MATRIX SAMPLING AND THE EFFECTS OF USING HOT DECK IMPUTATION

Susan M. Hinkins, Internal Revenue Service

From an annual sample of U.S. Corporate Tax Returns, the Internal Revenue Service provides estimates of population and subpopulation totals for several hundred financial items. The basic sample design is highly stratified and relatively complex [1, 5]. Starting with the 1981 sample, the design was modified to include matrix sampling--those items not observed in the subsample are predicted using an imputation procedure. This paper gives some further results of that recent modification [3].

1. MATRIX SAMPLING

Approximately three million corporate tax returns were filed in tax year 1981 and the Statistics of Income sample contained approximately 95,000 of these returns. Prior to the 1981 sample, the sample design resulted in the usual rectangular data base. The modification to include matrix sampling was prompted primarily by budget and resource constraints.

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 abstracted; they are also corrected and redistributed to compensate for taxpayer errors. This process of abstracting, correcting, and redistributing the tax return data is referred to as "editing" the return. The cost of editing varies by the complexity of the return; it may take only twenty-five minutes to edit a very simple return and as long as a week to edit a complicated one. The quality of the editing is vital to our estimates, as these checks reduce, but do not eliminate, the nonsampling error.

The population is highly skewed; a relatively small number of very large corporations dominate the estimates and are selected with certainty. Nonsampling errors on these records can have a significant effect on our estimates.

Consequently, in order to distribute our time and effort more effectively, stratified matrix sampling was introduced for the smaller returns, i.e., certain data items are retrieved on only a subsample of the sampled returns. The stratification is two-dimensional: first, in terms of which schedules can be subsampled and, secondly, in terms of which returns are subject to subsampling.The sample now has a form similar to that shown in Figure 1. The definition of the records as "crucial" (Group A returns) versus "less important" (Group B) is directly related to the choice of the schedules being subsampled. Group A, the "crucial" corporations, includes not only the very large corporations but also corporations, of any size, for which we believe these schedules are significant. Group B returns, i.e., the residual, are the only ones subject to subsampling.

When a schedule is edited, one often finds that the taxpayer has incorrectly classified certain amounts. For example, the Other Income schedule is simply a detailed list of the income that the taxpayer put in the catch-all variable "Other Income." It may be that the taxpayer included $500 in income from business receipts that should be included under the item Receipts, rather than Other Income. We would subtract $500 from the variable Other Income and increase the variable Receipts by $500. Seven such schedules are being subsampled [3]. The variables of interest are the final amounts in these fields. For example, the final amount in Other Income is equal to the original amount minus the changes due to editing the Other Income schedule. That is, Final Other Income equals:

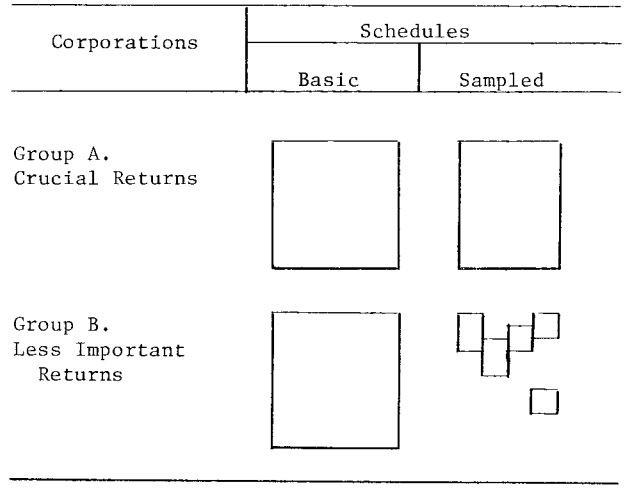
$$\begin{bmatrix}\text{Original}\\ \text{Other Income}\end{bmatrix} \text{ minus } \begin{bmatrix}\text{Changes due}\\ \text{to editing}\end{bmatrix}.$$

Similarly, Final Receipts equals:

$$\begin{bmatrix}\text{Original}\\ \text{Receipts}\end{bmatrix} \text{ plus } \begin{bmatrix}\text{Changes in Receipts}\\ \text{due to editing}\\ \text{Other Income schedule}\end{bmatrix}.$$

The original amounts are observed for every return. The variables being subsampled for returns in Group B are the changes that would be made if the Other Income schedule were edited. Group B includes returns for which we believe that this change is relatively small; i.e., cases where the final amount is either very small or is dominated by the original amount.

