SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE 1981, VOL. 7 NO. 1

IMPUTATION IN SURVEYS : COPING WITH REALITY[1]

I.G. Sande[2]

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. While imputation is not a particularly good solution to any specific estimation problem, it does permit the production of arbitrary estimates in a consistent way.

The survey statistician may have to cope with a mixture of numerical and categorical items, subject to a variety of constraints. He should evaluate his technique, especially with respect to bias. He should make sure that imputed items are clearly identified and summary reports produced.

A variety of imputation techniques in current use is described and discussed, with particular reference to the practical problems involved.

1. INTRODUCTION

Everyone who has been involved in surveys knows that life would be very easy if only the respondent had read the textbook. If he had, he would know that he is allowed to respond correctly and completely, or not to respond at all. He is not allowed to respond incorrectly or incompletely. Unfortunately, the respondent has not read the textbook. Furthermore, if you call him back to correct the data or fill in missing information, he may not be very co-operative. More often than not, the cost of calling back is simply too high to be carried out generally.

---

[1] Presented at the annual meeting of the Statistical Society of Canada, Halifax, 23-26 May, 1981.

[2] I.G. Sande, Business Survey Methods Division, Statistics Canada.

So reality might look like this:

TABLE 1

IMPORTANT CANADIAN SURVEY

Record No	Identification Classification	Weight	Variables 1	2	3	4	5
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	-	0.0	2.4
10	X	W_{10}	-	b	y	-	1.4

Edits: $A \wedge a \Rightarrow$ Not x.

$B \wedge b \Rightarrow$ Not y.

Var 4 + Var 5 $\leq$ 10.

Var 4 $\geq$ 0, Var 5 $\geq$ 0.

This survey has both categorical and numeric items, and there are three constraints (edits) on the items which must be satisfied. 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: a low value of Variable 5 is associated with a missing Variable 4.

Our primary problem is that we have to produce tabulations of population estimates, e.g. Variable 1 x Variable 2 x Classification Variables, or Variable 4 x 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 which we have not anticipated.

How, then, are we to deal with the partial non-response? The possibilities are:

(i) 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.

(ii) Publish "unknowns" as a category. This is a little better than (i); but still ignores the partial information about the missing value which 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.

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

(iv) Fill in the blanks in each record with plausible and consistent values. This is called imputation.

To sum up, partial non-response arises in two ways:

(i) 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.

(ii) 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").

To cope with the "missing value" problem in an expeditious manner, values are frequently imputed for the missing items so that the data set is "completed".

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, though imputation is not an optimal solution to the "missing value" problem in surveys when any particular estimates are considered, it may just be the least bad of the feasible solutions for general purposes.

2. THE GENERAL IMPUTATION PROBLEM

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

(i) 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 which do not fail. Does one change the least number of items, as recommended by Fellegi and Holt [9], or adopt a policy of "least change", whatever that means? Or does one adopt the "principle of expedience" : deleting that configuration which makes imputation easy?

These are non-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 [11]) in the Census of Population. The parallel work for numerical data with linear edits has been carried out by Gordon Sande at Statistics Canada using optimization techniques [20] 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, i.e. the completed record must be consistent. This requirement often eliminates the mathematically elegant imputation schemes and reduces the mathematical tractability of the problem to zero. Since complex edits make the imputation procedure hard, the theoretical analysis of such procedures is virtually impossible. Therefore edits are usually ignored in theoretical work on the properties of imputation techniques.

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

(iii) 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 (I do 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:

(iv) 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 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 respects. For example we may believe that X is proportional to Y on the basis of historic data; but the proportion $\frac{X}{Y}$ may change from year to year. On the other hand, the circumstances governing the joint occurrence or non-occurrence of X and Y may be similar over time, a fact which can be exploited in testing imputation procedures.

(v) 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 if one can optimally estimate a particular $\underset{\sim}{\theta}$ using some (correct) distributional assumptions and a (correct) model, one hasn't solved the problem for $\underset{\sim}{\phi}$. One has to start again. If one combines $\underset{\sim}{\theta}$ and $\underset{\sim}{\phi}$,

one may have an unwieldy problem. By the time one has optimally estimated all the parameters one can think of, one may have a set of estimates which is not consistent with any possible data set. And then someone may find a $\underset{\sim}{\psi}$ to be estimated. By imputing a consistent value for each missing item one can estimate any of the usual population parameters (means, totals, ratios, differences, proportions, correlations) very easily, although possibly with no guaranteed precision.

