Survey Methodology, December 1986
Vol. 12, No. 2, pp. 181–196
Statistics Canada

Hot Deck Imputation Procedure Applied to a Double Sampling Design

SUSAN HINKINS and FRITZ SCHEUREN[1]

ABSTRACT

From an annual sample of U.S. corporate tax returns, the U.S. Internal Revenue Service provides estimates of population and subpopulation totals for several hundred financial items. The basic sample design is highly stratified and fairly complex. Starting with the 1981 and 1982 samples, the design was altered to include a double sampling procedure. This was motivated by the need for better allocation of resources, in an environment of shrinking budgets. Items not observed in the subsample are predicted, using a modified hot deck imputation procedure. The present paper describes the design, estimation, and evaluation of the effects of the new procedure.

KEY WORDS: Double sampling; Hot deck; Imputation.

1. INTRODUCTION

When the U.S. Internal Revenue Service (IRS) is mentioned, the first words to cross one's mind may not be "sample surveys." But every April, those of you from the U.S. take part in at least one of our administrative "surveys" and file an individual income tax return. We sample this administrative data annually for statistical purposes. Another of our major programs is an annual sample of U.S. corporate tax returns; that is the sample survey discussed here.

The primary interest at a Symposium like this is in non-response or other undesirable missing data. Despite our extensive enforcement efforts, we at IRS also have such non-response problems. However, the present paper is concerned with a different type of missing data problem: missingness that is not unexpected, but is designed (see also, Strudler, Oh, and Scheuren 1986, for another example). We take the liberty of discussing these problems because we use techniques usually associated with non-response, e.g., hot deck imputation (Ford 1983). Our case allows an evaluation of the imputation procedure, since the underlying non-response mechanism is known.

Double sampling has been introduced in our corporate tax return sample in an effort to reduce costs with only a "tolerable" loss of information. Reweighting to account for the subsampling stage is a standard estimation approach in double sampling (e.g., Cochran 1977); however, in our application, we would have had to reweight almost on an item-by-item basis. This was judged unacceptable by our users, who require rectangular data sets. (For an analogous approach in a Canadian context, see Colledge *et al.* 1978.)

The imputation technique used – hot deck imputation – is procedurally simple. The need to discuss the application of such a relatively simple procedure may surprise theoreticians; but, as we will show, the problems of implementation within the setting of a large statistical operation are many.

[1] Susan Hinkins, Statistics of Income Division, Internal Revenue Service, P.O. Box 369, Bozeman, Montana 59771. Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, DC 20224.

In the remainder of the present paper, we describe in some detail the double sampling procedure and the imputation technique employed. Preliminary results on the impact of these procedures are also presented and the last section contains our conclusions and future plans. A brief theoretical discussion of the estimators we are using and their properties is given in an Appendix.

2. DESCRIPTION OF THE SAMPLING PROCEDURES

An annual sample of U.S. corporate tax returns is used by IRS to estimate National totals of both tax and economic variables. For example, approximately three million corporate tax returns will be filed for 1985, and the IRS sample will contain over 90,000 of these returns. (In Canada, there are two separate corporate tax return samples, each designed to meet narrower purposes. The Revenue Canada Taxation sample (e.g., Burpee and McGrath 1982) was developed for tax policy simulation purposes. The Statistics Canada sample (e.g., Ambrose 1985) is intended primarily to estimate economic aggregates. It is our belief that separate designs in the U.S., but not entirely separate processing systems, could lead to improvements in efficiency over the current procedures; however, the work done (Clickner *et al.* 1984) indicates that the problem is quite difficult and progress has been slow.)

The annual estimates obtained are for the entire corporate population and for subpopulations, usually defined by industrial activity and size. The underlying population is highly skewed. For most variables, a small proportion of the population accounts for a substantial fraction of the total dollar amount. Examples for 1982 corporations are given in Exhibit 1.

A highly stratified sample design is used; small corporations are selected with small probability and large corporations are selected with certainty (Jones and McMahon 1984). The strata are defined by industrial classification and the size of the corporation (i.e., in terms of assets and net income). Selection probabilities for each stratum are determined by employing a modified form of Neyman allocation. Almost all of the returns in the 100% strata (returns selected with certainty) have total assets of $50 million or more. A form of post-stratified raking ratio estimation is used to weight the sample results (Leszcz, Oh, and Scheuren 1983).

Retrieving the information from each sampled return is a time-consuming and expensive process. Over 600 items may be retrieved from a return, and these items are not simply extracted; they are also carefully checked and redistributed to compensate for taxpayer reporting variations. The complete process is referred to as "editing the return". The cost of "editing" varies by degree of complexity. It may take only twenty-five minutes to edit a fairly simple return but as long as a week to edit a really complicated one. The quality of the editing is vital to our estimates, as these checks reduce, but do not eliminate reporting inconsistencies.