Figure 1.--Data Pattern with Matrix Sampling



2. IMPUTATION PROCEDURE

The usual estimation technique associated with matrix sampling is to reweight the subsampled items [2]. However, in a large sample survey such as this, allowing different weights for different items is impractical. Therefore we

retain our rectangular data set by imputing the missing information, i.e., by imputing the changes.

These changes are imputed using a hot deck procedure within adjustment cells. A record with items to be imputed is matched to a record, in the same adjustment cell, with complete information (a donor). Since the original amounts are observed on all records, it seems logical to "hot deck" the relative change rather than the dollar amount of the change. In particular, the imputed value of Final Other Income, for record "i," would be obtained by the expression:

$$(1 - C_d) * \left[\text{Original Other Income}_i\right] ,$$

where C_d is the ratio

$$\frac{\text{Change on donor record d}}{\text{Original Other Income on record d}}$$

and "d" is the donor record, with complete information, that was matched to record i as part of the hot deck process. Using the relative change should reduce the coarseness of the hot deck procedure and should almost eliminate further corrections to balance the record. This procedure is described in more detail in [3].

Prior to the 1981 sample, the estimation procedure was design-based. The estimates were calculated by weighting the observed (sampled) values; inference was based solely on the distribution of the design indicators. The relative merits of this classical type of inference versus model-based inference have been discussed in the recent literature [6]. A model-based estimate predicts the unobserved values using an assumed model; inference depends on both the distribution of the design indicators and the model for the variable. Our estimation procedure is still primarily design-based, but it now contains a modelling aspect. Each record is assigned a weight, based on the original stratified sample design. Those items that are not observed in the subsequent matrix sample are imputed as described above.

Imputation is a model-based approach usually associated with nonresponse. Because this "nonresponse" mechanism is, in fact, part of the sample design, the mechanism is known and it is just another level of random sampling. The model associated with the imputation procedure is contained in the definition of Group B and in the definition of the adjustment cells. As mentioned earlier, stratified matrix sampling only subjects to subsampling those records that are likely to have small changes made due to editing the schedules. For those records subject to subsampling, the adjustment cells are defined so that they should contain records that are homogeneous with respect to the variable being imputed (i.e., the relative changes due to the schedule).

The underlying rationale for our approach is that if the amounts being changed will generally be small (with the original amount dominating the change), then the effect of the subsampling and imputation on our population estimates should be small, as well. Good results are not guaranteed, however.

Two cases need to be considered. For variables such as Other Income, there is some bound on the relative changes. The relative change must lie between 0 and 1, because the amount changed must lie between zero and the original amount claimed. For a variable such as Receipts, there is no intrinsic bound on the relative size of the change. The amount being added need not have any relationship to the amount originally in Receipts; an original zero amount can even be changed to a nonzero amount. So even if a small amount is added, it can result in a large relative change. There is, thus, a possibility of making significant changes in the microdata with potentially adverse consequences for estimates of subpopulations. For example, imputing $100 into Receipts when it was originally zero will not significantly change the estimated population totals. But if a user is interested in subpopulations defined by whether or not there are Receipts, the imputation may be a significant factor.

