Imputation in Surveys: Coping with Reality
Author(s): Innis G. Sande
Source: *The American Statistician*, Vol. 36, No. 3, Part 1 (Aug., 1982), pp. 145-152
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2683165
Accessed: 21-10-2015 17:28 UTC

Imputation in Surveys: Coping With Reality

INNIS G. SANDE*

In surveys a response may be incomplete or some items may be inconsistent or, as in the case of two-phase sampling, items may be unavailable. In these cases it may be expedient to impute values for the missing items. This article discusses the general problem of nonresponse and the problem of imputation. Methods of imputation are reviewed, and the evaluation of the effects of imputation on the survey estimates and the monitoring of the imputation process are discussed.

KEY WORDS: Imputation; Nonresponse.

1. INTRODUCTION

Textbooks on sampling theory rarely address the problems that arise when more than one item is collected in a survey. The assumption is that the respondent will answer all questions correctly and completely, or else will not respond at all. In fact, as the survey statistician soon finds out, respondents may often answer only some of the questions, resulting in a phenomenon known as partial nonresponse.

The obvious "cure" for partial nonresponse is to contact the respondent again and clear up the problem. Unfortunately, this is often impossible, impractical, or too expensive: the respondent may not know or may not wish to give the required information, there may be too many cases for an effective 100 percent call-back program, and all too often there are serious limitations of time and money. Furthermore, problems with the data are sometimes discovered when the survey processing is well advanced and it is too late to "stop the presses" while the respondent is consulted.

Imputation is the estimation of individual items missing in a survey response; that is, imputation "completes" an incomplete response. In recent years there has been a growing amount of theoretical and empirical work done in the area. This author does not attempt to survey all this activity, but attempts to give a (somewhat personal) guide to the options and the problems associated with them.

In the author's experience, surveys often involve interrelated items, and the choice of techniques for imputing in this situation involves considerations somewhat different from those when only a single variable is being imputed. When the real situation is very complex, as it often is, pragmatism becomes an important, if not overriding, factor in dealing with partial nonresponse.

In Section 2 the general problem of partial nonresponse is discussed. In Section 3 we discuss the problems facing the imputer. Methods of imputation are reviewed in Section 4. In Section 5 we touch on the issue of evaluating the effects of imputation on the survey estimates.

2. THE PROBLEM

Partial nonresponse arises in two ways:

1. A record (i.e., the total response for a single survey unit) contains one or more missing values because, after all possible checking and follow-up, the data are unavailable.
2. A record is inconsistent in the sense that its component items do not satisfy natural or reasonable constraints, known as edits, and one or more items are designated unacceptable and therefore are artificially "missing."

A realistic survey response file might look like Table 1. Each record has identification and classification items that come from the sampling frame, a sample weight, and the five variables collected by the survey. There are both numeric and categorical variables and these must satisfy the following constraints or edits:

1. If $V1 = A$ and $V2 = a$ then $V3 \neq x$,
2. If $V1 = B$ and $V2 = b$ then $V3 \neq y$,
3. $V4 + V5 \leq 10$,
4. $V4 \geq 0$, $V5 \geq 0$.

We notice that of 10 records, four (1, 2, 4, 5) are complete. If we look hard, we might also notice that the "missings" are informative: low values of $V5$ are associated with missing $V4$'s (3, 8, 10).

Our primary problem is that we have to produce tabulations of population estimates, for example, $V1 \times V2 \times$ Classification variables, or $V4 \times$ Classification variables. Although we might be able to write down all the estimates we think we have a need for in our publication, we know that after the publication comes out we are going to get a large number of requests for tabulations and estimates that we have not anticipated.

How, then, are we to deal with the partial nonresponse, assuming all possible follow-up and calling back of respondents have been done? The possibilities are

1. Ignore all the records with missing values. This may result in loss of a great deal of data, since many records may be affected. Furthermore, "missings" are seldom random and the procedure would almost certainly lead to biased estimates. For example, if only the complete responses were used for estimation in the data given in Table 1, only 40 percent of the responses would be used, and no responses with $V1 = C$ would be in-

*Innis G. Sande is Senior Methodologist, Business Survey Methods Division, Statistics Canada, Ottawa K1A 0T6, Canada.

Table 1. Important Canadian Survey

Record Number	Identification Classification	Weight	Variables V1	V2	V3	V4	V5
1	X	W_1	A	a	y	3.1	4.3
2	X	W_2	A	a	z	4.6	2.8
3	X	W_3	A	b	y	—	1.1
4	X	W_4	B	b	z	2.3	4.6
5	X	W_5	B	c	y	4.9	2.3
6	X	W_6	B	b	—	3.2	3.6
7	X	W_7	C	—	x	3.0	—
8	X	W_8	C	—	y	—	1.2
9	X	W_9	C	a	—	.0	2.4
10	X	W_{10}	—	b	y	—	1.4

cluded, so that the estimated proportion of these cases would be zero.

