MATRIX SAMPLING AND THE RELATED IMPUTATION OF CORPORATE INCOME TAX RETURNS

Susan M. Hinkins, Internal Revenue Service

Based on an annual sample of corporate tax returns, the Statistics of Income Division (Internal Revenue Service) publishes estimates of income and other financial items. This is an expensive process due largely to the manual abstraction, review, and correction of data items from each sampled tax return.

As in every operation, we are constantly trying to improve our data base, while working within a limited and often reduced budget. Starting this year, we are using matrix sampling; that is, we are retrieving certain data items on only a subsample of the sampled returns. We are currently imputing the results for the other records using a hot deck procedure within adjustment cells. This will reduce our costs with what we hope will be only a minor loss of information or precision.

The procedure being employed involves two related modelling problems:

(1) Determining which records should be subsampled; and,

(2) Imputing the missing information for the records not selected for complete editing.

A brief overview of the problem is given in Section I. Section II describes the problem in terms of double sampling and Section III describes the mechanics of the imputation procedure. The results of some preliminary analyses are given in Section IV. Section V outlines our future plans and expectations.

I. OVERVIEW

There are certain items on the tax return for which the taxpayer must supply additional information on an attached schedule. One such item on the corporate return is "Other Income." If an amount is reported as Other Income, a schedule must be attached showing further detail. We "edit" this schedule to determine if the taxpayer is correct, or if some of the items claimed as Other Income should be shown elsewhere or combined with more clearly defined income items.

FIGURE 1.--HYPOTHETICAL OTHER INCOME SCHEDULE

Taxpayer's Description	Amount	Correct Field
Other Income, total .	$1600	
Carrying Charges	500	Other Interest
Bank Deposits	400	Other Interest
Interest, U.S. Gov't obligations ..	200	Interest on U.S. Gov't obligat-tions
Claims Income	500	Other Income

An example is given in Figures 1 and 2. Other Income was reported as $1600 and the taxpayer attached the schedule shown as the first two columns of Figure 1. The third column of Figure 1 shows how each item should have been classified, according to the editor. For example, Bank Deposits should have been included under Other Interest instead of Other Income.

Figure 2 shows some of the items on this hypothetical tax return, as originally reported by the taxpayer and after the adjustment due to reviewing the Other Income schedule.

FIGURE 2.--INCOME STATEMENT FOR RETURN WITH OTHER INCOME SCHEDULE SHOWN IN FIGURE 1

Selected Fields	Original Tax Return	Recorded (After Review)
Total Income........	$9000	$9000
Business Receipts..	600	600
Total Dividends....	400	400
Other Interest	200	1100*
Interest on U.S. Gov't Obligations.	0	200*
Rents.............	5000	5000
:	:	:
Other Income.......	1600	500*

*Indicates amounts as changed after examining the Other Income schedule

There are seven schedules (such as Other Income, Other Liabilities, Other Assets, etc.) that are being subsampled. The purpose of looking at these schedules in only a subsample is to reduce the processing costs with as little loss of information or precision as possible. The schedules are attached on separate sheets of paper and they may consist of handwritten descriptions, with no standard form or length. Reviewing these schedules is a distinct, separable procedure. Ideally, we would like to review only those schedules that will result in a change. The basic plan is then to edit all schedules that have a high probability of redistributing some or all of that amount. The other records will be subsampled.

Unfortunately, prior to this year, we had no information regarding the type or amount of adjustments being made by editing these schedules. The editors only recorded the final result. The original fields, as claimed by the taxpayer, were not recorded. For example, in Figures 1 and 2, traditionally we would have recorded only the amounts in the last column of Figure 2.

Under the revised processing system, the abstraction of data from the tax return is now done in stages, and certain items are initially transcribed directly from the return. Using automatic tests, items or schedules are then flagged for abstraction or further scrutiny in later stages [2]. For the seven schedules of interest, this new strategy allows us to do two things:

(1) Retain original taxpayer information as reported so that the amount of change can be evaluated.

(2) Decide whether or not to review these schedules based on initial information transcribed for each record.

Consider the Other Income example: if it were processed this year, all of the information in Figure 2 would be available.

II. DOUBLE SAMPLING

Our problem falls naturally into the framework of double sampling for stratification. Other Income will be used for illustration.

In our initial data capture, we record certain variables, say u and x, from each record. The variable u includes descriptive or stratifying items such as industrial classification. The variable x is the original amount claimed as Other Income. Let y denote the change that would be made to Other Income due to editing the schedule.