3. EFFECTS ON POPULATION ESTIMATES

Consider first the estimates of population totals or subpopulation totals. Unlike most applications of a hot deck procedure, in this problem the nonresponse mechanism is known. In the terminology of missing data, the data are missing at random [8]. This is not an assumption, but a consequence of the design. Because the data are missing at random, there is no bias due to the nonresponse generated by the matrix sampling. However, because the change is estimated as a ratio, the imputation procedure will introduce bias, unless, within adjustment cells, (1) the expected value of the change is a constant multiple of the original amount, or (2) the change is independent of the original amount. Neither of these possibilities is likely. However, we expect the bias due to this technique to be insignificant, because only a relatively small component of the final amount is being imputed.

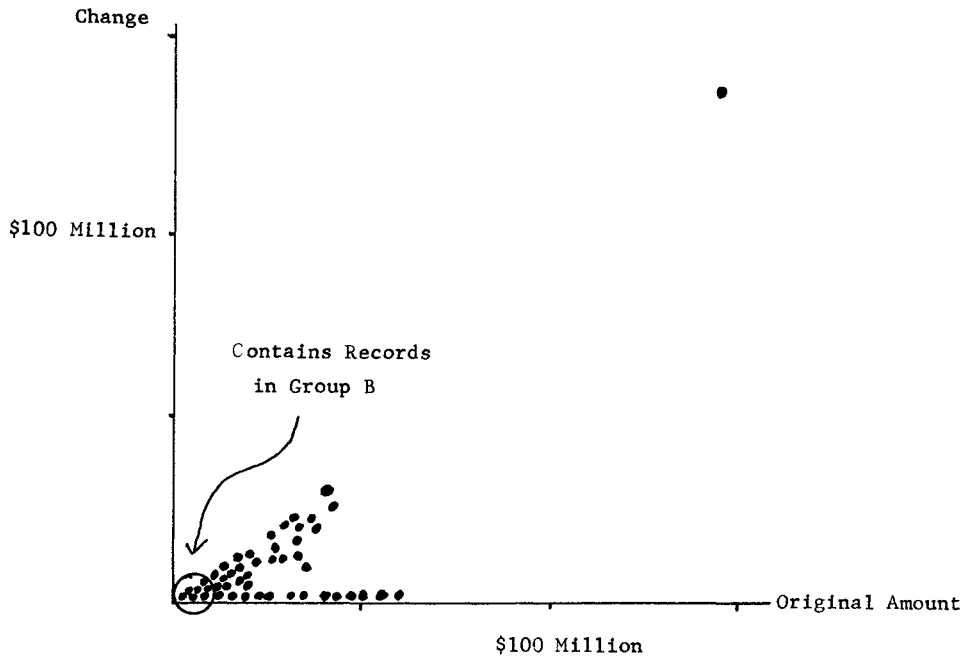
Using matrix sampling would increase the variance of our estimates because of the decrease in sample size. Recall again that the variable being estimated, z, has two components; for example, if

$$\text{Final} = \text{Original} - \text{Change},$$

then $$z = x - y,$$

and the increase in variance due to subsampling is a function only of the variance of y within Group B [3]. Using a hot deck imputation procedure, instead of reweighting the observed items, further increases the variance. Estimating the additional variance due to imputation is one of the difficulties associated with imputation procedures. If a user treats the data set as if it were completely observed, the variance of the estimate will be underestimated and incorrect inferences may result [9]. A better estimate of variance can be obtained using multiple imputation [7, 9].

Figure 2.--Changes Due to the Schedule: All Records with Other Income, 1981



An Example.-- Using multiple imputation procedures, estimates of the bias and estimates of the increase in mean square error due to the subsampling and imputation are being calculated. Summaries should be available in a subsequent report this year. However, we expect both the bias and the increase in variance to be relatively small, because only a small fraction of the Final Amount is being imputed.

For example, last year the results of a pilot study were given [3]. The results from the complete data set are now shown in Figure 2. This is a plot of the amount changed vs. the amount originally claimed as Other Income, for records in both Group A and Group B. Most of the records have an original Other Income less than $50 million, but there are some larger records. One record had an original amount of $144 million and $140 million was removed after editing the schedule.