Exhibit 1

Degree of Concentration of Selected Corporate Variables

Selected Items	Assets Under $50 Million	Assets $50 Million or more
Number of Returns	99.6%	0.4%
Total Assets	16.3	83.7
Total Receipts	39.3	60.7
Total Income Tax	25.9	74.1

Source: Internal Revenue Service, 1985.

Indeed, nonsampling error is a serious concern in the data "editing" process, particularly for the largest corporations. In order to spend proportionately more resources on reducing the nonsampling error for the large returns, we introduced stratified double sampling for the smaller returns; specifically, certain data items were retrieved on only a subsample of the returns (i.e., a subset of returns with assets under $50 million). Although this change would increase the error for some variables on the small returns, we expected that the procedure would have little adverse effect on the estimates of national totals, or on the subdomain estimates of primary interest to our major users. There were two main reasons for this conjecture:

- As already noted, corporate *returns* with total assets of $50 million or more were not subject to the extra sampling step.
- The information loss due to the subsampling was reduced by the choice of the *items* or variables to be subject to subsampling.

By and large, as will be shown, the results obtained so far confirm our expectations.

Items Selected for Subsampling

When certain miscellaneous items on a return are nonzero, the taxpayer must attach a schedule providing additional information. For example, if the item "Other Income" is nonzero, the corporation must describe what was included under this category. The schedules are attached on separate sheets of paper and have no standard form or length. The process of editing a schedule has several parts: finding the schedule, deciding whether the taxpayer included appropriate amounts in "Other Income", and making changes if there are errors.

Beginning with the tax year 1981 corporate program, the statistical editing of data from the tax return was done in stages, and certain items were initially transcribed for statistical use directly from the return. Employing automatic tests, items or schedules could then be "flagged" for abstraction or further scrutiny in later stages (Cys *et al.* 1982). This new strategy allowed us to:

- Retain original taxpayer information as reported so that the amount of editing change could be evaluated. Prior to the 1981 sample, we had no information regarding the extent of the adjustments being made by editing. The editors only recorded the final result. (See Powell and Stubbs 1981.)
- Decide whether or not to review a particular schedule based on the initial information transcribed. (Again, prior to the 1981 program, editors were, of course, required to completely edit all schedules.)

For the 1981 and 1982 corporate programs, seven items and their associated schedules were picked for subsampling: schedules for Other Income, Other Deductions, Other Costs of Goods Sold, Other Current Assets, Other (Noncurrent) Assets, Other Current Liabilities and Other (Noncurrent) Liabilities.

The reported amounts on a corporate return may be modified substantially as a result of the editing. For example, consider the "Other Income" schedule shown in Exhibit 2. The original amounts (in column 1) are observed initially for every return. The variables being subsampled are changes that would be made if the Other Income schedule were edited (column 2). In this hypothetical case, we have an original Other Income amount of $1,600, which, when examined by the editor, could be reclassified as including $900 from Business Receipts, $300 in Rents and $400 that really belongs in Other Income. The variables of interest are, of course, the final ("corrected") amounts for each item.

Before implementing the new processing system, an experiment was run comparing the amount of time it took to do the reduced, initial transcription and the amount of time it took to do the complete editing (reading all schedules). As expected, the reduced edit was

Exhibit 2

Illustration of Editing Other Income

Income Type	Original Amounts($)	Change Amount($)	Final Amounts($)
Other Income	1,600	−1,200	400
Receipts	500	+900	1,400
Rents	0	+300	300
Interest	700	0	700

significantly faster (and therefore, cheaper). Considerable resources could be saved by subsampling. (Conservatively, we extrapolated 1981 cost savings of at least $300,000, assuming only limited use of the subsampling technique.)

Double Sampling

We are now ready to describe the basic two-dimensional stratification chosen for our double sampling. The returns are stratified into "crucial" returns (Group A) versus the remaining returns (Group B). "Crucial" returns include all returns with total assets of $50 million or more, thereby including the important "large" returns and most returns selected into the sample with certainty. In addition, crucial returns should include corporations of any size for which the likelihood of an editing change was high. What we want, obviously, is a subsampling plan that has us edit all schedules that have a high probability of a change (especially a large change) and lets us subsample the rest.

In an attempt to predict which schedules are likely to change, a record is included in Group A if the original amount in Other Income, to continue our illustration, is unusually large compared to the amount in Total Income.

Also, since we do not want to impute large amounts, cases where Other Income is above a certain dollar value should be included in Group A, as well. (Unfortunately, this was done only indirectly.) By inference, Group B is supposed to include only small returns which we believe are likely to have little or no change made as a result of editing. (See Barker *et al.* 1982, for details.)

