^*)\}$, estimated without bias, by $\hat{s}^2(\mathbf{X}^*)$;
d. the sampling variance of the estimator $\hat{s}^2(\mathbf{X}^*)$ is of smaller order of magnitude than s^4.

Proper imputation in b. means that the sources of uncertainty about the parameters in the model for missing values as well as about the missing values given the model parameters are reflected in the process of generating the plausible values. In most settings, this entails two sources of uncertainty: about parameters of the posited model for non-response and about the outcomes given the values of the model parameters and covariates.

Denote by $\mathbf{X}_m$ the data set completed by the *m*-th set of plausible values, and by $\hat{\theta}_m = \hat{\theta}(\mathbf{X}_m)$ the estimate based on the *m*-th completed data set. If the assumptions a.–d. are satisfied the MI-estimator

$$\tilde{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

is approximately unbiased and its sampling variance is estimated with at most small bias by

$$\tilde{s}^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{s}^2(\mathbf{X}_m) + \sum_{m=1}^{M}(\hat{\theta}_m - \tilde{\theta})^2$$

(Rubin, 1987).

Except for a., the assumptions are natural and usually satisfied, although condition b. prescribes how to generate the plausible values. Assumption a. is widely quoted as a hindrance to an effective application of MI. In particular, the assumption of missing at random (MAR), on which most MI procedures are based, is quoted as difficult to justify. After all, SI is associated with no such assumption explicitly. This argument is fallacious. A SI is equivalent to *M* identical imputations in MI, implying a model with no between-imputation variance, that is, no uncertainty about the missing values. In most settings, this is a very poor choice of a model for imputation and it destroys the strategy of using SI until the model for missing values can be specified correctly. Using SI does not absolve us from concerns about MAR and the appropriate conditioning (model). Its specification is a task common to SI and MI.

To combat the reluctance to use MI, we advocate a strategy of improvement on the default SI method by formulating the model it implies. A simple generic example of this approach is to replace the 'bring the last value forward' (BLVF) method for a longitudinal dichotomous variable with (multiple) draws from a model that assumes a (small) probability of change from one time point to the next. The probability is estimated from the complete records, and each set of plausible values is generated as draws from a plausible probability. A plausible probability is drawn from the (estimated) distribution of the estimator of the probability. In most settings, normality of the estimator involves an acceptable approximation. In this MI procedure, the sets of plausible values (to be imputed) differ in two aspects: due to the vagaries of the Bernoulli process and because the probabilities differ between the sets (imputations). Concerns about MAR can be addressed by introducing more detail in the model: assuming that the probabilities differ for groups of subjects, across the time points, depend not only on the previous but also on earlier outcomes (when available), and so on. With a multinomial variable, the conditional distribution of the new state (given change from the previous time point) can be modelled similarly.

How detailed a model should we formulate? (How far should we go with conditioning or stratification?) Although Rubin (1996) recommends as detailed a conditioning as possible, this should not be regarded as an imperative. It is much better to have less conditioning than no conditioning at all. So, the fear of not satisfying an ultimatum should not turn anybody away. In any case, there comes a point when the model for missing data is so rich in parameters that there is just too much uncertainty about the missing values. The analyst's comfort with all the details should not be neglected.

The following three sections outline applications of MI with the perspective described above. Each section is a condensed version of a paper describing a case study.

3. THE UK LABOUR FORCE SURVEY

The UK Labour Force Survey (LFS) is a continual survey of residential addresses in the UK. Each address selected to the sample at a time point is retained for one year, and contacted soon after selection and 3, 6, 9 and 12 months later. The key outcome variable recorded is the employment status, as defined by the International Labour Organisation (ILO). It has four categories: child (CH, aged 1–15 years, inclusive), employed (EM), unemployed (UN), and economically inactive (IA). The media headline analysis based on the survey is the estimated rate of unemployment among the working-age residents in the country (aged 16–64 for men and 16–59 for women).