Ideally we would like the records subject to subsampling (Group B) to have both a small amount changed and a small change relative to the original amount. The records in Group B are entirely contained in the small circle around zero; the dollar amounts of possible changes have been reasonably well controlled because only "small" records are in Group B. Therefore, the bias and increase in variance should be small, at least for population estimates.

On the other hand, we have been less successful in predicting which records will have a small relative change. That is, the relative change appears to have the same range (between zero and the original amount) for records in Group B as for records in Group A. We would have preferred that the relative changes on records subject to subsampling would have been smaller, closer to zero.

In Figure 3, the Group B donors are plotted separately for financial records and for nonfinancial. Financial corporations include banks, real estate, insurance, and holding companies. Nonfinancial corporations encompass everything else - manufacturing, agriculture, etc. Since for the Other Income Schedule there is a distinct difference between these two categories, we have found that the type of corporation is a reasonably good predictor of which records will change. Financial records are likely to change due to editing the schedule; nonfinancial records are not.

4. TWO EXAMPLES OF THE IMPACT OF THE IMPUTATION

Two aspects of the imputation have been discussed:

(1) the absolute size of the change (are large amounts being imputed?), and

(2) the relative size of the change, compared to the original amount (are we making changes of 0%, 10%, or 200%?).

For the Other Income Schedule we saw that the dollar amounts of possible change have been reasonably well controlled. In this section we will look at two examples of the change relative to the original amount.

Two classes of records are chosen, one to exemplify a potentially "bad" case and the second class to illustrate a "good" case. We saw in