(vi) It is generally hard to know how to estimate the variance of estimates when some data is imputed. If the amount of imputed data is very small, the usual estimates will do. In some circumstances, mathematical or empirical studies in a vaguely related situation may be available.

(vii) The imputer is faced with ethical problems if the microdata are ever going to be given out. At the very least, he 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, he may choose to give out edited, but unimputed, versions of the data set. In this case, the secondary users may do their own imputations and get results which 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. And we could argue that if the data are so bad that the presence of imputed data could influence the analysis significantly, then the data are not worth analyzing.

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

(i) will impute plausibly and consistently provided only that the non-missing data satisfy the edits;

(ii) will preserve the underlying distributions in the data or, at least, reduce the response bias and preserve the relationships between items as far as possible;

(iii) will work for (almost) any pattern of missing items;

(iv) can be set up and tested ahead of time;

(v) can be evaluated in terms of data quality and impact on precision of the estimates.

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

3. 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 a correlated variable 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 his 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.

(i) Use of ad-hoc values. Each case may be treated differently in a manual procedure, or a few rules of thumb are formulated on the basis of "experience" and hunches, and often without the encumbrance of real facts. These are used to fill in the blanks.

For example, in a business survey we may have the rule for imputing the value of closing inventory: if gross income (GI) is less than or equal to $25,000, set closing inventory (CI) to 0; if GI is greater than $25,000, set CI equal to 5% of GI minus net income or 0, whichever is larger. In many ways this rule appears quite reasonable, provided GI and net income are always available, especially if the 5% 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 non-response bias is not necessarily reduced. Evaluation may be impossible.

(ii) Post-stratify and use the post-stratum marginal mean or another typical value (e.g. the mode in the case of a categorical variable), making sure that there are sufficient data in each post-stratum. In the numeric case, this is equivalent to item by item reweighting.

In the closing inventory example of (i), we might post-stratify by gross income, net income, industry, region, etc. 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 (see [8]).

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 post-strata). 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. However, in the numeric case variance estimates are generally available.

(iii) Model the relationships between the variables. A popular idea has been to use the conditional mean given the items present, modified to account for the information in the incomplete records assuming normality, or some generalization of this idea (e.g. [3], [7], [12]). However, normality is not usually a plausible assumption and it does not take the edit structure into account. I have not seen any theory worked out for non-normal cases and I am not aware of any application to missing survey data except for test purposes (e.g. Huddleston and Hocking in [1], pp. 88-93).

In one survey at Statistics Canada, about 160 items are collected (from administrative documents) for a fairly small sample of businesses and 5 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 non-sampled businesses. A ratio-type imputation is used, after stratification by size and industry:

$$\hat{x}_i = \frac{\sum_P x_j}{\sum_P Y_j} Y_i$$

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 which 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 model would be prodigious and one may well never achieve a good fit.

Thus modelling is an elegant solution which will probably reduce bias. 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. 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.

(iv) 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 which are not expected to change often. 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.

(v) Use a 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 which can be used as a proxy for the missing survey item (e.g. [10]).

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 in [1], pp. 108-113). The idea of statistical matching is closely related to the hot deck and nearest neighbour techniques discussed in (vi) and (vii) below.

(vi) 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 which uses records from the current survey to supply missing values and involves a random or pseudo-random 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 Current Population Survey and 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 other related fields (B,C,D...) is defined. For each cell in this classification, that value of A is retained which occurred in the last record processed with the corresponding values of B,C,D.... . Thus, as the file is processed, the values in the individual cells of the B,C,D... matrix change. When a record lacking a value for A occurs, it receives the value currently in the cell of the matrix which matches its own values of B,C,D... If two such records (missing A, but with the same values of B,C,D...) 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 not be advantageous 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 together.

Because different fields are imputed from different imputation matrices, several donors may be involved in completing a single deficient record and 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 used by the Canadian Census and Labour Force Survey. 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 [11].

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 (e.g. [2]). However, there may be a reduction in bias.

(vii) 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 in (vi), these procedures are appropriate for use with numeric data. I shall call them nearest neighbour procedures rather than hot deck procedures because the value 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 (vi) 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, e.g. age, into intervals and coding the intervals; but consider the problem of imputing the age of a child from the age of a parent.

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 neighbours is made. If the tentatively completed recipient record passes the edits, the imputation is complete. Otherwise, the next nearest neighbour is tried, and so on. If none or the m neighbours will do, the imputation fails and further processing is required [20].

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 is not crucial and one which is simple computationally is advisable.

It is possible that particular records will be used as donors much more often than others. Another nearest neighbour 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 [5].