2. Publish "unknowns" as a category. This is better than possibility (1); but still ignores the partial information about the missing value that may be available in the other variables. Frequently, the users of the data will make adjustments for the "unknown" categories without being able to look at the microdata and with little knowledge of the data collection process.

3. Adjust (reweight) each table or estimate, ignoring the missings in each case. This is a variation of possibility (1) that that may give rise to inconsistent tables or estimates in the sense that no complete data set corresponds to the set of estimates because of the constraints on the data.

For example, suppose a survey collects data on three variables, $V1$, $V2$, and $V3$, with the constraint $V1 + V2 \leq V3$. We would expect the population averages, $AV1$, $AV2$, and $AV3$, to satisfy the same constraint. Now suppose the survey data looks like Table 2. If we estimate the population average of $V2$ using only the first record, we estimate $A\hat{V}2 = 5$. Similarly, $A\hat{V}3 = 5$. Of course, $A\hat{V}1 = 1$. The constraint is then violated for the estimates, that is, $A\hat{V}1 + A\hat{V}2$ is not less than or equal to $A\hat{V}3$, and no set of complete data (where all records would satisfy the constraint $V1 + V2 \leq V3$) could exist that would result in this set of estimates.

4. Fill in the blanks with plausible (i.e., realistic) and consistent values. This is called imputation. In Table 2, the missing value of $V2$ in the first record would have to be imputed by a value between 0 and 4, while an imputation for the missing value of $V3$ in the second record would have to be at least 6.

The estimation of individual values in a data set is not a new problem. It is the direct descendant of the "missing observation" problem in ANOVA and the "incomplete data" problem in multivariate analysis. However, these problems have been directed towards the estimation of specific parameters under specific models and not towards the general problem of providing estimates of arbitrary parameters "after the fact."

Table 2. Survey of Widget Producers

Record Number	Weight	V1	V2	V3
1	10	1	—	5
2	10	1	5	—

3. THE GENERAL IMPUTATION PROBLEM

What are the "facts of life" facing the unwilling imputer? No matter what method of imputation he or she opts for, the following problems must be dealt with:

1. The close relationship between editing and imputation.

a. If a record fails an edit it is not always obvious which fields are faulty, but some basis must be established for deciding which fields to change. Does one change all the fields involved in a failed edit? Some of them may be involved in other edits that do not fail. Does one change the least number of items, as recommended by Fellegi and Holt (1976), or adopt a policy of "least change," whatever that means? Or does one adopt the "principle of expedience": deleting that configuration that makes imputation easy?

These are not trivial problems. The mathematical analysis of edits and the identification of fields to be changed when several edits have been failed is a very subtle problem. Fellegi and Holt did the first systematic work on categorical or coded data and their methods have been implemented at Statistics Canada and used (with modifications; see Hill 1978) in the Census of Population. The parallel work for numerical data with linear edits was carried out by Gordon Sande at Statistics Canada using optimization techniques (Sande 1979), and the development of techniques for the combined numerical and categorical data problem is seen as feasible.

b. When it has been decided which fields must be imputed (because they are missing or must be changed), it is obvious that the imputed data must satisfy the edits; that is, the completed record must be consistent. This requirement seriously reduces the applicability of some imputation procedures.

2. The marginal and joint distributions of responses are almost certainly different from those of the parent population. In the case of numeric data, such distributions are unlikely to be normal. Transformations to normality (or just less pronounced skewness) usually result in transformations of the edits that make them more difficult to deal with.

3. The pattern of missing fields varies from record to record. In an n-field record (excluding the identifiers and classification variables), there are $2^n - 1$ possible patterns of fields to impute. Some imputation schemes (the author does not know if any have been seriously implemented) seek to specify a separate imputation procedure for each pattern; but if n is large, this idea soon gets out of hand.

4. The imputer does not usually have much time to fiddle with the data after they have come in. Most survey data should be processed promptly to be useful and in some cases (such as many at Statistics Canada) the

time constraints are severe. Therefore the method of imputation should be precisely specified, so that no further experimentation with the data is necessary before the processing begins. Furthermore, the statistician usually has little, if any, test data to work on before the data collection begins. Historic data cannot always be trusted to look like current data in any but the most general aspects. For example we may believe that X is proportional to Y on the basis of historic data; but the proportion X/Y may change from year to year. On the other hand, the circumstances governing the joint occurrence or nonoccurrence of X and Y may be similar over time, a fact that can be exploited in testing imputation procedures.