Based on the values of the variables u and x, the population can be stratified into two groups, say A and B, where we believe that group A will contain records that are likely to be changed due to editing the Other Income schedule ($y = 0$) or records that are especially important (such as large, well-known corporations). For example, records with an amount in Total Assets of $250 million or more would be put in group A. Or if the original amount in Other Income is large compared to the amount in Total Income, the record is in group A. Since Other Income is one component of Total Income, we assume this is an indication that the taxpayer has incorrectly designated items as Other Income; and editing this schedule is likely to result in some redistribution. We do not want to impute relatively large amounts.

So far our criteria for defining the two groups has been based entirely on subject matter expertise. (A more complete description of the current definition of group A is available in [10]).

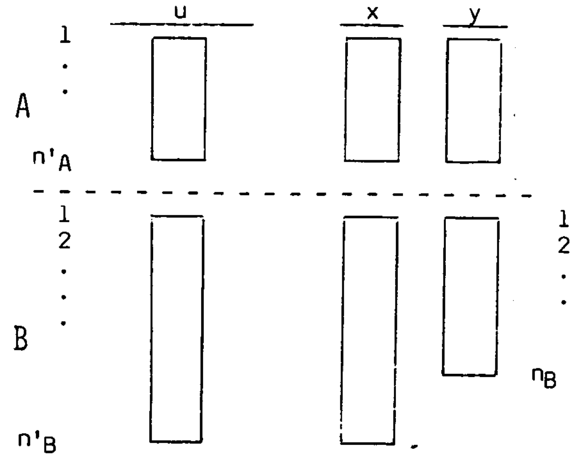
Assume our original sample of size n', containing u and x, has now been stratified into two groups A and B. The variable y (changes due to editing Other Income) will be recorded for all units in group A and for a random subsample of units in group B. This is the classical double sampling for stratification [1]. Following Cochran's notation, let

n'_A , n'_B = number of units in the first sample falling in strata A and B respectively

n' = $n'_A + n'_B$

n_B = number of units in the second sample drawn from stratum B

Therefore, our sample will consist of:



We are interested in estimating $\bar{Z} = \bar{X} - \bar{Y}$, the final, "corrected" amount assigned to Other Income. Let

$1/K = n_B/n'_B$, the sampling proportion

$\bar{x} = \Sigma x_i/n' \quad i=1,2,\ldots n'$

$\bar{y}_{ds} = (\Sigma y_{Aj} + K\Sigma y_{Bm})/n'$

$j=1,2,\ldots n'_A$

$m=1,2,\ldots n_B$

Then $\bar{y}_{ds}$ is the usual double sampling estimate of $\bar{Y}$. The associated estimate of $\bar{Z}$ is then

$$\bar{z}_{ds} = \bar{x} - \bar{y}_{ds}$$

and is unbiased.

Let Var(Z) denote the population variance of Z. Let

N_B = number of units in population falling in stratum B

$P_B = N_B/N$, proportion of population falling in stratum B

$\bar{Y}_B$ = population mean in stratum B

$\text{Var}_B(Y) = \Sigma(Y_{Bi} - \bar{Y}_B)^2/(N_B-1)$, $i=1,\ldots N_B$, the variance of Y in stratum B.

The sampling proportion, 1/K is assumed fixed (in our application 1/10). It follows [1] that the unconditional variance of $\bar{z}_{ds}$ is

$$\left(\frac{1}{n} - \frac{1}{N} \right) \text{Var}(Z) + P_B\ (K-1)\ \text{Var}_B(Y)/n'$$