Nearest neighbour procedures can be converted into hot deck procedures by choosing the donor record at random from m nearest neighbours instead of taking the nearest satisfactory record. Both types of procedure can be regarded as a form of non-parametric regression.

With numeric matching, the variance would be hard to calculate since the match is deterministic given the data.

(viii) Use of hybrid methods. In fact, to my knowledge, no complex imputation problem is handled by a single imputation procedure. 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. In one case [22], the occurrence of zeros in a particular variable was modelled. Those missing cases not imputed as zero through the model were imputed by hot deck.

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

(i) 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.

(ii) 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 neighbour procedures, the distannce function function is often better formulated in terms of transformed variables than the originals which may be very skew. In terms of the original variables, "nearness" in one part of the space may be quite different from "nearness" in another.

(iii) 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 [5]. A related device is to attempt a global imputation first and, where this fails, to try a stage by stage imputation [11]. If all else fails, we can end with an ad-hoc procedure to tie up the loose ends.

IV EVALUATION OF IMPUTATION PROCEDURES

In evaluating an imputation procedure, the relevant concerns are bias and variance of the estimates (means, ratios, etc.) not the ability of an imputation procedure to guess missing values of individual items correctly.

The theoretical treatment of imputation procedures is generally confined to fairly simple cases, ignoring edit constraints (e.g. Bailar and Bailar in [2] and [15], pp. 422-447; Schaible in [15], pp. 170-187; Platek and Gray, [17]; Szameitat and Zindler, [23]). Empirical work deals either with the comparison of different imputation methods (e.g. [6], [8], [22]); or with the performance of a particular technique under different conditions ([5], [10]). Various edit and imputation strategies are compared by Nordbotten in [13]. Other studies simply attempt to examine the impact of imputation [14], or summarise current practice [18].

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 re-ordering the data set rather than by regenerating a complete set of "missings" as required by nearest neighbour techniques.

Rubin [19] 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 and the effects of different models can be studied. The method which is applicable to only a limited variety of imputation techniques, including hot deck, has been used experimentally.

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 which 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), statistics breaking down the imputations into those due to item non-response 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 neighbour 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 which precedes the imputation).

V. CONCLUSION

This is not the first paper on imputation in surveys (e.g. [4], [16], [18], [23]) nor will it be the last. The activity has been going on for a long time under such disguises as "automatic error correction" and used to be considered as part of data processing rather than statistical methodology. Now the survey statisticians are getting involved and the subject is being discussed in the literature and at meetings. Predictably, the open discussion of imputation has dismayed some of the more classical statisticans.

Reality does not consist of the data at the end of the chapter (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 co-operative; but if the information requested isn't handy or has been forgotten, they omit 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 which are made on some data sets. These should be satisfied in a consistent way. And reality is the fact that 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.

I believe that the real problem of imputation is the interaction with editing. Very little of the literature deals with this problem. Szameitat and Zindler [23] and Nordbotten []3] touch on the subject. The "Canadian School" led by Fellegi and Holt ([9]; [5], [11], [20] and even [21]) discuss it (with little empirical work), while, by and large other writers do not, preferring to simplify the problem so that it is amenable to mathematical analysis or empirical study. This does not suggest to me that the effort is wasted, but that the problem of studying the properties of imputation procedures under realistic conditions is a very difficult one. And one must admit that there <u>are</u> some one-question surveys to which the available results might be applicable.

I hope that we will see more empirical work on data sets with complex edit constraints. We need to know much more about how imputation procedures compare with each other and we need guidance about how to optimize the performance of a specific type of procedures. So far, we have only scratched the surface.

RESUME

Dans les enquêtes, il arrive qu'une réponse soit incomplète ou que certains éléments soient incompatibles ou encore, que des éléments puissent manquer, comme dans le cas de l'echantillonnage à deux phases. Il peut alors être utile d'imputer des valeurs aux éléments manquants. Même si cette méthode n'offre pas une solution particulièrement bonne à un problème d'estimation donné, elle permet cependant la production d'estimations arbitraires d'une façon cohérente.

Le statisticien enquêteur sera peut-être aux prises avec un mélange d'éléments numériques et qualitatifes qui seront assujettis à une variété de contraintes. Il doit évaluer sa technique, en particulier en ce qui concerne le biais, et veiller à ce que les éléments imputés soient nettement identifiés et que des rapports sommaires soient produits.

L'auteur décrit diverses techniques d'imputation utilisées à l'heure actuelle et elle accorde une attention particulière aux problèmes pratiques en cause.