Although the elementary sampling unit is an address, the information is collected from (and about) individuals. Thus, an address can have the same, slightly altered, or a completely different list of occupants from one contact to the next. There is a narrow window of two weeks when an address can be contacted, so as to satisfy a tight publication schedule. Thus, it may be impossible to establish whether an address is unoccupied or the interviewer failed to make a contact with the residents. Of course, subjects can refuse any (further) cooperation with the survey on behalf of some or all the residents at the address. A well-established protocol contributes to the reduction of item-level non-

response. Response by proxy is accepted (from an adult), so data may be collected about a subject not at home at the time of the interview. Sometimes no proxy respondent is available.

In the operation, the survey uses imputation for the ILO employment status by BLVF, if the status was established three months ago. In particular, no imputation is applied for missing status at the first planned contact. The database is organised by quarters, containing the records of all subjects who were contacted (directly or by proxy) in the particular quarter. The basic socio-demographic variables are recorded for all the subjects in the database (otherwise they are not included). A quarter's database contains about 140,000 individuals, 80,000 of whom are of working age. Among the 26,369 subjects in the database for March–May 2001 whose addresses were in the follow-up for the fifth time, the status was imputed for 1418 subjects (5.4%) and was left missing for 562 subjects (2.1%).

3.1 From BLVF to hot deck

We are concerned with imputation for missing ILO status of adult residents. The model implied by BLVF is that of transition among the three states (EM, UN and IA), with high probabilities of no change. A simple inspection of the data reveals that changes of the employment status among young adults 16–24 are much more frequent (10–20%, depending on how the subpopulation is defined) than among middle-aged or elderly. The non-response rate of the young adults is much higher than the overall rate, and the non-response rate of elderly is very low.

The first step in improving on BLVF is to estimate the transition probabilities among the three states EM, UN and IA. The stratification by age provides an obvious improvement; the sets of probabilities differ among the age groups (16–24, 25–39 and 40+). Insubstantial analytical complexity is introduced by additional stratification on sex and marital status (single or not). However, these factors are not as useful as earlier employment status (two, three and four quarters earlier), when available. Since the number of strata (categories) thus introduced may be excessive, the earlier states can be collapsed by counting the number of changes or classifying subjects to those who have had a change while in the survey and those who had not. As a convention, a missing status at an earlier time point is regarded as a change.

A practical implementation of such an imputation scheme is by hot deck. Within each stratum (combination of the categories), we identify a pool of donors — subjects with the particular combination whose current status is established, and the recipients — subjects with the same combination whose current status has not been recorded. For each recipient we draw at random a donor and his/her status is imputed for the recipient.

Hot deck is a random process of drawing a value of the status with the probabilities equal to the representation in the donor pool. The deficiency of this process is that the *estimated* multinomial probabilities are used instead of the underlying probabilities. This is remedied by drawing the probabilities from their estimated sampling distribution, using the normal approximation. This ensures that the plausible values have appropriate between-imputation variation, in addition to the differences due to random drawing from a multinomial distribution. See Longford (2002a) for further background and details of the imputation scheme and analysis (small-area statistics).

Arguably, imputation for missing status in the LFS database should be regarded as having lower priority than dealing with address-level non-response. When a sampled address is not contacted it does not appear in the database. The rate of address-level non-response exceeds 30%. It is easy to establish that young and single people are under-represented in the LFS database. Since their UN rate is higher than the rates for the middle-aged and elderly, the non-response brings about a bias in several important analyses. A MI procedure for dealing with address-level non-response would impute entire households for each non-contacted address, catering for the possibility that the address is not occupied, is occupied by the same residents as three months ago, by an altered set of residents, or by an entirely different set. Setting up such a model represents a much greater challenge, but that is the price for approaching the ideal of efficient estimation (using all the available information) and honest assessment of the precision (unbiased estimation of the sampling variance). This includes distinguishing between the planned and realised sampling designs. An integral part of this effort is modelling the dominant nuisance features of the data collection, among which address-level non-response stands out.

4. NATIONAL SURVEY OF HEALTH AND DEVELOPMENT ALCOHOL CONSUMPTION

The National Survey of Health and Development, funded by the Medical Research Council (UK), is a longitudinal survey of subjects born in Great Britain in the week March 3rd–9th, 1946. The original sample included all single legitimate births to wives of non-manual or agricultural workers and a 1:4 simple random sample of the single legitimate births to wives of manual workers. The subjects were followed up regularly from birth through adolescence and adulthood, until in 1989 they were contacted for the 19th time. The 20th follow-up has taken place recently. The original sample of 5362 has over the years been reduced by loss of contact, refusal, emigration and death to 3262 (61%).