For the crucial returns in Group A, all variables (items) are always completely observed. Only returns in Group B are subject to the subsampling of the seven schedules mentioned earlier. Even for Group B returns, the original amounts for all items are always recorded; therefore, some information is obtained for every item. The information not obtained for some records in Group B is the change due to editing a schedule. It is these changes that are being imputed using the procedure described in the next section. Not all variables are affected by the subsampling. For example, of the 600 items picked up for the 1981 corporation program, only 56 were in any way affected by the double sampling; however, of the approximately 100 major income and balance sheet items, nearly one half could be affected.

3. THE IMPUTATION PROCEDURE

The missing information (i.e., changes from editing) in Group B was imputed using a hot deck procedure within adjustment cells. A record with schedules to be imputed was matched to a donor record, in the same adjustment cell, with these same schedules edited. (The formation of adjustment cells is described later in this section.)

In 1981, the subsampling rate was 10% for the returns subjected to subsampling: one out of ten was selected systematically for editing (these were the hot deck "donors") and the other nine were left to be imputed. In 1982, the subsampling rate was kept at 10% for non-financial returns (trade, manufacturing, etc.) but was raised to 20% for financial returns (banks, insurance companies, etc.)

Within an adjustment cell, the number of returns, n', can be divided into the number of donors, n'', and the number of imputes, $n' - n''$. Because of the small subsampling rate, the number of donors is almost always smaller than the number of imputes. In particular, let $n' - n'' = rn'' + t$ where r and t are nonnegative integers and $0 \leq t < n''$. Then the hot deck procedure selects all n'' donors r times, and selects the remaining t units by simple random sampling without replacement.

To continue our illustration, recall that the item of interest is Z, the final "corrected" amount for Other Income; Z can be written as $Z = X - Y$, where X is the original taxpayer amount in Other Income and Y is the change made due to editing the Other Income schedule. It is only the change, Y, that is unobserved and must be estimated for a subset of the returns in Group B.

If we simply employ a conventional hot deck procedure and estimate the unobserved y_i value, on record i, with the observed value y_j from donor record j, then the resulting estimate of the final value z_i may not satisfy the edit checks. For example, assume the donor record had \$30,000 originally as Other Income, and \$15,000 was removed when the schedule was edited. Suppose that on the record to be imputed, the original amount in Other Income is \$10,000, then the imputed change of \$15,000 would result in a negative estimate for other income:

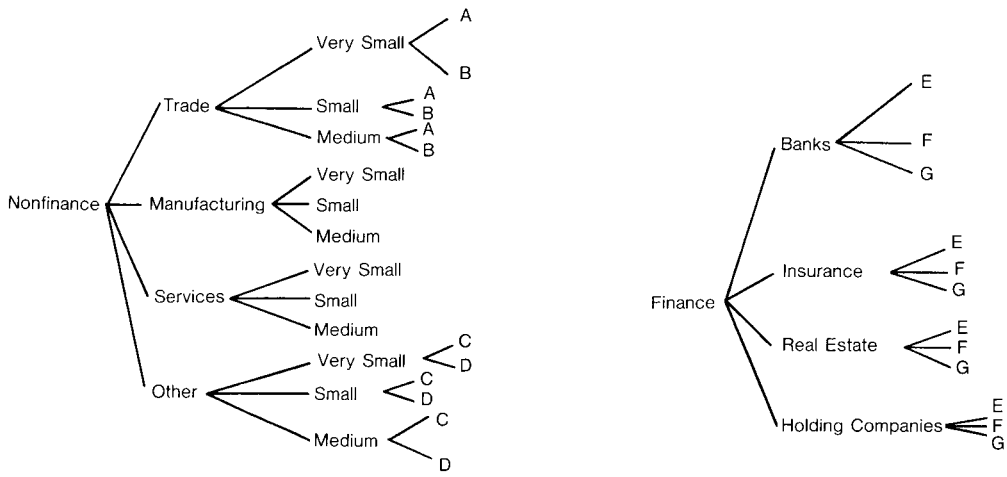
$$\hat{z}_i = x_i - \hat{y}_i = 10{,}000 - 15{,}000 = -5{,}000.$$

Since the amount for Other Income must be nonnegative, edit checks would fail and additional adjustments would have to be made to the record. (See Sande 1982, for a general discussion of this problem.) Since the original amount is always observed, it seemed more reasonable to "hot deck" the relative change $R = Y/X$ rather than the actual change Y. In this example, since the donor record had one half of the amount in Other Income removed after reading the schedule, then 1/2 should be removed on the imputed record. The estimated final amount in Other Income is then

$$\hat{z}_i = x_i - \hat{y}_i = 10{,}000 - (1/2)10{,}000 = +5{,}000.$$

In addition to satisfying the edit checks, we expected the ratio procedure to reduce the variance of our estimates relative to the basic hot deck approach; however, the variance of the estimator is not analytically tractable and must be measured empirically. We have not yet verified in our corporation application the smaller variance that we conjecture; but simulation results do support the approach we have taken. However, by introducing the ratio, our estimators are now biased. We conjectured that the biases would be small and in fact they were, for the most part, as we shall show.