5. Imputation does not solve any specific estimation problem more satisfactorily than classical estimation techniques for incomplete data, and it may do a lot worse. The trouble is that there is usually a very large number of parameters to be estimated, and the mathematical problem, assuming one has a correct model to work with, is formidable. If one estimates one subset of these parameters at a time, there is, in general, no guarantee that the resulting set of of estimates would be consistent in the sense that a complete data set might exist that would give rise to the same set of estimates. The problem is compounded when estimates for new parameters are required (as alluded to at the end of Section 2).

By imputing a consistent (in the sense that the edits are satisfied) value for each missing item, one can estimate any of the usual population parameters (means, totals, ratios, differences, proportions, correlations) very easily, although possibly without specified precision (i.e., the precision of these estimates cannot be predicted on the basis of the sample design and estimates obtained from pilot data or other sources).

6. The use of imputed data in the estimates will make them less reliable than if they were based on a complete set of real data. The usual estimates of variance are inadequate since they do not include the error due to imputation. Some theoretical work for single variables has been done for particular imputation schemes in restricted circumstances (e.g., Bailar and Bailar 1978, 1979; Ernst 1978, 1980; see Section 5); this may be illuminating, but the imputer is unlikely to find a formula that applies exactly to his or her case.

7. The imputer is faced with ethical problems if the microdata are ever going to be given out. At the very least one must plan to identify the imputed items on all copies of the data and publish the proportions of imputations in each field as part of a discussion of data quality when the primary results are published. Alternatively, one may choose to give out edited, but unimputed, versions of the data set. In this case one must be prepared for the fact that secondary users may do their own imputations and get results that are inconsistent with each other and the original.

Which data set should be analyzed? The question really is "What do you mean by analysis?" If one wants to explore relationships between variables, the use of imputed data could be prejudicial, not to mention misleading. For simple estimation purposes, as we have pointed out, the imputed set reduces the headache.

After considering these problems we may conclude that the imputer needs a procedure that

1. will impute values which are consistent with the edits provided only that the nonmissing data satisfies the edits;
2. will reduce the nonresponse bias and preserve the relationships between items as far as possible;
3. will work for (almost) any pattern of missing items;
4. can be set up ahead of time;
5. can be evaluated in terms of impact on the bias and precision of the estimates.

Particular techniques of imputation vary in their ability to meet these requirements.

4. METHODS OF IMPUTATION

Planning ahead is to be recommended: if one can guess the fields most likely to cause problems, it will pay to pick up (one or more) correlated variables on the questionnaire or from auxiliary sources. For example, it may be hard to get information about household income, but easy to get an estimate of square feet of living space or some other correlate of income. The store manager may not want to disclose gross income; but one can count the number of cash registers. How this information is used depends on the circumstances.

Techniques of imputation vary from naive to sophisticated.

1. Use of ad hoc values. Each case may be treated differently in a manual procedure, or a few rules of thumb may be formulated on the basis of "experience" and hunches. These are used to "fill in the blanks."

For example, in a business survey we may have the following rule for imputing the value (in dollars) of closing inventory (CI) using the values for gross income (GI) and net income (NI), assumed to be present (or already imputed):

$$\text{If GI} \le 25{,}000, \text{ set CI} = 0;$$
$$\text{if GI} > 25{,}000, \text{ set CI} = \max\{0, .05\,(\text{GI} - \text{NI})\}$$

In many ways this rule appears quite reasonable, provided GI and NI are always available, especially if the 5 percent came from last year's survey. If it is dirty, it is at least quick and not too damaging if only a small percentage of the records are affected.

Rules of this type can be formulated to force compliance with the edits. They are also compatible with the simplest of data-processing systems. However, they are subjective and may not reflect reality. The effects on the underlying distributions are often unpredictable and nonresponse bias is not necessarily reduced. Evaluation may be impossible.

2. Poststratify and use the poststratum marginal mean or another typical value (e.g., in the case of a

categorical variable, the mode), making sure that there are sufficient data in each poststratum. For the estimation of means and totals in the numeric case, this is equivalent to item-by-item reweighting. However, in the numeric case, use of the marginal mean may result in impossible values, such as 2.3 children in a household.

In the closing inventory sample of Technique 1, we might poststratify by gross income, net income, industry, region, and so on. If we create too fine a grid or too many data are missing, some collapsing may be necessary to ensure that there are enough good data in each cell.

This technique may run into trouble with the edits. If this seems likely, some modification may be in order (such as letting the edits define the poststrata). Like the method of ad hoc values, it is very simple, if it works; but will create spikes in the marginal distributions and may be biased if the data are not missing at random within each cell.