Therefore, the increase in variance due to double sampling is

$$P_B\ (K-1)\ \text{Var}_B(Y)/n' \ = \ \text{Var}_B(Y)/n_B$$

This increase in variance is the price paid for

FIGURE 3.--CHANGES DUE TO THE SCHEDULE: ALL RECORDS WITH OTHER INCOME

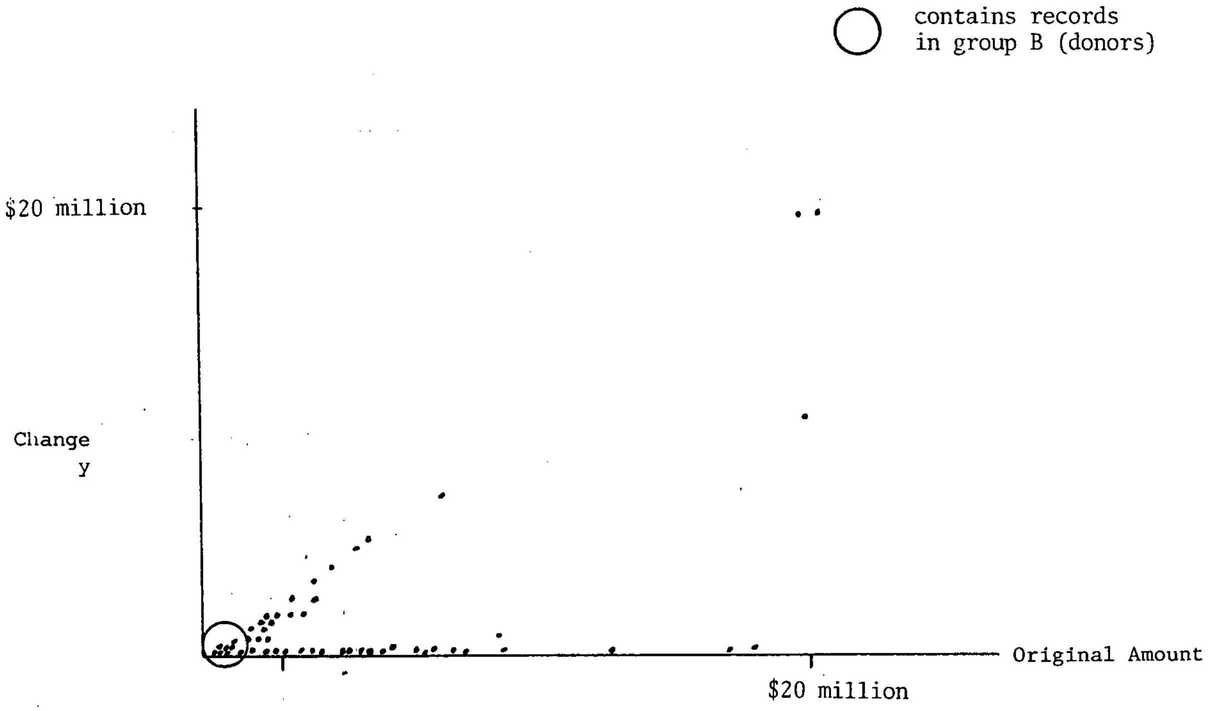
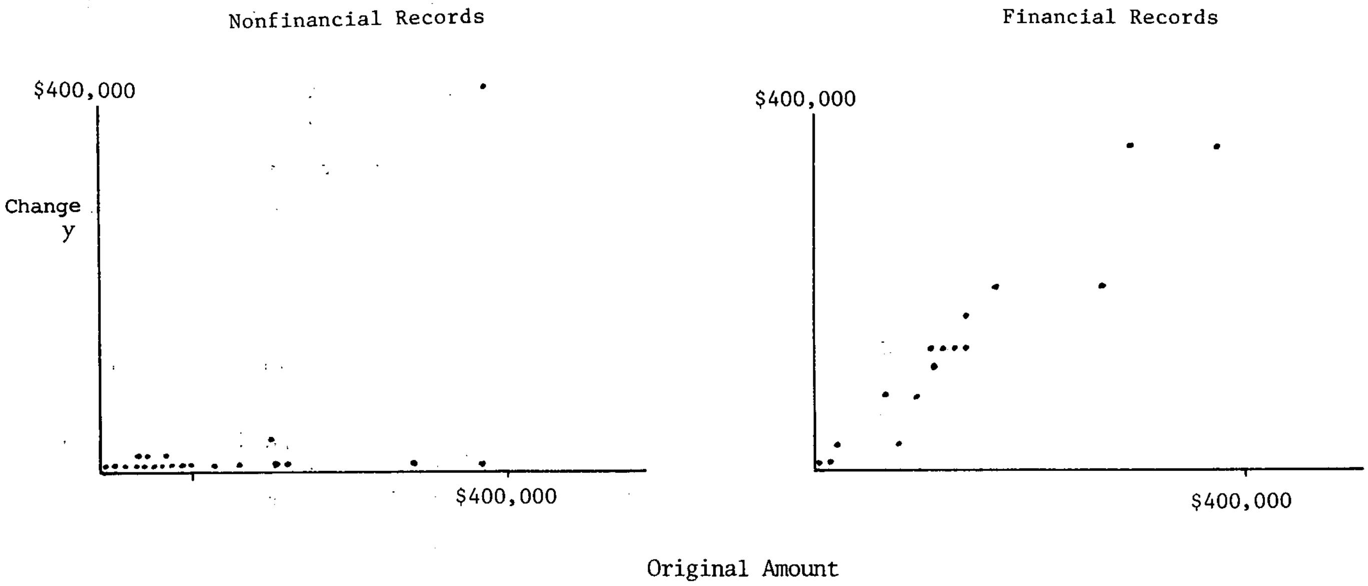
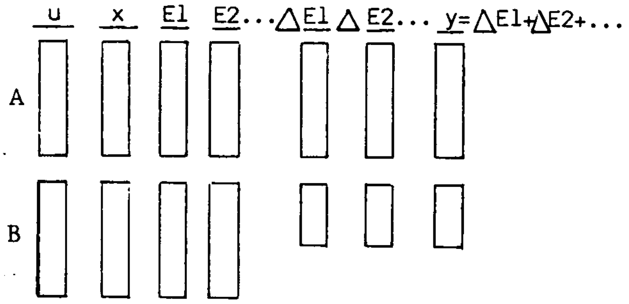