Among other questionnaire items, the subjects were requested in 1989 to complete a one-week diary of all food and drink they consumed. The diary was completed for the first two days during the visit by a health professional, although some subjects either refused to cooperate, or the interview was stopped for another reason while completing the diary, or even earlier. The focus of the study described in Longford *et al.* (2000) is alcohol consumption, and consumption in excess in particular. The study narrowed down the concern about incomplete data to the universe represented by the 3262 partially or fully cooperating subjects. Ignoring the subjects lost in the earlier follow-ups is not appropriate, although those who have died or have emigrated can be regarded as no longer belonging to the population of interest. But reducing the analysis to the 2002 subjects with complete diaries would be even less appropriate.

Alcohol is consumed according to a variety of patterns — never, occasionally in small quantities, in bouts of drinking of varying frequency, regularly in a range of quantities, and the like. Thus, we can 'learn' from a short segment of the diary about what the consumption is likely to be on other days. Further, there is a fair amount of good auxiliary information — variables highly correlated with alcohol consumption. Sex and body mass are obvious choices, although it may be profitable to look for auxiliary information beyond the variables we would consider as covariates in a traditional regression setting. In NSHD in 1989, the subjects were also asked to recall how much of four types of alcoholic beverages (beer, wine, sherry, and liqueurs) they had consumed during the previous week.
Such a set of questions is responded within a very short time, after a cursory mental recall. So the responses are less reliable, and there is ample evidence of under-reporting, but non-response is rare. On the other hand, diary data are much more reliable when they are recorded completely. So, while the recall data are not a very good substitute for the missing diary records (especially for non-empty incomplete ones), they are useful as auxiliary variables — effective *informants*. The practical way of exploiting this association is by formulating a multivariate model for the quantities at recall and on the seven diary days, and drawing plausible values for the 'missing' days from the plausible conditional distribution of the missing diary quantities given the recorded part of the diary and recall quantities. Regression can be introduced in this multivariate model by a stratification on other relevant variables: body mass, smoking status, and CAGE, a four-item questionnaire about problems with alcohol.

In the study, the generation of plausible values was organised in stages. First, plausible values were generated for height and body mass. Next, plausible values were generated for the recall, conditioning on body mass, CAGE and smoking status. Since the recall quantities have a distribution well approximated by a log-normal distribution with boosted zeros, the sign (zero or positive) of the consumption and the log-quantity (ignored if the sign is zero) were generated separately, using distinct four-variate models (for the four types of beverages).

Finally, plausible signs and log-quantities were generated for the alcohol consumed on each missed day. See Longford *et al.* (2000) for details. Rubin (1996) and Schafer (1997) argue that it is not essential for the data used as outcomes in a model for missing values to be normally distributed. By taking their advice, the process of generating the plausible values could be simplified somewhat.

The goal of the study was to estimate the percentage of middle-aged Britons who consume alcohol in excess. In the analysis, we have to ignore the week-to-week variation in alcohol consumption, as well as the systematic differences among the 43-year-olds and those in the neighbouring age cohorts. Depending on how we specify excessive consumption (possibly by different quantities for men and women, or quantities specific to body mass), an

incomplete record may inform about the outcome of interest with certainty — if a subject consumed more than the weekly 'quota' of alcohol on the first two days, the consumption on the remaining days is immaterial.

Trivial SI methods, such as BLVF, mean imputation, imputing zero for each missing value, used in similar settings, are not appropriate, and their deficiencies show up particularly clearly in estimating a tail probability, as in our case. Details are documented in Ely (2003).

Imputation for the subjects lost in the earlier follow-ups presents a considerable logistical challenge of sifting through extensive (incomplete) information, some of it not in electronic form or in a computer format that is difficult to process. We have to exclude subjects who have died or have emigrated because they are no longer members of the population of interest. Note that if a subject returns to the UK from emigration, he/she does not rejoin the Survey. The status (living in the UK, died or emigrated) of those subjects with whom contact has been lost is not usually known. The problem of missing not at random in addressing such non-response is acute, and both extensive conditioning on information from previous follow-ups and sensitivity analysis are essential.