The model associated with our imputation procedure is based on the definition of the double sampling strata being used and on the definition of the adjustment cells. Several constructive steps were taken to make the approach reasonable. In the initial stratification, an attempt was made to subsample only those records that were likely to have no changes or only small changes. Also, the adjustment cells were *subjectively* chosen to be homogeneous with respect to the magnitude of the relative editing change that might be made. In particular,



The coded tree branches above correspond to the following:

A = Retail, B = Wholesale, C = Transportation and Utilities, D = Other, E = Very Small, F = Small, G = Medium.

Figure 1. Hierarchy of Ratio Hot Deck Adjustment Cells

the adjustment cells are defined in terms of industrial classification, corporation size and the pattern of items present on the return. There were thirty categories defined by various industrial and size criteria (see Figure 1). In addition, sixteen item patterns were treated separately, defined by the presence/absence of Other Income (2 classes), the presence/absence of either Other Deductions or Other Costs of Goods Sold (2 classes), Other Current Assets or Other Assets (2 classes) and, finally, Other Current Liabilities or Other Liabilities (2 classes). The maximum number of adjustment cells was $30 \times 16 = 480$.

For each item pattern, a hierarchical structure was developed so that collapsing could be done when there were an insufficient number of donors for use in the imputation (see Figure 1). The first division is into financial returns (banks, insurance companies, etc.) versus non-financial records; cells are not collapsed across this division. The next levels of the hierarchy separate cases according to fairly broad industrial classes and according to the size of the corporation, in terms of assets and net income. Recall that the largest corporations are not subject to subsampling and, so, should not need imputation; hence, broad industrial and size groups seemed sufficient.

The quality of our estimation depends on how much collapsing takes place. In 1981, we had 36,586 returns with at least one schedule to impute, and 3,989 donors. For the non-financial returns we never collapsed across the major industrial classification, and, in fact, we always had some size distinction. Many cells were not combined at all, but maintained the maximum detail possible. In contrast, for financial returns the size variable was often lost by combining all cells, and major industries were sometimes combined (Hinkins 1983). For one pattern, all financial returns were combined into the same cell.

Based on our 1981 experience, several changes were made in the 1982 double sampling design:

- Due to the extensive collapsing of cells for financial returns in 1981, the subsampling rate for small financial returns was doubled to improve the estimates (from 10% to 20%, as noted earlier).

Table 1

Selected Statistics on Hot Deck Ratio Imputation, 1981-1982

Item	Tax Year 1981		Tax Year 1982	
	Financial	Non-financial	Financial	Non-financial
NUMBER				
Donors	908	3,081	1,806	4,697
Imputes	7,912	28,674	10,719	43,477
Adjustment Cells	113	238	142	260
DONOR CELL SIZE				
Average	8	13	13	18
Maximum	68	58	126	98
Minimum	1	1	2	2
DONOR-TO-IMPUTE RATIOS				
Average	.11	.11	.17	.11
Maximum	1.00	.25	2.00	.28
Minimum	.05	.05	.05	.05

Note: For 1982, cell sizes of 2 donors each were required in order to make possible the calculation of the variance.

- In 1981, the double sampling procedure was not applied across the entire sample, but was restricted to certain processing centers. Other processing centers collected all information, as before. In 1982, the procedure was applied across the whole sample. The relative number of records in1982 with some items imputed was 63 percent, compared to 40 percent in 1981.
- In order to estimate the hot deck imputation variance (Oh and Scheuren 1980; Rubin and Schenker 1986), an additional restriction was imposed on the 1982 design, in that we required that there be at least two donors in each adjustment cell. (See Table 1.)

In 1982, there were 54,196 records to be imputed from 6,503 donors, and there was considerably less collapsing of adjustment cells (Hinkins 1984). In particular, for financial records, 94 percent of the records imputed in 1982 were in adjustment cells defined with some size distinction, compared to 75 percent in 1981. Table 1 provides a selection of other statistics on the operation of the 1981 and 1982 systems.