3. Model the relationships between the variables. In fact, all imputation procedures imply the existence of a model of some kind, either explicit or implicit. For example, in the case of Technique 2, use of the poststratum mean implies a model of data missing at random within the cell.

If it is believed that there is a relationship between the variable Y and the variables $\mathbf{X}$ that are assumed to be present, then a model, such as $E(Y) = f(\mathbf{X})$, can be fitted from those responses where Y is present and used to predict the value of Y in a response where Y is missing. Of course, this assumes there is no nonresponse bias in Y given $\mathbf{X}$.

Similarly, if some of the (numerical) response variables $\mathbf{Y} = (Y_1, Y_2 \ldots Y_k)$ are missing and there are no adequate predictor variables $\mathbf{X}$ that are always present, one can model the joint distribution of $\mathbf{Y}$ using complete (in $\mathbf{Y}$) responses and then, for each response in which some of the $Y_1, Y_2 \ldots$ are missing, predict the missing variables using the marginal expectation of the missing Y's given the Y's that are present; that is, for the ith observation,

$$\hat{\mathbf{Y}}_{i,\ \text{missing}} = E(\mathbf{Y}_{\text{missing}} \mid \mathbf{Y}_{\text{present}} = \mathbf{Y}_{i,\ \text{present}})$$

(where we have partitioned $\mathbf{Y}$ into $\mathbf{Y}_{\text{missing}}$ and $\mathbf{Y}_{\text{present}}$). An extension of this idea modifies the estimate of the distribution of $\mathbf{Y}$ to include the information from the partially incomplete observations (Beale and Little 1975; Dempster, Laird, and Rubin 1977; Hocking and Marx 1979). However, this method, in practice, requires the assumption of normality, which is not usually a plausible assumption, and it does not take the edit structure into account. This author has not seen any theory worked out for nonnormal cases and is not aware of any application to missing survey data, except for test purposes (e.g., Huddleston and Hocking 1978).

However, simple ratio or regression techniques are often used. In one survey at Statistics Canada about 160 items are collected (from administrative documents) for a fairly small sample of businesses and five major items are collected from other sources for the entire population. For various reasons (mainly the ease of arbitrary tabulation of estimates), it is desired to impute the 160 items for the nonsampled businesses. A ratio-type imputation is used, after stratification by size and industry:

$$\hat{x}_i = \left(\sum_P x_j \Big/ \sum_P Y_j\right) Y_i,$$

where x is related to major item Y and the ith record requires imputation. P is the sample of complete records with all 160 items present. Because of the structure of the data, the edits are automatically satisfied; but the imputations do not reflect the real structure of the data that have a lot of zero values. In other words, the imputed records are not realistic and the marginal distributions are distorted. On the other hand, the principal estimates, which are just ratio estimates, are quite acceptable and permit variance estimation. In this case the ratio-type imputation is used because it is easy and convenient, not because it is a good model. The effort that would go into fitting a good model would be prodigious, and one might well never achieve a good fit.

Modeling is a mathematically appealing solution that will probably reduce bias in the estimation of means and totals. On the other hand, achieving a good fit may require a great deal of effort, or one may have to tolerate a bad fit, and there may be problems with edits. Also, the distributional properties of the data may be distorted so that, for example, a correlation coefficient may be poorly estimated. Furthermore, one may find that the assumed model becomes "built into" the data and may be recovered by other researchers later, unless steps are specifically taken to prevent this.

4. Use of historic data, such as last month's or last year's response for the same unit, if available. This technique is in common use in monthly surveys where the same units are surveyed in consecutive months, for variables that are not expected to change often (see, e.g., Ashraf and Macredie 1978). Of course, the assumption is that one did get a response for the particular item at some stage (and when one has carried a value forward for several months in a row, one perhaps ought to do some investigation into what is going on).

5. Use of proxy data from another source. This means that another file, perhaps of administrative data such as medical or tax records, is available with the unique identifiers required for matching to the survey file and that this file includes an equivalent item that can be used as a proxy for the missing survey item (see, e.g. Schieber 1978; Oh, Scheuren, and Nisselson 1980).

If an exact match is not available (possibly because the identifiers have been removed for reasons of confidentiality), one may be content with a statistical match on classification fields such as age, sex, and place of birth. For example, one may use last year's sample survey as a source of data for statistical matching and imputation for this year's survey.

Most statistical matching is used for linking different data files to extend data sets (see, e.g., Radner 1978). The idea of statistical matching is closely related to the hot deck and nearest neighbor techniques, the next two techniques to be discussed.