REFERENCES

[1] Aziz, F. and Scheuren, F., Imputation and Editing of Faulty or Missing Data, 1978, U.S. Department of Commerce, Bureau of the Census (papers presented at the 1978 meetings of the American Statistical Association, almost all appearing in the Proceedings of the Section of Survey Research Methods).

[2] Bailar, J.C. III and Bailar, B.A., "Comparison of Two Procedures for Imputing Missing Survey Values". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, pp. 462-467. Also [1], pp. 67-75.

[3] Beale, L.M.L., and Little, R.J.A., "Missing Values in Multivariate Analysis", Journal of the Royal Statistical Society, Series B, Vol. 37, 1975, pp. 129-145.

[4] Chapman, D.W., "A Survey of Nonresponse Imputation Procedures". Proceedings of the Social Statistics Section, American Statistical Association, 1976, pp. 245-329.

[5] Colledge, M.L., Johnson, J.H., Pare, R., and Sande, I.G., "Large Scale Imputation of Survey Data". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, pp. 431-436. Also, Survey Methodology, Statistics Canada, 1978, Vol. 4, No. 2, pp. 203-224; and [1], pp. 102-107.

[6] Cox, B.C. and Folsom, "An Empirical Investigation of Alternate Item Nonresponse Adjustments". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, pp. 219-223; also [1], pp. 51-55.

[7] Dempster, A.P., Laird, N.M. and Rubin, D.B., "Maximum Likelihood from Incomplete Data via the E M Algorithm". Journal of the Royal Statistical Society, Series B, Vol. 39, 1977, pp. 1-11.

[8] Ernst, L.F., "Weighting to Adjust for Partial Nonresponse", Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, 1978, pp. 468-472. Also [1], pp. 87-91.

[9] Fellegi, I.P., and Holt, D.A., "Systematic Approach to Automatic Edit and Imputation". Journal of the American Statistical Association, Vol. 71, 1976, pp. 17-35.

[10] Ford, B.L., "Missing Data Procedures: A Comparative Study". Proceedings of the Social Statistics Section, American Statistical Association, 1976, pp. 324-329.

[11] Hill, C.J., "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, 1978, pp. 474-479. Also Survey Methodology, Statistics Canada, Vol. 4, 1978, pp. 178-202; and [1], pp. 82-87.

[12] Hocking, R.R., and Marx, D.L., "Estimation with Incomplete Data: an Improved Computational Method and the Analysis of Mixed Data". Communications in Statistics - Theory and Methods, Vol. A8, 1979, pp. 1155-1182.

[13] Nordbotten, S., "The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means of Improving the Quality of Statistics". Bulletin of the International Statistical Institute. Proceedings of the 35th Session, Vol. 16, 1965, pp. 417-441.

[14] Ono, M., and Miller, H.P., "Income Nonresponse in the Current Population Survey". Proceedings of the Social Statistics Section, American Statistical Association, 1969, pp. 277-288.

[15] Panel on Incomplete Data of the Committee on National Statistics/ National Research Council, Symposium on Incomplete Data: Preliminary Proceedings, 1979. U.S. Department of Health, Education and Welfare, Social Security Administration Office of Policy, Office of Research and Statistics.

[16] Platek, R., "Causes of Incomplete Data, Adjustments and Effects". Survey Methodology, Statistics Canada, Vol. 6, 1980, pp. 93-132.

[17] Platek, R., and Gray, G.B., "Nonresponse and Imputation", Survey Methodology, Statistics Canada, Vol. 4, 1978, pp. 144-177.

[18] Pritzker, L., Ogus, J., and Hansen, M.F., "Computer Editing Methods - Some Applications and Results". Bulletin of the International Statistical Institute. Proceedings of the 35th Session, Vol. 16, 1965, pp. 442-465.

[19] Rubin, D.B., Multiple Imputations in Sample Surveys - A Phenomological Bayesian Approach to Nonresponse". Proceedings of the Section on Survey Research Methods, American Statistical Association, 1978, pp. 20-28.

[20] Sande, G., "Numerical Edit and Imputation". International Association for Statistical Computing, 42nd Session of the International Statistical Institute, December 1979.

[21] Sande, I.G., "A Personal View of Hote Deck Imputation Procedures". Survey Methodology, 1979, Statistics Canada, Vol. 5, pp. 238-258.

[22] Schieber, S.J., "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, 1978, pp. 212-218. Also, [1], pp. 44-50.

[23] Szameitat, K., and Zindler, H.J., "The Reduction of Errors in Statistics by Automatic Corrections". Bulletin of the International Statistical Institute. Proceedings of the 35th Session, Vol. 16, pp. 395-417.