4. INITIAL EVALUATION OF BIAS

The evaluation of the 1982 double sampling system is still underway, but some initial results are available on the potential biasing effects of the imputation. Bias should be small if R, the ratio of the editing change to the original amount, is always small, or if R is constant within adjustment cells. We have taken the approach of looking for the "worst" cases of bias by looking for examples where R is neither small nor constant. We confine attention to only two variables: Other Income and Business Receipts.

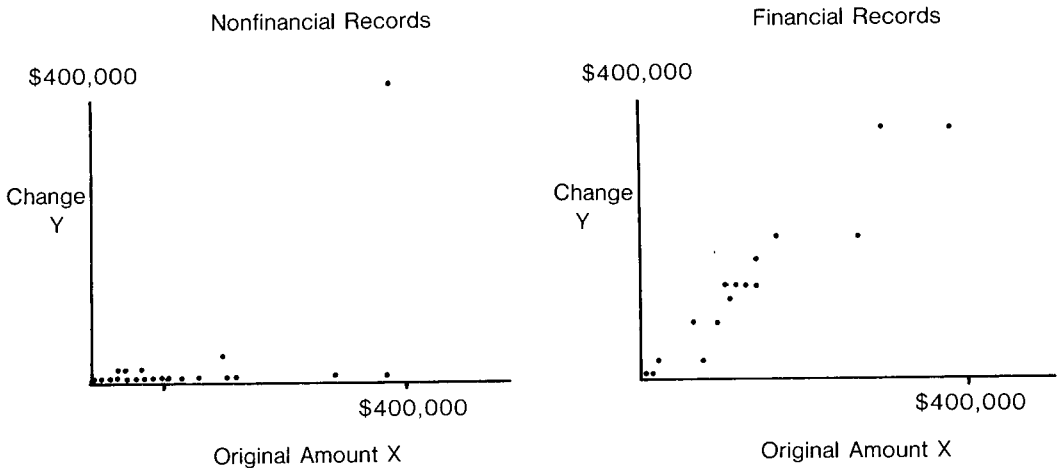


Figure 2. Changes in Other Income: Group B Donors only

Unbiased Model

The ratio bias in the hot deck imputation we are using would be zero if the relationship $Y = RX$ were to hold for all members of each adjustment cell chosen. An overall plot of the data might be useful, to look at the degree to which this model holds for Other Income. In Figure 2, therefore, we have plotted the Group B donors separately for financial and nonfinancial corporations. There is a distinct difference between these two categories. Nonfinancial returns are much less likely to change; in 1982, 14 percent of the nonfinancial donors had a change made to Other Income, compared to 59 percent of the financial records. Also, for financial returns at least, it looks as if the model $E(Y) = RX$ might be appropriate. Further work along these lines is intended, but the scatterplot encourages us to believe that, by and large, existing biases would be small.

Actual Bias Measures

Table 2 provides relative bias measures for selected worst case industries. These are shown for all returns in that industry and returns with assets under $25 million (i.e., for corporations likely to be most affected by the new procedures). Of the items changed in the double sampling the Other Income schedule showed some of the largest values of R and the most disperse distributions of R. The greatest change as a result of editing Other Income was made in the Business Receipt amount. It should be noted that the bias estimates in Table 2 are subject to considerable sampling error (Czajka 1986). Except for the very smallest amounts, however, it is conjectured that the estimates shown probably have the correct sign and are of the appropriate order of magnitude.

These examples indicate that within small subpopulations, there can be noticeable bias effects. However, even within a major industry, selected for its potential problems, the bias across all sizes is relatively small.

Table 2

Estimated Relative Biases for Business Receipts and Other Income by Selected Minor Industries, 1982

Selected Minor Industries	Business Receipts		Other Income	
	All Returns	Assets Under $25 Million	All Returns	Assets Under $25 Million
	(Biases as percent of applicable total)			
WHOLESALE TRADE				
Machinery, Equipment and Supplies	−1.40	−2.6	0.4	0.6
Miscellaneous Trade	−0.30	−0.5	−1.3	−2.4
RETAIL TRADE				
Auto Dealers and Service Stations	−0.30	−0.5	3.3	4.6
FINANCE AND INSURANCE				
Banking	−0.02	−0.7	0.1	2.4
Credit Agencies Except Banks	−0.50	−2.2	−0.9	−9.0
Insurance Agents	−0.60	−0.7	1.2	2.3

Note: All calculations are based on design-weighted estimates of the biases involved. The industries were selected to represent worst case examples.

Czajka's results (1986) indicate that for global estimates (across all industries), the bias effect of the imputation is small (less that 1% in all cases; considerably less than .05% in most cases).

There is no question that some of the biases in Table 2 appear large and warrant concern; however, it is important to realize that the overall effect on the root mean square error of the bias is small for all returns, generally 5% or less. These results give us strong evidence that the procedures employed did little or no harm to the data needed by our users; that, however, is not to say that major improvements, like those envisioned for 1985 and 1986, should not be made.

