Imputation of Balance Sheets for the 1992 SOI Corporate Program

Bertrand Überall, Internal Revenue Service

For each tax year, a sample of U.S. corporate tax returns is selected by the Internal Revenue Service's Statistics of Income Division (SOI), from which income and financial data are compiled. As in most samples, the issues of item nonresponse and data consistency and quality arise.

This paper examines the problem of imputing data items missing from the balance sheets of sampled corporate tax returns. A corporation's balance sheet consists of two portions. The first portion contains the asset items, including the Total Assets field. The second portion contains the liability items, including the Total Liabilities field, as well as the components of Shareholder's Equity. For the balance sheet to be considered complete, Total Assets must be equal to Total Liabilities plus Shareholder's Equity. Also, the sum of the asset items (with some items subtracted) must be equal to Total Assets, and the sum of the items in the second portion (with some items subtracted) must be equal to Total Liabilities plus Shareholder's Equity. For simplicity's sake, and to be consistent with the terminology used for SOI processing, we will refer from this point on to all the fields in the second portion as "liability items," and we will refer to the total of the fields in the second portion as "Total Liabilities." A list of the balance sheet items is provided in Figure 1.

Figure 1.--Balance Sheet items

Asset items:

(1) Cash
(2) Trade notes and accounts receivable
(3) (-) Bad debt allowance
(4) Beginning inventories
(5) Inventories
(6) U.S. Government obligations
(7) Tax-exempt securities
(8) Other current assets
(9) Loans to stockholders
(10) Mortgages and real estate loans
(11) Other investments
(12) Depreciable assets
(13) (-) Accumulated depreciation
(14) Depletable assets
(15) (-) Accumulated depletion
(16) Land
(17) Intangible assets
(18) (-) Accumulated amortization
(19) Other assets
(20) Beginning total assets
(21) Total assets

Liabilitiy items (including shareholder's equity):

(1) Accounts payable
(2) Mortgages, etc. less than one year
(3) Beginning P-insurance-liabilities
(4) P-insurance-liabilities
(5) Other current liabilities
(6) Loans from stockholders
(7) Mortgages, etc. one year or more
(8) Other liabilities
(9) Common stock
(10) Preferred stock
(11) Capital stock
(12) (+/-) Paid-in capital surplus
(13) Beginning retained earnings appropriated
(14) Retained earnings appropriated
(15) Beginning retained earnings unappropriated
(16) (+/-) Retained earnings unappropriated
(17) (-) Cost of treasury stock
(18) Total liabilities (plus shareholder's equity)

Since corporations are not required to provide a complete balance sheet, and since none of the balance sheet items have a tax consequence, it is possible for a corporation to have items missing from its balance sheet. Because assets are used for stratification for the SOI corporate program, it is necessary to impute the missing items.

For all years up to 1991, corporation data aggregated at the major industry level from the previous year were used as a basis for imputation. These aggregate amounts are included in the *Corporation Source Book* published annually by SOI. For simplicity, we will refer to these data from this point on as *"Source Book"* data.

Beginning with the 1992 program, data from a corporation's return in the previous year are used as a basis, if this prior year return is available and valid, in the hope of obtaining more accurate estimates for the absent items. After describing the extent of the missing data problem in 1992 and the methodology used to impute these data, this paper describes a simulation designed to evaluate the effect of two different imputation methods on estimates of balance sheet items missing from a company's tax return. This simulation is carried out on returns that did not initially require imputation, so that actual reported data are available for comparative purposes. The first method uses the *Source Book* totals (for the company's major industry) in year T-2 as a basis to impute items missing from its balance sheet in year T (the current year). The second method is a cold-deck imputation method, that uses prior year information from the corporation's return in year T-1, which resembles the current year's balance sheet more closely. If this prior year return is not valid or is nonexistent, *Source Book* totals from year T-2 are used as a basis for imputation. Examples of prior year records that would not be considered valid as a basis for imputation include incomplete records, imputed records, added records, rejected records, and duplicate records. For the 1992 program, it is necessary to revert to data from two years prior when using *Source Book* totals as a basis for imputation, because 1992 is also the first year that the imputation of corporate balance sheets at SOI is performed on-line during editing (and no longer in batch mode), and the complete *Source Book* for year T-1 is not yet available when editing for the current year begins.