Figure 3.--Changes Due to Schedule: Group B Donors Only, 1981

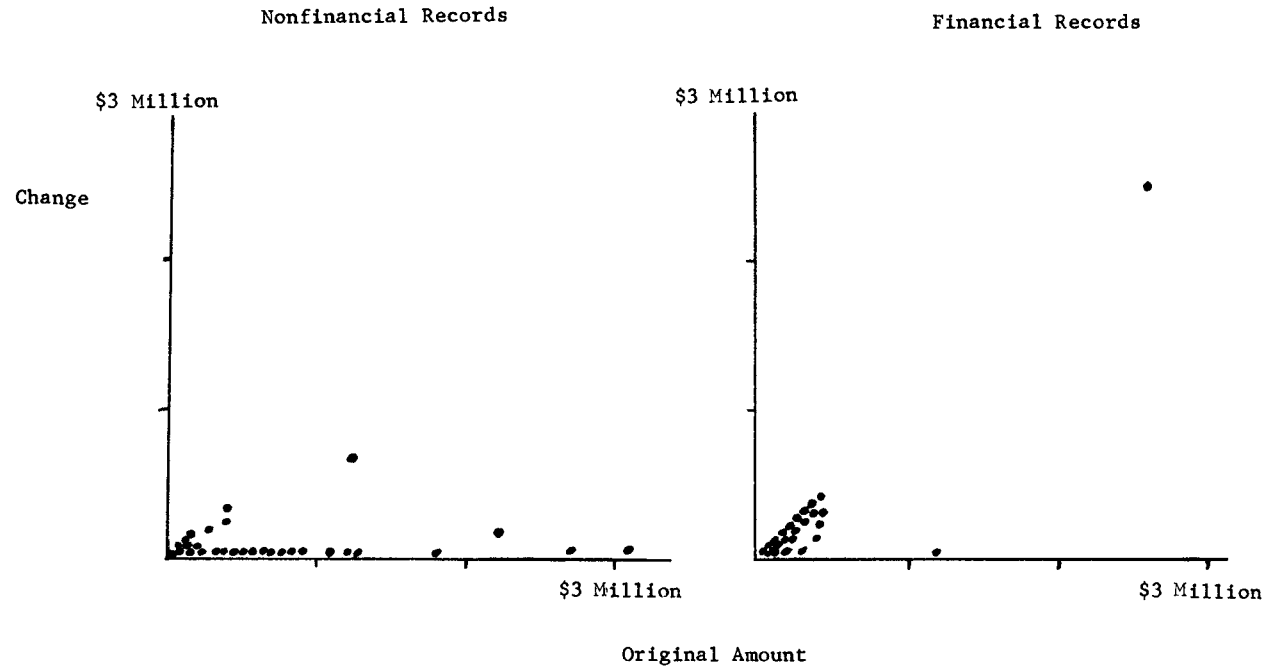


Figure 3 that, for donor records (Group B), the financial records are very likely to have a change made; approximately 2 out of 3 records had a change due to editing the Other Income Schedule. Within the financial classification, the largest major industrial classification was banks: of the 369 donor records classified as financial, 241 were banks. The banks are even more likely to have a change made to Other Income; 9 out of 10 banks had a change made due to editing Other Income. Therefore, banks were chosen as the first example, an example of potential problems of distorting the microdata because of making large relative changes.

In Figure 3, we also saw that the nonfinancial records were unlikely to have a change made in Other Income; for donor records, only 1 out of 10 records had a change. Therefore, the nonfinancial records were selected as our "good" case. (This is a substantially larger class, with 1441 donor records.)

Finally, instead of considering the relative change in the variable Other Income, we look at the relative change in Receipts. This is done for several reasons. When Other Income is changed, amounts are moved out of Other Income and into other variables. The variable most often changed in this way is the variable Receipts. Unlike Other Income, there is no bound on the relative change made to Receipts, which is defined as

$$\frac{\text{Change in Receipts}}{\text{Original Receipts}},$$

except that it cannot be negative. A change can be made even when the original amount in Receipts is zero. (When this happens, the relative change is considered to be "infinite.")

Example 1. - Banks. -- The distribution of the relative changes made to Receipts, for Group A and Group B, is shown is Figure 4. The values of the relative change have been grouped into intervals and the last category is for records that have changes made to an original zero amount. An asterisk indicates a percentage less than 1%.

For example, in Group A (records not subject to subsampling), 10% of the records had no change, 2% of the records had a relative change greater than 0 but less than or equal to 1/2, and on 86% of the records the original amount was zero and a change was made.

For banks, the only noticeable difference between the two groups is that for records not subject to subsampling (Group A), there are a few records with relative changes of 100 and more. However, the important point is that even in the records subject to subsampling, most records have Receipts changed from zero to a positive amount.

Figure 4.--Change in Receipts, for Banks, 1981

Matrix Sampling Stratum	Interval Values of the Relative Change (Upper Endpoint)						
	0	½	1	20	100	Over 100	Infinite (Change 0)
Group A---	10	2	1	*	*	*	86%
Group B---	12	3	0	1			84%

* indicates less than 1%

Therefore, for banks, the imputation procedure may:

(1) cause distortion of the distribution within microdata, and
(2) even have a significant effect on the population estimates, if the amounts changed are large.

However, at least the latter does not happen.

Figure 5 shows the mean values for selected income items for banks in Group B. As one would expect, the dominant income item is Interest. Receipts is a small source of income, and the amounts being moved into Receipts are smaller still; they include items such as charges on checking accounts, late charges, etc. For the variable Receipts, the original amount still dominates the change over aggregates. Also, the banks subject to subsampling are small (average Total Income less than $3 million). Therefore, though the microdata for (small) banks may be impacted by the imputation, most population and subpopulation estimates should not be significantly altered.

Figure 5. -- Income Items on Banks Subject to Subsampling, 1981

Income Items	Average Value in Dollars
Receipts - Original Amount	121,000
Amount Changed	79,000
Interest	2,394,000
All Other Income Items	258,000
Total Income	2,852,000

Example 2. - Nonfinancial Records. -- The nonfinancial records exemplify, for the variable Receipts, what we hoped would happen. The distribution of the relative change is summarized in Figure 6. For nonfinancial records relatively few changes were made due to the Other Income Schedule, so there should be few changes to Receipts; what is of interest is whether there are large relative changes. In Group A, the change is dominated by the original amount for most records; the relative change is less than or equal to 1/2 for 96% of the records. But on 3% of the records a change was made when the original amount was zero, and there are several records with very large relative changes. For example, there is one record where editing the Other Income Schedule added over 500 times what was originally in Receipts.