5. THE SCOTTISH HOUSE CONDITION SURVEY

The Scottish House Condition Survey (SHCS) is a survey of the housing stock in Scotland. It was conducted in 1991, 1996 and 2002. The 1996 survey employed a stratified clustered design, with adjustments (boosts) that ensure sufficient subsample sizes in certain geographical areas. SHCS engages professional surveyors to assess the sampled dwelling units. The assessment comprises a large number of *elements* (items), such as dwelling type, presence of central heating and the extent of disrepair of various parts and features of the dwelling. The latter variables are scored on an 11-point scale (0–10), interpreted as 0, 10, …, 100% of the replacement cost required to bring the part/feature to the established standard. These scores are converted to total cost for the dwelling (*visible repair cost*) and the cost required to maintain the standard for the next 10 years (*comprehensive repair cost*). This conversion is based on extensive tables and formulae that take into account the size, type and location of the dwelling and economies of scale (savings when a lot of repairs are required), and the like.

The assessment by the surveyors, although of high quality, is not perfect. As a form of quality control, a non-informative sub-sample of the dwellings was surveyed second time, and the pairs of surveys compared. The differences between the assessments of a dwelling raise a concern about the estimated precision of the key reported estimators, because they are evaluated assuming that the survey assessments are perfect. The relative importance of the differences is difficult to quantify because for some elements disagreements are more frequent and/or tend to be greater, whereas for others the impact on the cost is more substantial.

The sample size of SHCS in 1996 was about 16,000, with 575 dwellings surveyed twice each.

5.1 Misclassification and MI

For each element and dwelling, we define the *ideal* assessment — what the assessment would have been with a perfect surveyor, and regard it as the missing information. The realised assessments are high-quality auxiliary information about the ideal assessments, and the pairs of assessments are useful for defining an appropriate model from which plausible values (assessments) are generated.

For an ordinal scale we define the neighbourhood of a score in the natural way. For instance, for scores 0– 10, the neighbours of $0 < k < 10$ comprise k–1 and k+1. We consider two kinds of imperfections in the surveyors' assessments:

- discrepancy, resulting in assigning the inspected element to a category neighbouring the ideal one;
- gross error, resulting in assigning the inspected element to an arbitrary category.

The element-specific probabilities associated with these two kinds of imperfections are denoted by p_d and p_g, respectively. Prior information suggests that p_g is much smaller than p_d, and the latter is no greater than a few per cent. The assessment of an element may be subject to both discrepancy and gross error. If a category k has L_k neighbours the probability of assigning a dwelling to the ideal category is

$$\mathrm{P}(X = X^* \mid X^* = k) = 1 - L_k p_k - (K-1) p_g + (K-1) p_d p_g ,$$

where X denotes the realised assessment and X^* the ideal assessment. The probabilities p_d and p_g are estimated, separately for each element that contributes to the assessed cost, by moment matching (as the root of a quadratic equation) and their sampling variances from the Taylor expansion; for details, see Longford (2002b). For generating a set of plausible ideal scores, we require a plausible conditional distribution of the ideal score given the assigned score, $\mathrm{P}(X^* \mid X)$. This is obtained by the Bayes theorem, using a plausible pair of probabilities $(\tilde{p}_d, \tilde{p}_g)$. A set of plausible (ideal) scores are obtained by simulating discrepancies and gross errors according to $\tilde{p}_d$ and $\tilde{p}_g$, respectively. The simulations are independent for each of the 50+ variables involved in the cost calculations. The plausible costs are evaluated from each set of plausible scores (and a few other variables). The population quantities of key interest are certain large sub-domain totals, such as the total comprehensive repair cost for the dwelling in each type of dwelling (detached, semi-detached, block of flats, etc.) and in small areas (local administrative authorities). In some analyses the outcome variable (cost) is on the original scale, in others on the logarithmic scale. Without MI, different methods would have to be developed for the two kinds of analyses. With MI, the plausible values generated by the data constructor are for universal application — for *any* analysis.