6. Use of the current survey data as a source of matched individual data records from which one, the donor, is selected at random to supply values for missing items in a particular deficient record. Procedures of this type are often called hot deck procedures; but there is no agreement on the definition of hot deck procedures in the literature. I will take it to mean an imputation procedure that uses records from the current survey to supply missing values and involves a random or pseudorandom choice. There seem to be two main variants currently in use, both directed mainly at categorical data:

a. The sequential hot deck, used in the U.S., for example, in the Census of Population. Here the data are processed one record at a time. To impute a field or group of fields A, a cross-classification (matrix) of several related fields $(B,C,D\ldots)$ is defined. For each cell in this classification, that value of A is retained that occurred in the last record processed with the corresponding values of $B,C,D \ldots$. Thus, as the file is processed, the values in the individual cells of the $B,C,D \ldots$ matrix change. When a record lacking a value for A occurs, it receives the value currently in the cell of the matrix that matches its own values of $B,C,D \ldots$. If two such records (missing A, but with the same values of $B,C,D \ldots$) occur consecutively, the same value of A will be imputed in each case.

The ordering of the file may not be random, so that the record used as a donor is not chosen at random. In fact, it may be advantageous not to randomize the file, thereby exploiting the correlations between nearby records to improve the imputation.

The matching fields (and therefore the imputation matrix) vary with the fields to be imputed, so that many matrices must be maintained. In those cases where imputation of a single field might result in an edit failure after imputation, a set of related fields is deleted and imputed as a group.

Because different fields are imputed from different imputation matrices, several donors may be involved in completing a single deficient record; this may be a source of some concern.

Each imputation matrix must be initialized, using historic data or ad hoc values. On the other hand, the imputation can be done at one pass and is not difficult computationally.

b. The random choice procedure, which is used by the Canadian Census and Labour Force Survey, and by the Current Population Survey in the U.S. for the imputation of items in the March Income Supplement. Here an imputation matrix is not maintained; but the set of records with the required values in the matching fields is identified and the donor is chosen at random from these to supply the missing items to the deficient record.

In the Canadian Census, an attempt is made to impute all missing items on a deficient record using a single donor. If this fails, a field-by-field hot deck is tried in which several donors may be involved (Hill 1978).

In the case of the income items in the CPS, a large number of economic and demographic characteristics are identified as matching fields. Matches at successively lower levels are attempted, and the donors are chosen systematically and (as far as possible) without replacement from within an imputation cell (Oh and Scheuren 1980; Welniak and Coder 1980).

The choice of matching fields in both sequential and random choice procedures must be made considering likely sources of variation, linkage through edits, and the number of complete or eligible records available as potential donors in each cell. If too many fields are used for matching, the number of potential donors may be too small; if too few fields are used for matching, there is a risk of a poor match or edit failure in the imputed record.

With hot deck methods, the variance of the estimates in simple cases is known to be larger than the variance of the usual expansion estimates of means and totals (see, e.g. Ernst 1980). However, there may be a reduction in bias. Compared to some other methods of imputation, such as the use of models, hot deck methods should produce imputed data sets that appear more realistic and do a better job of reflecting distributional properties.

7. Use of the current survey data as a source of individual data records with similar characteristics to supply values for missing items. Unlike the hot deck procedures just discussed, these procedures are appropriate for use when one is matching with numeric data. They will be called nearest neighbor procedures rather than hot deck procedures because the values in the matching fields must be similar (not the same) and the element of randomness in the choice of donor may be absent.

The hot deck procedures discussed in technique 6 run into trouble when numeric fields are linked by edit constraints and matching must be done on them. Occasionally the problem can be dealt with by splitting the range of the variable, for example, age, into intervals and coding the intervals; but this may not always work. For example, consider the problem of imputing the age of a child from the age of its mother. Natural edits might be:

$$\text{(Age of mother)} - \text{(age of child)} \geq 14, \text{ and}$$
$$\text{(Age of mother)} - \text{(Age of child)} \leq 50$$

If the coding is in 10-year intervals (0–9, 10–19, etc.), the problem arises in expressing the edits in terms of the codes. A 29-year-old mother might well have an 11-year-old child; but would one want to use this record for imputing the age of a child when the mother is 23? If one matched on the coded data, this could very well happen.

For purely numeric data with linear edits, a prototype system at Statistics Canada locates the m "nearest" complete records to a particular deficient record. An attempt to complete the deficient record using fields from the nearest of the m neighbors is made. If the tentatively completed recipient record passes the edits, the imputation is complete. Otherwise, the next nearest neighbor is tried, and so on. If none of the m neighbors will do, the imputation fails and further processing is required (Sande 1979).

In this type of imputation, the use of suitable data transformations can make the imputation proceed more smoothly. It also helps to insert additional edits so that extreme observations are not admitted as donors (special arrangements can be made for them). The method requires an efficient search algorithm; but the choice of distance function, given that the data have been suitably transformed, does not appear crucial and one that is simple computationally is advisable.