FIGURE 4.--CHANGES DUE TO SCHEDULE: GROUP B DONORS ONLY



the reduction in cost due to editing the schedule for only a subsample of the records in group B. This increase in variance would probably be relatively small for this variable. We expect Var(X) to dominate Var(Y) which should dominate $\mathrm{Var}_B(Y)$.

However, the problem is more complicated in that we really have



where E1, E2 ... indicate the fields into which the Other Income can be redistributed. Recalling Figures 1 and 2, these are fields such as "Other Interest and "Interest on U.S. Government Obligations." The real variables of most interest are then

$$\begin{gathered} E1+\Delta E1 \\ E2+\Delta E2 \\ \vdots \\ E12+\Delta E12. \end{gathered}$$

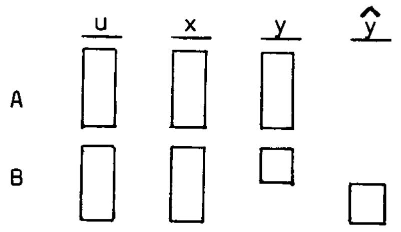
It is an open question whether the relative increase in variance for these variables would be significant.

III. IMPUTATION

It is not practical in our situation to calculate estimates using the weighting technique associated with double sampling. Each record is quite lengthy and it would be complicated to allow a different weight for each item on the record. In fact, at least one of our users has vetoed the idea of having more than one weight per record. For this reason, the missing values for the units that are not subsampled for editing will be imputed and estimates will be calculated from all

$$n' = n'_A + n'_B$$

records. That is, our data will be of the form



As with our model for stratifying the records into group A or B, our imputation procedure is only an initial attempt based primarily on subject matter expertise; we expect to refine and improve this procedure subsequent to the analysis of the first year's data.

The initial plan is to use hot deck imputation within adjustment cells. A record with schedules to be imputed will be matched to a record in group B with these same schedules edited. The record with the schedules edited is called the donor, because the amounts to be imputed are calculated from this edited record. We are not using the traditional hot deck procedure; instead of "hot decking" the amount of change, we are using the percent change.

The adjustment cells were subjectively chosen so that they should define groups of records that are relatively homogeneous with respect to the variable of interest, namely the relative change made due to the schedules. The adjustment cells are defined in terms of three characteristics of the record:

(1) Pattern of schedules needing imputation;

(2) Industrial classification; and,

(3) Total Assets and Net Income size [11].

In order to ensure that a cell has enough donors, we have provided for a relatively simple strategy of collapsing cells.

There are many important statistical issues regarding such a procedure (such as cell size, the effect on variance and variance estimation) which need to be taken into account [8]. For the most part, however, the cell definitions and the hierarchical structure are based primarily on subject matter opinion and expertise. A brief summary of the cell sizes and the cell collapsing for this year is included in the next section.

IV. PRELIMINARY RESULTS

Once processing began, we looked at a sample of about 3,000 records to see how well our prediction procedures worked for Other Income.