Figure 6.--Change in Receipts for Nonfinancial Records, 1981

Matrix Sampling Stratum	Interval Values of the Relative Change (Upper Endpoint)						
	0	½	1	20	100	Over 100	Infinite (Change 0)
Group A---	82	14	*	*	*	*	3%
Group B---	92	7					*

* indicates less than 1%

Notice however that for the records subject to subsampling, the nonfinancial records exhibit the desired property - the original amount dominates the change. The relative change is less than 1/2 for over 99% of the records. The imputation should not severely distort the distribution, even within microdata sets.

5. SUMMARY AND FUTURE PLANS

Matrix sampling and the subsequent imputation of the unrecorded amounts were introduced in order to expend our resources more efficiently. Based on preliminary analyses, we do not expect the imputation to significantly effect the estimates of important population and subpopulation aggregates. Estimates of bias and variance are being calculated.

We also hoped that the imputation procedures would not severely distort the distributions within microdata sets. While we expect this to be true for most variables and most subpopulations, it is not true for all. (Clearly for some small banks the amount changed dominates the original amount.)

Naturally, we have many future plans, in addition to improving our models. A serious consideration is the dual problem of having enough complete records (donors) in an adjustment cell and performing the imputation calculations early in the processing. In the first year of matrix sampling (the 1981 sample), the number of donors was so inadequate that the adjustment cells were very broad, especially for financial records. That is, there was severe collapsing of cells in order to attain an adjustment cell with at least one donor. This was discussed last year and a summary of the collapsing was given [3]. For the 1982 sample, more records were subject to subsampling and the subsampling rate was doubled for financial records. This was successful in that much less collapsing occurred and each cell had at least two donors [4]. However, the ratio of donors to imputes can still be quite small. Also, because the ratio of donors to imputes is small, the imputation processing must wait for all records to be available. This can delay production by several weeks. Increasing the percentage of donors by editing more records has the disadvantage of increasing costs. We would like to use the prior year's complete records to impute the current year's records. If this is reasonable, it would increase the number of complete records in the adjustment cells, and allow the imputation calculations to be done in the mainstream of the processing.

In conclusion, while there are many improvements to make, we feel encouraged to continue with this type of sample design and imputation technique.

ACKNOWLEDGMENTS

The author would like to thank Karen Cys and Dave Barker, for their continued collaboration on the work presented here; Wendy Alvey and Beth Kilss, for help in preparing the ASA presentation and paper; and Linda Daniel, for typing the paper.

REFERENCES

[1] Clickner, R.P., Galfond, G.J. and Thibodeau, L.A. (1984). "Evaluation of the IRS Corporate SOI Sample," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1984.

[2] Cochran, W.G. (3rd ed., 1977), Sampling Techniques, Wiley, New York.

[3] Hinkins, S. (1983). "Matrix Sampling and the Related Imputation of Corporate Income Tax Returns," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1983.

[4] Hinkins, S. (1984). "Matrix Sampling and the Effects of Using Hot Deck Imputation," Statistics of Income and Related Administrative Record Research: 1984, 1984.

[5] Jones, H. and McMahon P. (1984). "Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1984.

[6] Little, R.J.A. (1982). "Models for Nonresponse in Sample Surveys," Journal of the American Statistical Association, 77.

[7] Oh, H.L. and Scheuren, F.J. (1980). "Estimating the Variance Impacts on Missing CPS Income Data," American Statistical Association, Proceedings of the Section on Survey Research Methods, 1980.

[8] Rubin, D. (1976). "Inference and Missing Data," Biometrika, 63.

[9] Rubin, D. (1980). Handling Nonresponse in Sample Surveys by Multiple Imputations. U.S. Department of Commerce, Bureau of the Census Monograph.