It is possible that particular records will be used as donors much more often than others. This will increase the variance while possibly reducing the bias of the estimate. Another nearest neighbor type of imputation system, developed at Statistics Canada for the imputation of mixed numeric and categorical data, incorporates the number of times a particular record has been used as a donor into the distance function, so that the distance increases with the number of previous donations (Colledge et al. 1978).

In numeric matching, the match is deterministic given the data. This makes the statistical properties of the procedure difficult to study. In particular, the variance is hard to calculate, even in simple cases. Nearest neighbor procedures can be converted into hot deck procedures by choosing the donor record at random from m nearest neighbors instead of taking the nearest satisfactory record. Both types of procedure can be regarded as forms of nonparametric regression, that is, regression without an explicit model.

8. Use of hybrid methods. In fact, to this author's knowledge, no complex imputation problem is handled by a single imputation procedure. Most imputation procedures start by poststratifying the data into more homogeneous groups. Some ad hoc imputations are usually combined with more sophisticated methods so that the job gets done expediently. Typically, some items are imputed one way and others another way, and then some cleaning up is done.

Sometimes, a two-stage procedure may be employed. For example, this is particularly appropriate when a variable shows a significant occurrence of zeros. The occurrence of zeros is imputed with a model or by hot deck, for example, and then the nonzero values are imputed separately by another method (e.g., Schieber 1978). Another example is the use of a regression model, followed by the addition of a residual value chosen by hot deck or some other random process. In this case, the imputation is

$$Y_{\text{imp}} = \hat{Y} + \hat{\epsilon},$$

where $\hat{Y}$ is the predicted value obtained from the fitted model, which is based on the complete observations; and $\hat{\epsilon}$ is the estimated residual that may be obtained by hot deck from the actual residuals of the fitted values or randomly generated using the estimated distribution of the residuals.

Various devices may be employed to expedite the imputation. Among these are

1. Formulation of the edit procedures to reduce the number of possible missing configurations. More fields than necessary are deleted, but consistent imputation is easier. For example, if the edit is $A + B + C \leq X$, failure of the edit may result in the deletion of all fields A, B, C, and X or just A, B, C rather than only one of these fields. Obviously this is an option to be used with extreme caution, since information is destroyed.

2. Transformation of the data. It is sometimes more natural to impute proportions than absolute numbers, and often the edits transform neatly to permit this. For the purpose of numerical hot decks or nearest neighbor procedures, the distance function is often better formulated in terms of transformed variables than in terms of the originals, which may have very skew distributions. Thus, the data may be much more dense in some parts of the sample space than in others. "Nearness" in a part of the space where the data are sparse may be quite different from "nearness" where the data are dense. For example, economic data are often highly skewed toward the low end; but the difference between $10,000 and $11,000 is $1,000 or 10 percent, while the difference between $100,000 and $105,000 is $5,000 or 5 percent: in some sense, the members of the second pair are more similar than the first.

3. Dividing the record into segments and imputing one segment at a time. Each pass is conditional on the preceding ones being complete. This makes the imputation task less formidable and, in those cases where matching is required, allows different appropriate matching procedures to be used at each stage (Colledge et al. 1978). A related device is to attempt a global imputation first, and, where this fails, then to try a stage by stage imputation (Hill 1978). If all else fails, we can end with an ad hoc procedure to tie up the loose ends.

5. EVALUATION OF IMPUTATION PROCEDURES

Imputation obviously affects the quality of the estimates and the responsible statistician should put some effort into determining, even if only approximately, the nature and magnitude of the "imputation effect." In evaluating an imputation procedure, the most relevant concerns are the bias and variance of the estimates (means, ratios, etc.). The ability of an imputation procedure to guess the missing values of individual items correctly is of lesser importance; although the better the imputation does this, the smaller will be the error due to imputation.

The theoretical treatment of imputation procedures is generally quite difficult. The easiest case is the use of the poststratum mean under simple random sampling, where the variance conditional on the poststratum sizes is easy to find and the unconditional variance is slightly more complicated. For other imputation procedures, the complexity increases rapidly. Nevertheless, a considerable amount of theoretical work has been (and is being) done, although most of it is confined to single variable cases under simple random sampling, where edit constraints are ignored. This reflects the extreme difficulty of the problem, not a deficiency in the research (e.g. Bailar and Bailar 1978,1979; Schaible 1979; Platek and Gray 1978; Ernst 1978,1980; Kalton and Kish 1981; Santos 1981). Imputation under complex sampling procedures requires special thought and care: Cox (1980) has modified the sequential hot deck to accommodate weighted data (see also Cox and Folsom 1981).