5. FUTURE PLANS AND SUMMARY

Double sampling and imputation were not used for the 1983 and 1984 samples because of processing constraints. They will be used again starting with the 1985 sample. As part of reinstituting the imputing process, we are planning to make several changes:

- It will no longer be necessary to initially transcribe certain items for statistical purposes before subjecting the records to double sampling. The fields needed are now being obtained directly from the IRS revenue processing system, so they are available before we begin reading and editing the tax return; thus, before editors first look at a return, we can designate whether or not they should review certain schedules. This makes the use of stratified double sampling even more appealing; the savings should increase.
- However, because of the new processing system, only three schedules are now available for subsampling. The schedules for 1985 are Other Income, Other Deductions and Other Costs of Goods Sold; the remaining four schedules used in 1981 and 1982 had to be dropped from the subsampling design.

- Despite the modest success of the 1981 and 1982 procedures, changes will be made for 1985 in the imputation methods. For example, the current definition of the adjustment cells could be improved, and separate imputation depending on the pattern of items represented needs to be reconsidered. The possible use of predictive mean matching within adjustment cells also bears examination (Little 1986). For 1986, refinements in the subsampling plan will need to be looked at too.
- Finally, we would like to base our estimates, in some way, on previous years' data, so as to be able to impute missing information earlier in the processing. In order to minimize the collapsing of adjustment cells, the 1981 and 1982 imputation processing had to wait for all records to be available. This delayed production by several weeks. We could avoid this problem by further increasing the number of donors; but, the editing of more records has the obvious disadvantage of increasing costs. On the other hand, by basing our approach in part on the previous year's data, we might not only improve the estimation, but also allow the imputation calculations to be done in the mainstream of processing.

Overall Summary

In this paper, we have described the reasons we had for making major changes in our statistical processing of corporate returns:

- The traditional complete data estimate was rejected in favor of double sampling because of cost considerations.
- The usual double sampling estimator (reweighting the complete data) was rejected because it did not result in a rectangular data set.
- A conventional hot deck approach was rejected because the resulting estimates could fail the edit checks.

Instead, the relative change was estimated using ratio hot deck imputation within adjustment cells.

We conjectured that because the double sampling procedure was restricted to a subset of the "small" corporations, the estimates of interest to our major users should be virtually unaffected; indeed, these estimates could even be improved, by better allocating our resources to validate and correct the records of the larger corporations. Our results so far largely vindicate these conjectures.

Compared to the traditional complete data estimator, the use of double sampling and hot deck imputation increased the mean square error of estimates in two ways; bias was introduced, and the variance of the estimator was increased. Our preliminary results indicate that there could be a significant bias effect for some estimates; however, the examples were chosen because they appeared to be cases where the hot deck ratio method would be weakest. Even so, the estimated overall effect of the procedure on the root mean square error appears relatively small. Looking at the increase in variance, the largest component is usually due to the decrease in sample size (double sampling). This increase in variance also turned out to be relatively small, since only one component of the final amount (the change) is imputed; the variance of the original values appears to dominate the variance of the changes.

In conclusion, while there are improvements to make, we feel encouraged to continue with our current double sample design and imputation technique. Perhaps at another Conference of this type we will be able to report on the further results of our research.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the considerable help they have received from the staff members in the SOI Division, whose day-to-day responsibilities are covered by the material presented here. We would also like to thank David W. Chapman and John L. Czajka for their many constructive comments, especially in clarifying our exposition. We, of course, accept full responsibility for any remaining obscurities or errors.

APPENDIX: SOME BASIC THEORY

This appendix provides some technical details on the double sampling procedure as applied in our particular situation. We contrast several potential estimators for the double sampling design we chose. An overall summary of the bias and variance expressions for these different approaches is found in Table A.

For this discussion, we ignore the underlying stratified sample design and act as if a simple random sample had been taken, or equivalently we consider estimates within a sampling stratum. To do otherwise would make the notation exceedingly complex, but would not change the main points we wish to make.

Let us again consider just one of the items subject to subsampling, namely Other Income as before. The variable of interest is Z, the final, corrected value of Other Income, and Z can be decomposed as

$$Z = X - Y,$$

where X = the original taxpayer (or revenue processing) value of Other Income,
Y = the change made to Other Income after reviewing the schedule.