Figure 3 shows the plot of the change made to Other Income y, versus the original amount x. The stratification into groups A and B is also shown. It appears that stratification is successfully catching records with large (absolute) changes to Other Income, because it is putting records with large original Other Income into group A. However, so far we have been unsuccessful in predicting which records will not be changed (y=0)--ideally, we would like all those records to be in group B.

Looking only at donors (Group B), we consider the adjustment cells. Figure 4 shows the plot of y vs x, rescaled, for all donors. Two basic subdivisions are shown: financial and nonfinancial records. This is the broadest possible subdivision. However, we can see that being in a financial industrial class is a good predictor that Other Income will be changed. Although the nonfinancial records generally have no changes made, there is an indication again of the two distinct populations, "change" and "no change."

The hierarchy of the cells and the collapsing strategy are defined so that, at its worst, the adjustment cells are defined by the pattern of schedules to be imputed and by whether they are classified as financial or nonfinancial. That is, we will never combine records across these variables.

After the break into financial or nonfinancial classes, the next level of the hierarchy separates records according to fairly broad industrial classes. The records are further classified according to the size of the corporation, in terms of assets and net income: tiny, small or medium size. Recall that the largest corporations were not subject to subsampling and so should not need imputation. For two major industrial classes within the nonfinancial sector, there was one more level of detail; the size classes were subdivided according to (two) minor industrial classifications. (See Figure 5.)

The quality of our estimation depends on how much collapsing takes place. This year we had 36,586 records with at least one schedule to impute, and 3,989 donors. There were 7,912 financial records to be imputed and 28,674 nonfinancial. Tables 1 and 2 summarize the results of the collapsing for this year, across all 15 patterns. Since there were significantly more nonfinancial records than financial, there was a severe problem with collapsing on the financial side. (We have increased the subsampling rates for financial records so that next year we will not have this problem.)

For the nonfinancial records (Table 1), we never collapsed across the major industrial classifications, and in fact we never combined all the size classes--i.e., we always had some size distinction. We had many cells that were not combined at all, but maintained the maximum detail possible.

In contrast, for financial records (Table 2), the size variable was often lost by combining all cells, and major industries were sometimes combined. In fact for one pattern, the financial records collapsed as far as possible. That is, all financial records were combined into the same cell; this cell contained 505 records to impute.

The tables also show the maximum and minimum of two variables:

(1) The number of donors in a cell; and,

(2) The ratio of donors to imputes.

For example, in Table 1, 15.9% of the nonfinancial records to be imputed were in an adjustment cell containing only records for small corporations classified as Trade. The number of donors in such cells ranged from 19 to 88 and the ratio of donors to imputes ranged between .09 and .25. As one can see, in a few cases in Table 2, there is only one donor in a cell. This was allowed in order to improve the adjustment cells by minimizing the collapsing. However, this will effect our estimation of variance.

V. CONCLUSIONS AND FUTURE PLANS

The problem of nonresponse and the use of a static hot deck procedure for imputation raise several issues and difficulties:

(1) Extent of potential bias;

(2) Variance of the estimate;

(3) Estimating the variance; and,

(4) Preserving relationships between reported and imputed values.

These have been discussed in various places in the literature [e.g., 4, 5, 7-9]. Imputation of the missing information will increase the variance of the estimator. Using the standard estimate of variance does not take into account the component of variance due to imputation, and may result in a gross underestimation of the variance [9]. We plan on estimating this additional component of variance using multiple imputation [3] (and, hence, the missing data have already been imputed twice, using two different random starts within each adjustment cell).

Our initial attempt to minimize the distortion of the relationships between reported and imputed items is limited to imputing all the missing information on a record from one donor record, and the hierarchy of adjustment cells. This year's effort was based primarily on our subject matter opinion and expertise. Our subsequent analysis of the results will undoubtedly lead to some changes and improvements to the system. In particular, we will model an indicator variable (change versus no change) and next year we hope to include this variable in our definition of the adjustment cells.

In conclusion, we are moving toward more computer-assisted data capture. The computer has some obvious advantages - it is fast and relatively cheap - but it certainly cannot take over all our decisions in reviewing and correcting data items. However, we believe that we have identified a part of our population, a part of our problem, where the computer can do almost as well as an editor. We hope to reduce our cost with relatively little loss in precision. Finally, by reducing our cost in one such area, we may be able to afford to put more emphasis, more resources, on other critical areas, such as on the largest corporations, which dominate the estimates. Employing an imputation strategy for the smaller corporations allows us to control the nonsampling error in this portion of the sample, at the same time freeing up resources to reduce the nonsampling error elsewhere.