■ Extent of the Missing Balance Sheet Data Problem for the 1992 Corporate Program

The number of corporate returns with incomplete balance sheets usually only represents a small fraction of the total number of returns. For the 1992 program, only 221 returns have at least one imputed balance sheet item, from a total of 82,646 returns sampled (these totals include subsidiaries of parent corporations). Of these, 77 are imputed using data from the corresponding 1991 return as a basis, including six with missing Total Assets, and 144 using data from the 1990 SOI *Corporation Source Book*, including 30 with missing Total Assets.

By far the largest number of imputations (98) are in the Insurance industry (which must have Total Assets present), especially Form 1120-L Life Insurance returns (84 for Life Insurance and stock companies, almost half of which are prior year imputations, and five for Mutual Life Insurance companies). The reason for this large number is that Form 1120-L records do not have a balance sheet schedule. Instead, they are required to provide certain amounts on Schedule K, as well as attach an annual statement which may be incomplete, or become separated from the return during processing.

The only other industries with a significant amount of imputed balance sheets are Banking, with 22 imputed returns, all of which have Total Assets present, and 18 of which are *Source Book* imputations; and Other Services, with 16 imputed returns, seven of which have missing Assets and 13 of which are *Source Book* imputations.

■ Description of the Imputation Method

For both imputation methods described in the introduction, the different steps are as follows.

We first decide which balance sheet items (including Total Assets) need to be imputed. We also determine which items are excluded from the imputation process because they are not present on the particular 1120 series Tax Form being considered. If Total Assets (or, equivalently, Total Liabilities) are reported, missing balance sheet items can be imputed using prior year or *Source Book* amounts. If Total Assets are not reported, this item must be imputed first, using the return's receipts or deduction amounts. Then, once Total Assets are obtained, the other missing balance sheet items can be imputed. Note that Total Assets must be reported for insurance companies.

The following notation is used:

X_i: asset items (excluding Total Assets)
Y_j: liability items (excluding Total Liabilities)
TA: prior year or Source Book total for Total Assets
A_i: prior year or Source Book total for item X_i
L_j: prior year or Source Book total for item Y_j
TR: prior year or Source Book total for Total Receipts
GR: prior year or Source Book total for Business Receipts.

The indices i and j can be included in any of three disjoint subsets:

I = set of items to be imputed
E = set of items not reported and excluded from imputation because of Form type
R = set of reported items.

Finally, we define the sum of reported assets (SRA) and the sum of reported liabilities (SRL) by:

$$SRA = \sum_R X_i \text{ and } SRL = \sum_R Y_j .$$

Imputing Total Assets

When imputing Total Assets, two cases may occur.

Neither the sum of reported assets nor the sum of reported liabilities is positive.

This will be the case if no balance sheet items are reported, or if the only reported items are items that are subtracted when computing the sums of asset and liability items.

For a financial corporation:

Case 1: If TR $\neq$ 0 and Total Receipts > 0 set:

Total Assets = Total Receipts (TA/TR).

Case 2: If TR $\neq$ 0 and Total Receipts = 0:

- The balance sheet is suppressed if Total Deductions < 50,000
- If 50,000 $\leq$ Total Deductions < 1,000,000 set:
  Total Assets = 1,000 (TA/TR).
- If 1,000,000 $\leq$ Total Deductions set:
  Total Assets = Total Deductions (TA/TR).

For a nonfinancial corporation:

Case 1: If GR $\neq$ 0 and Gross Receipts > 0 set:

Total Assets = Gross Receipts (TA/GR).

Case 2: If GR $\neq$ 0, Gross Receipts = 0 and Total Receipts > 0 set:

Total Assets = Total Receipts (TA/GR).

Case 3: If GR $\neq$ 0 and Gross Receipts = Total Receipts = 0:

- The balance sheet is suppressed if Total Deductions < 50,000
- If 50,000 $\leq$ Total Deductions < 1,000,000 set:
  Total Assets = 1,000 (TA/GR).
- If 1,000,000 $\leq$ Total Deductions set:

Total Assets = Total Deductions (TA/GR).