The population values and parameters are indicated by upper-case letters and the sample statistics by lower case. The population parameters of interest are the finite population mean and variance, i.e.,

$$\bar{Z} = \sum Z_i/N = \bar{X} - \bar{Y},$$

$$S^2(Z) = \sum (Z_i - \bar{Z})^2/(N - 1).$$

Complete Sample – Prior to the introduction of double sampling, the estimates were calculated from a complete sample of size n', and the unbiased estimator of $\bar{Z}$ was

$$\bar{z} = \sum z_i/n'$$

$$= \bar{x} - \bar{y}.$$

Ignoring the finite population correction (N is large), the variance is

$$\text{Var}(\bar{z}) = S^2(Z)/n'.$$

Table A

Selected Properties of Alternative Estimators

Estimator	Bias	Variance	Satisfy Edit?
Complete Sample	0	$\text{Var}(\bar{z})$	Yes
Double Sample	0	$\text{Var}(\bar{z}) + c_1 S_B^2(Y)$	Yes
Hot Deck			
Amount (Y)	0[a]	$\text{Var}(\bar{z}) + c_1(1 + c_2) S_B^2(Y)$	No
Ratio (R)	b_1	$\text{Var}(\bar{z}) + V_1$	Yes
Combined Ratio	b_2	$\text{Var}(\bar{z}) + V_2$	Yes

[a] In general, the basic hot deck procedure is unbiased only when it results in final values that satisfy the edit checks.

In Table A, we use the properties of $\bar{z}$ as a benchmark, to compare among alternative estimators.

Double Sampling Estimation - Using Cochran's notation (Cochran 1977, 12.2), the original sample of size n' has now been stratified into the two groups A and B, with n_A' and n_B' units respectively. A subsample of size n_B is selected from group B. The original taxpayer amount X is recorded for all $n' = n_A' + n_B'$ records. The changes due to editing Other Income, Y, will be recorded for all n_A' units in group A and for the random subsample of n_B units in group B.

Since the double sampling procedure only applies to variable Y, within group B, the double sampling estimator of $\bar{Z}$ is

$$\bar{z}_d = \bar{x} - \bar{y}_d$$

$$= \bar{x} - \left(\sum y_{Ai} + (n_B'/n_B) \sum y_{Bj} \right)/n'$$

and $\bar{z}_d$ is unbiased.

Let N_B = number of population units falling in stratum B,
P_B = N_B/N, proportion of population falling in stratum B,
$\bar{Y}_B$ = population mean in stratum B,
$S_B^2(Y)$ = $\Sigma (Y_{Bi} - \bar{Y}_B)^2/(N_B - 1)$, $i = 1, 2, \ldots, N_B$,
$1/K$ = the subsampling proportion = n_B/n_B'.

If the sampling proportion, $1/K$, is assumed fixed (in our application, $1/K = .10$ or $.20$), it follows (Cochran 1977) that the unconditional variance of $\bar{z}_d$ is, ignoring the fpc,

$$\text{Var}(\bar{z}_d) = \text{Var}(\bar{z}) + c_1 S_B^2(Y),$$

$$= [S^2(Z) + P_B(K - 1) S_B^2(Y)]/\text{n}',$$

where $c_1 = P_B(K - 1)/n'$.

Therefore the price paid for the reduction in cost due to not editing every schedule, is the increase in variance due to double sampling. This increase in variance looks potentially damaging because K is large. However, recall that $Z = X - Y$, and the increase in variance is a function only of the variance of Y within subpopulation B. We expect $S^2(X)$ to dominate $S^2(Y)$, which should further dominate $S_B^2(Y)$, i.e.

$$S^2(X) >> S^2(Y) >> S_B^2(Y).$$

This is because the size of the variance is related to the mean value, and Y should be small compared to X. (For most items, we expect the amount misclassified to be small, compared to the original amount). Therefore we expect $S_B^2(Y)$ to be so much smaller than $S^2(Z)$ that $P_B(K-1)S_B^2(Y)$ will still be relatively small compared to $S^2(Z)$, and so the increase in variance due to subsampling will be relatively small. This is not guaranteed, but Czajka's results bear this out, for most items (Czajka 1986).

Hot Deck Imputation - Hot deck imputation was used, within adjustment cells, to reconstruct a rectangular data set. In particular, a return with schedules to be imputed was matched to a donor in group B, in the same adjustment cell, with these same schedules edited.

Imputing the missing values of y with a hot deck procedure, using simple random sampling, further increases the variance over using the double sampling estimate ($\bar{z}_d$). However the additional increase in variance due to using hot deck imputation is small compared to the increase due to double sampling. This relative increase in variance due to imputing, denoted as c_2 in Table A, is bounded and in our case is small. (When $K \geq 2$, $c_2 \leq 0.125$. See, for example, Hansen, Hurwitz, and Madow 1953).