ACKNOWLEDGEMENTS

The author would like to thank David Barker and David Reboussin for their contributions and technical support on this project, Wendy Alvey for her help in developing the ASA presentation and this paper, and Cindy Thornley for typing this paper.

NOTES AND REFERENCES

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

[2] Cys, K., S. Hinkins and V. Rehula (1982), Automatic and Manual Edits for Corporation Income Tax Returns, American Statistical Association, Proceedings of the Section on Survey Research Methods.

[3] Herzog, T.N., and C. Lancaster (1980), Multiple Imputation Modeling for Individual Social Security Benefit Amounts, American Statistical Association, Proceedings of the Section on Survey Research Methods.

[4] Kalton, G. (1983), Compensating for Missing Survey Data, Research Report Series, Institute for Social Research, The University of Michigan.

[5] Little, R.J. (1982), Models for Nonresponse in Sample Surveys, Journal of the American Statistical Association, Vol. 77, pp. 237-250.

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

[7] Oh, H.L., and F.J. Scheuren, Weighting Adjustments for Unit Nonresponse, Incomplete Data: The Theory of Current Practice, National Academy of Sciences, Panel on Incomplete Data (In Press).

[8] Incomplete Data: The Theory of Current Practice, National Academemy of Sciences, Panel on Incomplete Data (In Press).

[9] Rubin, D.B. (1981), Handling Nonresponse in Sample Surveys by Multiple Imputations, Monograph for the Census Bureau.

[10] Barker, D., S. Hinkins and V. Rehula, et al., 1981 Corporation Validation Tests, Statistics of Income Division, Internal Revenue Service (Unpublished).

[11] The size of the corporation (in terms of assets and income) is also one of the variables defining our sample strata. The imputation adjustment cells do not coincide with the sampling strata, so the units in an adjustment cell may have different weights, different probabilities of selection. However in our application, the data are missing at random, because we control the "missingness" process, so the weights will not be related to the pattern of missing data and, hence, are ignorable [9].

FIGURE 5.--HIERARCHY OF ADJUSTMENT CELLS.