Either the sum of reported assets or the sum of reported liabilities is positive.

This will be the case if at least one positive balance sheet item is reported.

Assuming the denominators are $\neq 0$, we set Total Assets in the following way:

Case 1: If SRA > SRL set:

$$\text{Total Assets} = (\text{SRA}).\ (\sum_{I} A_i + \sum_{R} A_i) / (\sum_{R} A_i);$$

Case 2: If SRA < SRL set:

$$\text{Total Assets} = (\text{SRL}).\ (\sum_{I} L_j + \sum_{R} L_j) / (\sum_{R} L_j).$$

Imputing Asset and Liability Items

Assuming the denominators are $\neq 0$, the imputed asset and liability items are given by:

$$X_i = (\text{Total Assets - SRA}).\ A_i / (\sum_{I} A_i) \text{ and}$$

$$Y_j = (\text{Total Assets - SRL}).\ L_j / (\sum_{I} L_j).$$

■ Description of the Simulation

Our goal is to analyze and compare the effects of the two imputation methods described in the introduction.

To do this, we simulate how both imputation methods perform when applied to a selected group of returns in the 1992 SOI Corporation file that did not require imputation and for which balance sheet fields are set to zero for the purposes of the simulation. Three cases can be considered: no balance sheet items are reported, Total Assets is the only item reported, and Total Assets are not reported but other items are. For this paper, we consider the situation where Total Assets are reported and all other items are missing. For the two other cases, further research will need to be conducted.

The group of returns selected consists of Life Insurance returns which have assets between $50,000,000 and $250,000,000. This group was selected because it seemed representative of the imputation process: 6.6% of Form1120-L returns are imputed, and this is the industry with the most number of imputed returns. There were 103 returns in this class to begin with, and 78 returns were used for the simulation, since imputed returns, returns with invalid prior year records, and consolidated returns needed to be excluded. All fields were set to zero except Total Assets, and these 78 returns were imputed using both methods.

■ Results of the Simulation

For both imputation methods, the imputed balance sheet amounts are compared to the true reported amounts for the selected group of returns. For each balance sheet item, the relative error is measured and the results for the two methods are compared. Box plots are graphed for both methods as a visual comparison.

For all 15 balance sheet items that are significant for Form 1120-L returns, we observe that the relative errors are significantly lower and more centered around the median for prior year imputation than for *Source Book* imputation. For most fields, the errors tend to be positive, indicating that, in general, imputed amounts are somewhat higher than reported amounts. Large outliers are observed, especially for *Source Book* imputation. Also, some errors cannot be computed. This occurs if the item is, in fact, reported as zero, but imputed anyway. This is indicative of a limitation of the imputation method due to the nature of our data: the editors cannot distinguish between an item that is missing due to ommission by the taxpayer or an item that is legitimately not present on the tax form because it is, in fact, zero. By default, this amount is imputed anyway.

We examine four balance sheet items in particular: Capital Stock, U. S. Government Obligations, Other Assets, and Retained Earnings (unappropriated). Figures 2 (and 3) list the medians and quartiles of the relative errors for these four fields when prior year and *Source Book* imputation is used, respectively.

Figure 2.--Results from Prior Year Imputation

Item	Median	Q1	Q3
Capital Stock	0.090	-0.100	0.186
Government Obligations	-0.095	-0.351	0.146
Other Assets	0.096	-0.239	0.768
Retained Earnings (unappropriated)	0.005	-0.199	0.226

Figure 3.--Results from *Source Book* Imputation

Item	Median	Q1	Q3
Capital Stock	1.374	0.466	2.755
Government Obligations	0.296	-0.442	1.532
Other Assets	3.250	0.734	9.080
Retained Earnings (unappropriated)	-0.580	-1.378	0.210

The absolute value of the median for these four items is much larger for *Source Book* imputation than for prior year imputation. The quartiles also demonstrate how much larger the errors are for the *Source Book* method. This is especially apparent for Other Assets.