There has also been a fair amount of empirical work comparing different imputation procedures (e.g., Cox and Folsom 1978; Ernst 1978; Schieber 1978) or studying the performance of a particular technique under different conditions (e.g., Colledge et al. 1978; Ford, 1976).

Since the scope for theoretical work is limited to fairly simple data and imputation procedures, it seems that, in general, imputation procedures must be evaluated by simulation. This usually means selection or creation of a clean data set (no items missing) to act as a population, the creation of artificial "missings" in biased and unbiased modes and at different rates, and studying the performance of the imputation process over several replicates of each case.

The quality (bias, variance), in relationship to the rate and bias of "missings," of the resulting estimates may then be assessed. Particular imputation procedures will allow variants of this basic recipe: for example, in a sequential hot deck, replicates may be generated by reordering the data set rather than by regenerating a complete set of "missings" as required by nearest neighbor techniques.

From such empirical studies, the imputer can gauge the extent to which imputation affects bias and variance, and he can develop relationships that will allow the estimation of the imputation effects in the actual survey.

Rubin (1978) advocates the routine production of several sets of imputed values under different models or sets of assumptions, as part of the regular data processing. This leads to estimates of the "imputation error," that part of the error due to imputation, in the actual data, so the effects of different models can be studied. The method is applicable only to imputation techniques that employ some random component, such as hot deck or Bayesian procedures. It has been used experimentally and has been found to be useful and informative (e.g. Oh and Scheuren 1980; Cox and Folsom 1981; Herzog and Lancaster 1980; Herzog 1980).

In general, the estimation of the "imputation error" under normal production conditions will be very difficult; but it is better to use approximations obtained from a simulation study than nothing at all.

Whatever the method of imputation, the actual imputation process should be carefully monitored. In the simplest cases this means recording data about the missing items that were subsequently imputed: the number of records in which any imputation is made, the number requiring one (two, three, etc.) item(s) to be imputed, the number of records missing specific variables (or possibly combinations of variables), and statistics breaking down the imputations into those due to item nonresponse and those due to edit failure. For imputations made using a decision tree (the imputation being conditional on other fields and the relationships between them), the number of imputations made in each branch of the tree should be recorded. For a nearest neighbor procedure one also wants to know, for example, how many times each record was used as a donor, which donor was involved in a particular imputation, how many attempts were required to complete a record and what the value of the distance function was. And of course one wants a listing of any records failing to be completed. (It is also equally important to monitor the editing process that precedes the imputation.)

At the very least, monitoring the imputation will record how much imputation was done, and where. It can also give information about the effectiveness of the edits and the imputation procedure, leading to improvements in subsequent versions of the survey or even in other surveys.

6. CONCLUSION

Reality does not consist of the data at the end of the chapter of some textbook (like the iris data) and normal distributions; it consists of 20,000 long forms filled out by 20,000 businessmen with other things on their minds, or several million census returns filled out by individuals who want to get back to the newspaper or the TV. These people want to be cooperative; but if the information requested is not handy or has been forgotten, they pass over the question or make up a response, and they also make mistakes. The survey people have to extract as much sense as possible from the results, and they try to do a respectable and ethical job.

Reality also consists of the almost unlimited and unpredictable demands that are made on some data sets. These should be satisfied in a consistent way. Furthermore, even the simplest survey, properly run, is a complex operation, and one does not want to increase the complexity any more than one has to.

Anyone faced with having to make a decision about imputation procedures will usually have to choose some compromise between what is technically effective and what is operationally expedient. If resources are limited, this is a hard choice. It is to be hoped that this article might be helpful to some in guiding that choice.

This author believes that the real problem of imputation is the interaction with editing. Very little of the literature deals with this problem in general, with the particular exceptions of Fellegi and Holt (1976) and Sande (1979). It is sometimes discussed in particular situations (Ashraf and Macredie 1978; Colledge et al. 1978; Hill 1978) but there has been little empirical work done. By and large, writers prefer to simplify the problem so that it is amenable to mathematical analysis or empirical study. This is not to suggest that the effort is wasted, but that the problem of studying the properties of imputation procedures under realistic conditions is a very difficult one. One must admit that there *are* some one-question surveys to which the available results might be applicable.

This article has not attempted to review the literature, which is extensive and growing. Aside from general papers such as this (e.g., Chapman 1976; Platek 1980), papers may view imputation on its own, in connection with editing, or against the background of the "incomplete data problem." This last area was thought to be sufficiently important to warrant a symposium sponsored by the Panel on Incomplete Data of the Committee on National Statistics (1979).