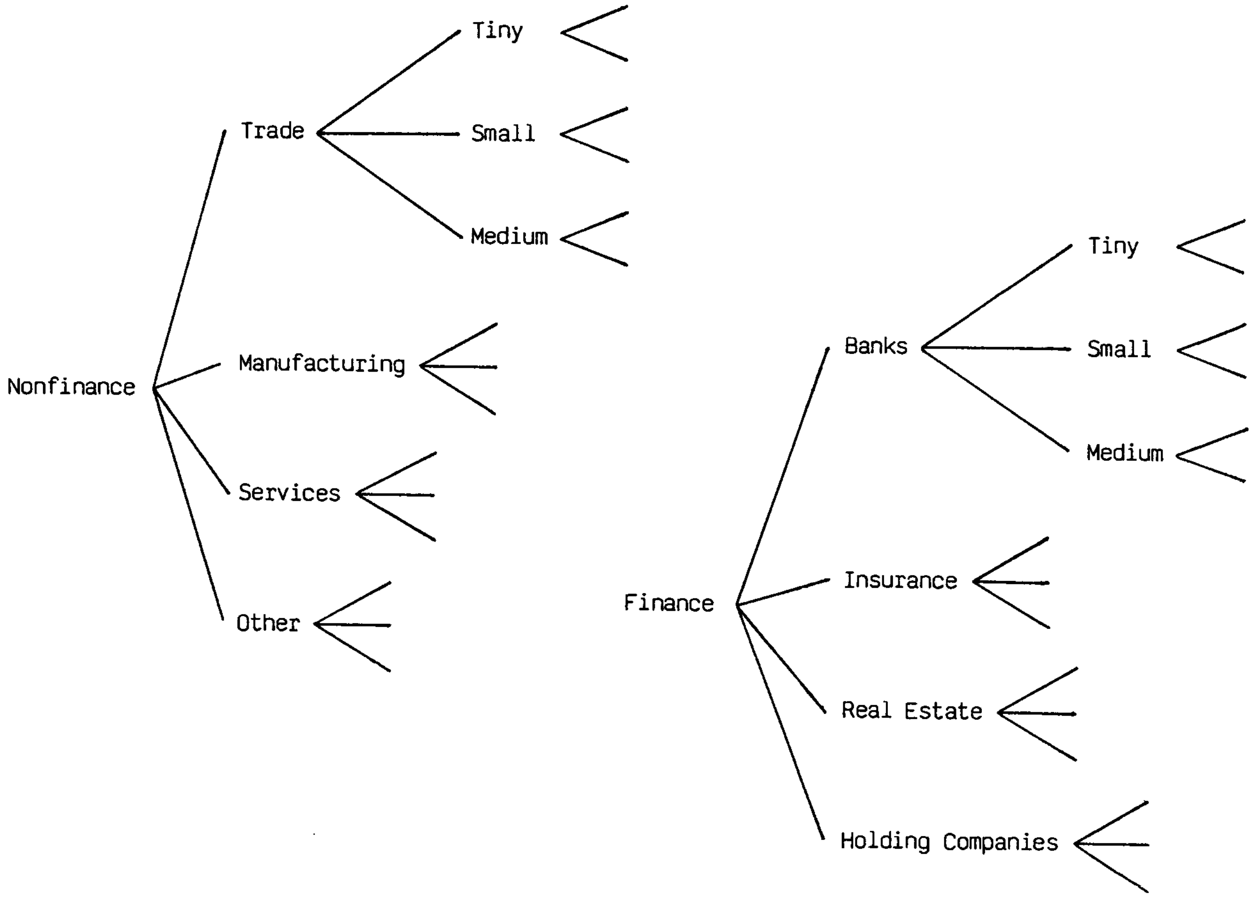


Table 1.--CELL COLLAPSING FOR NONFINANCIAL RECORDS

(28,674 records)

Cells	Percent of imputes at this level	Number of donors		Ratio $\left(\frac{\text{donors}}{\text{imputes}}\right)$	
		max	min	max	min
NONFINANCE........	0	--	--	--	--
TRADE.............	0	--	--	--	--
partial collapse.	4.2	47	24	.14	.08
tiny............	11.3	71	6	.16	.07
small...........	15.9	88	19	.25	.09
medium..........	7.3	41	4	.19	.07
MANUFACTURING.....	0	--	--	--	--
partial collapse.	2.0	18	9	.12	.08
tiny............	3.0	18	4	.25	.06
small...........	6.9	41	6	.19	.07
medium..........	9.0	58	4	.15	.07
SERVICES..........	0	--	--	--	--
partial collapse.	0.9	17	10	.13	.07
tiny............	8.3	39	4	.18	.07
small...........	4.5	27	4	.15	.07
medium..........	2.2	10	3	.18	.06
OTHER.............	0	--	--	--	--
partial collapse.	0.5	10	--	.07	--
tiny............	7.1	36	4	.16	.06
small...........	11.0	55	9	.15	.07
medium..........	6.0	24	6	.25	.05

NOTE: A total of 61 records were imputed in cells with just one donor.

Table 2.--CELL COLLAPSING FOR FINANCIAL RECORDS

(7,912 records)

Cells	Percent of imputes at this level	Number of donors		Ratio $\left(\frac{\text{donors}}{\text{imputes}}\right)$	
		max	min	max	min
FINANCE...........	6.4	51	--	.10	--
Combined 3 industries.......	3.3	27	--	.10	--
Combined 2 industries.......	4.1	11	6	.10	.05
BANK..............	1.0	9	1	.15	.06
partial collapse.	1.2	5	2	.13	.07
tiny............	1.7	4	2	1.00	.08
small...........	2.7	18	2	.40	.08
medium..........	33.0	68	1	.16	.05
INSURANCE.........	0.2	1	--	.07	--
partial collapse.	1.4	4	1	.17	.05
tiny............	1.8	5	1	.40	.06
small...........	0.6	6	1	.21	.11
medium..........	0.3	2	1	.50	.11
REAL ESTATE.......	0	--	--	--	--
partial collapse.	1.6	15	--	.12	--
tiny............	12.6	27	1	.19	.09
small...........	13.7	19	2	.18	.07
medium..........	5.1	10	3	.20	.06
HOLDING COMPANY...	3.6	12	1	.14	.06
partial collapse.	2.2	4	1	.11	.05
tiny............	0.5	2	1	.20	.11
small...........	0.8	7	4	.23	.13
medium..........	2.3	11	3	.33	.12

NOTE: A total of 212 records were imputed in cells with just one donor.