The between-imputation variance informs us about the inflation of the sampling variance due to imperfect assessment. While M=5 sets of plausible values might be sufficient for evaluating the MI estimator, it is advisable to use more sets when the focus is on planning a future survey.

The relative contributions of the elements to the between-imputation variance can be explored by the following approach. The plausible values of the focal element are replaced by the realised assessments, and the MI estimation conducted pretending that the assessments on this element are perfect. The estimated sampling variance, or its between-imputation component, is then compared with the estimated sampling variance of the original MI estimator. Similarly, the plausible values may be applied for only one element.

The M sets of plausible scores are sizeable data sets, but they are much smaller than the original database which contains many variables, most of them not involved in the cost formula.

6. PERIPHERAL ISSUES

Not so many years ago, the concerns about the data storage and the amount of computing required to implement the first two steps of an MI procedure might have been well founded. Since then, computing power and storage capacity have become so cheap and abundant that these concerns can be dismissed for all but extremely large databases. The extra data storage required in MI is for M sets of plausible values, possibly labelled by the subject and variable involved. For instance, for M=5 and 14% of the data items missing, the M additional data sets contain as many items as the original (incomplete) database. An organisation of the database much more friendly to a secondary user with limited software equipment is to provide the M completed data sets. When imputations are desired only for a few variables, their plausible values can be attached to the database as additional variables. Another convenient way of providing the plausible values to secondary users is in the form of a program that would generate a set of plausible values and impute them to complete the database.

Having to apply a (complete-data) procedure M times requires much less resources than the M-multiple of the resources required for a single analysis, because the main expense is on the design, construction and debugging of the program. In fact, the main virtue of MI is that the secondary users require no software tools or expertise other

than to apply the complete-data analysis. The computing or the real time expended by the M–1 additional analyses is not a factor of any importance.

Apart from an impact on the quantity of computing, the number of sets of plausible values has an impact on the sampling variance of the MI estimator. This variance is $W + B(1+1/M)$, where W is the within-imputation variance, the estimate of the complete-data sampling variance, and B is the between-imputation variance, interpretable as the amount of information lost due to data incompleteness. Further loss, B/M, is due to using only M imputations. In most surveys M=5 is sufficient, but the reasons for not having more imputations are well founded only for huge databases with large fractions of missing information.

7. CONCLUSION

MI is a practical method for dealing with incomplete information in large-scale surveys. In ideal circumstances, it is efficient and honest; it uses all the information in the incomplete records and assesses the precision of the estimators with little or no bias. When not in the ideal circumstances of having *the* correct model for non-response, we do not fare as well, but with the standard tools for modelling we have much greater opportunities to get close to the ideal. With any SI method we are much further behind because every SI method can be improved by making it MI — by introducing the sources of uncertainty about the imputed values in their generation, and replicating the process a few times. In some special cases, the between-imputation variance can be estimated directly and expressed (or approximated) analytically. This is a feasible proposition when only a few analyses are to be carried out. When a wide variety of complete-data methods are applied, MI is superior by delegating the work to the computer.

The scope of problems for which MI is applicable covers the entire missing-information agenda, as outlined in Dempster, Laird and Rubin (1977). MI involves a residual inefficiency because of a finite number of imputations, but relieves the secondary analyst from requiring expertise in handling missing data. In contrast, the EM algorithm is efficient, without any qualification other than some mild regularity conditions, but is specific to an analysis, is much more difficult to implement, and convergence problems may be encountered in its application.

REFERENCES

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood for incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B, 39, pp. 1–38.

Ely, M. (2003), "Comparison of methods for dealing with missing values in the National Survey of Development and Health", PhD. thesis. In preparation.

Longford, N. T. (2002a), "Missing data and small area estimation in the UK Labour Force Survey." Submitted.

Longford, N. T. (2002b), "Surveyor inconsistency in the Scottish House Condition Survey", unpublished manuscript.

Longford, N. T., Ely, M., Hardy, R., and Wadsworth, M. E. J. (2000), "Handling missing data in diaries of alcohol consumption", *Journal of the Royal Statistical Society*, Series A, 163, pp. 381–402.

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

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

Schafer, J. L. (1997), "Analysis of Incomplete Multivariate Data", New York: Chapman and Hall.