As discussed in the paper, there is a problem with using an ordinary hot deck approach. If we simply estimate the unobserved y_i value, on record i, with the observed value y_j from donor record j, then the resulting estimate of the final value z_i may not satisfy the edit checks. Additional corrections would have to be made to the record. Since the original amount is always observed, it seemed more reasonable to "hot deck" the relative change $R = Y/X$ rather than the actual change Y. In addition to satisfying the edit checks, we expected the ratio procedure to reduce the variance of our estimates relative to the basic hot deck approach; however the variance of our estimator is not analytically tractable and must be measured empirically. Also, by introducing the ratio, our estimators are now biased. We conjectured that the biases would be small and in fact they were, for the most part, as seen in Table 2. In practice, the hot deck imputation was done within adjustment cells, created by post-stratifying the records into what we hope are homogeneous cells. The effect of this post-stratification should be to reduce variance and bias effects, but that is dependent on our skill in defining the imputation cells (an area with ample room for additional work).

Ratio or Regression Estimation - We are also considering ratio (or regression) estimates within cells, instead of the hot deck estimates. For example, $\hat{z} = x_i - \hat{r}\,x_i$, where $\hat{r} = \bar{y}/\bar{x}$ is calculated within appropriate cells. Referring to Table A, the increase in variance, V_2, using the ratio estimator could be approximated using the formulas for the ratio estimator (e.g., Cochran 1977). However, these formulas are large sample approximations, and our sample sizes are almost always quite small. (In this case, the sample size is the number of donors, n_B, in an adjustment cell.) Therefore, empirical results are needed here.

Similarly, the bias, b_2, can be found using the results for ratio estimators. Unlike the hot deck ratio, the bias of the ratio estimator goes to zero as the sample size increases and in this sense the ratio estimator is more robust. In fact, the hot deck ratio estimator is unbiased only if the model $Y = \beta X$ is correct. (Of course, the bias of both estimators goes to zero as the fraction of missing data goes to zero). However, even if the model $Y = \beta X$ is incorrect, the ratio estimator is consistent.

There are of course many other options; multivariate regression models could be investigated. We are still in the early stages of this project and we certainly have our work cut out for us now and in the upcoming years.

REFERENCES

AMBROSE, P. (1985). Tax year 1985 business finance (T2) sample selection: detailed statement of requirement. Statistics Canada (Unpublished).

BARKER, D., HINKINS, S., and REHULA, V. (1982). 1981 corporation validation tests. Statistics of Income Division, Internal Revenue Service (Unpublished).

BURPEE, J., and McGRATH, A. (1982). Micro-model of corporation taxation sample design and estimates. Statistical Services Division, Revenue Canada Taxation (Unpublished).

CLICKNER, R.P., GALFOND, G.J., and THIBODEAU, L.A. (1984). Evaluation of the IRS corporate SOI sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 443-448.

COCHRAN, W.G. (1977). *Sampling Techniques,* (3rd ed.). New York: John Wiley and Sons, Inc.

COLLEDGE, M., JOHNSON, J., 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. (See also the paper by S. Michaud in this issue.)

CYS, K., HINKINS, S., and REHULA, V. (1982). Automatic and manual edits for corporation income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 443-448.

CZAJKA, J. (1986). Imputation of selected items in corporate tax data: improving upon the earlier hot deck. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* (in publication).

FORD, B.L. (1983). An overview of hot deck procedures. In *Incomplete Data in Sample Surveys,* Volume 2 - Theory and Bibliographies (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 185-207.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory,* Vol. II. New York: John Wiley and Sons, Inc.

HINKINS, S. (1983). Matrix sampling and the related imputation of corporate income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 427-433.

HINKINS, S. (1984). Matrix sampling and the effects of using hot deck imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 415-420.

JONES, H., and McMAHON, P. (1984). Sampling corporation income tax returns for statistics of income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 437-442.

LESZCZ, M.R., OH, H.L., and SCHEUREN, F.J. (1983). Modified raking estimation in the Corporate SOI Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 434-438.

LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. Presented at the Second Annual Census Research Conference, March 1986. To appear in the *Journal of Business and Economic Statistics.*

OH, H.L., and SCHEUREN, F.J. (1980). Estimating the variance impact on missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 416-420.

POWELL, W.T., and STUBBS, J.R. (1981). Using business master file data for statistics of income purposes. *Statistical Uses of Administrative Records: Recent Research and Present Prospects,* Vol. 1., Washington, DC: Internal Revenue Service, 157-167. See, especially, the Appendix by Alan Freiden.

RUBIN, D., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association,* 81, 366-374.

SANDE, I.G., (1982). Imputation in surveys: coping with reality. *The American Statistician,* 36, 145-152.

STRUDLER, M., OH, H.L., and SCHEUREN, F.J. (1986). Protection of taxpayer confidentiality with respect to the IRS Individual Tax Model. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* (in publication).