[*Received June 1981. Revised February 1982.*]

REFERENCES

ASHRAF, A., and MACREDIE, I. (1978), "Edit and Imputation in the Labour Force Survey," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 425–430.

BAILAR, J.C. III, and BAILAR, B.A. (1978), "Comparison of Two procedures for Imputing Missing Survey Values," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462–467.

——— (1979), "Comparison of the Biases of the 'Hot Deck' Imputation Procedure with an 'Equal Weights' Imputation Procedure," *Preliminary Proceedings of the Symposium on Incomplete Data, National Academy of Sciences, Panel on Incomplete Data*, 422–447.

BEALE, L.M.L., and LITTLE, R.J.A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*, Ser. B, 37, 129–145.

CHAPMAN, C.W. (1976), "A Survey of Nonresponse Imputation Procedures," *Proceedings of the Social Statistics Section, American Statistical Association*, 1976, 245–329.

COLLEDGE, M.L.; JOHNSON, J.H.; PARE, R.; and SANDE, I.G. (1978), "Large Scale Imputation of Survey Data," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 431–436.

COX, B.G., (1980), "The Weighted Sequential Hot Deck Imputation Procedure," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 721–726.

COX, B.G. and FOLSOM, R.E. (1978), "An Empirical Investigation of Alternate Item Nonresponse Adjustments," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 219–223.

——— (1981), "An Evaluation of Weighted Hot Deck Imputation for Unreported Health Care Visits," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 412–417.

DEMPSTER, A.P.; LAIRD, N.M.; and RUBIN, D.B. (1977), "Maximum Likelihood From Incomplete Data via the E M Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–11.

ERNST, L.F. (1978), "Weighting to Adjust for Partial Nonresponse," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 468–472.

——— (1980), "Variance of the Estimated Mean for Several Imputation Procedures," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716–720.

FELLEGI, I.P., and HOLT, D.A. (1976), "Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17–35.

FORD, B.L., (1976), "Missing Data Procedures: A Comparative Study," *Proceedings of the Social Statistics Section, American Statistical Association*, 324–329.

HERZOG, T.N. (1980), "Multiple Imputation of Individual Social Security Benefit Amounts—Part II," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 404–407.

HERZOG, T.N., and LANCASTER, C. (1980), "Multiple Imputation of Individual Social Security Benefit Amounts—Part I," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 398–403.

HILL, L.J. (1978), "A Report on the Application of a Systematic Method of Automatic Edit and Imputation to the 1976 Canadian Census," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 474–479.

HOCKING, R.R., and MARX, O.L. (1979), "Estimation with Incomplete Data: An Improved Computational Method and the Analysis of Mixed Data," *Communications in Statistics*, 1155–1182.

HUDDLESTON, H.F., and HOCKING, R.R. (1978), "Imputation in Agricultural Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 480–485.

KALTON, G., and KISH, L. (1981), "Two Efficient Random Imputation Procedures," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 146–151.

OH, H.L., and SCHEUREN, F.J. (1980), "Estimating the Variance Impact of the Missing CPS Income Data," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 408–415.

OH, H.L.; SCHEUREN, F.J.; and NISSELSON, H. (1980), "Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 416–420.

PANEL ON INCOMPLETE DATA of the Committee on National Statistics (1979), *Symposium on Incomplete Data: Preliminary Proceedings*, Washington, D.C.: U.S. Department of Health, Education, and Welfare.

PLATEK, R. (1980), "Causes of Incomplete Data, Adjustments and Effects," *Survey Methodology, Statistics Canada*, 6, 93–132.

PLATEK, R., and GRAY, G.B. (1978), "Non Response and Imputation," *Survey Methodology, Statistics Canada*, 4, 144–177.

RADNER, D.B. (1978), "The Development of Statistical Matchings in Economics," *Proceedings of the Social Statistics Section*, American Statistical Association, 503–508.

RUBIN, D.B. (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20–28.

SANDE, G. (1979), "Numerical Edit and Imputation," International Association for Statistical Computing, 42nd Session of the International Statistical Institute.

SANTOS, R. (1981), "Effects of Imputation on Regression Coefficients," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 140–145.

SCHAIBLE, W.L. (1979), "Estimation of Finite Population Totals from Incomplete Sample Data: Prediction Approach," *Preliminary Proceedings of the Symposium on Incomplete Data*, Washington D.C.: National Academy of Sciences, Panel on Incomplete Data, 170–187.

SCHIEBER, S.J. (1978), "A Comparison of Three Alternative Techniques for Allocating Unreported Social Security Income on the Survey of the Low-income Aged and Disabled," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 212–218.

WELNIAK, E.J., and CODER, J.F. (1980), "A Measure of the Bias in the March APS Earnings Imputation System," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 421–425.