The box plots for these four balance sheet items are given in Figure 4 through Figure 7. Large outliers have been reset to amounts larger than Q3 to ensure that plots are on the same scale. The difference between the two methods can easily be observed. In all cases, the interquartile range is much smaller for the prior year imputation method. For Retained Earnings, the balance sheet amount is generally underestimated. It is usually overestimated for the other three items.

Figure 4.--Relative Errors for Capital Stock

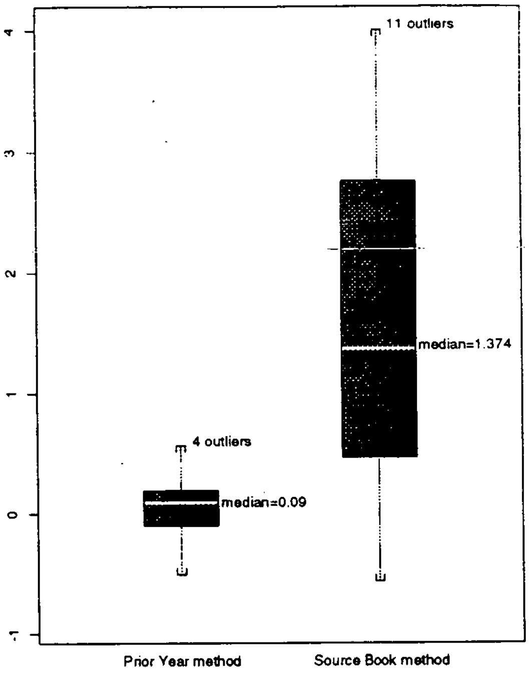


Figure 5.--Relative Errors for Government Obligations

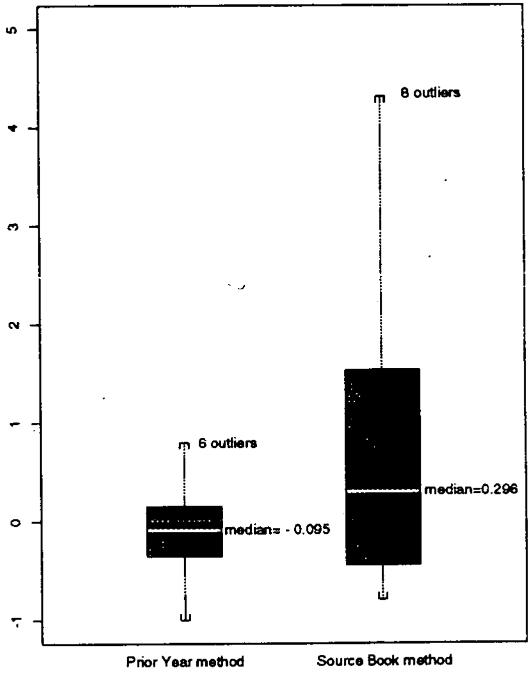


Figure 6.--Relative Errors for Other Assets

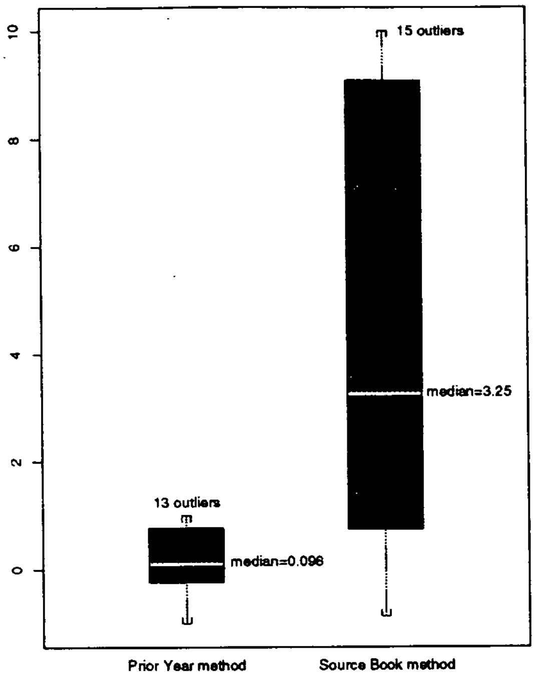
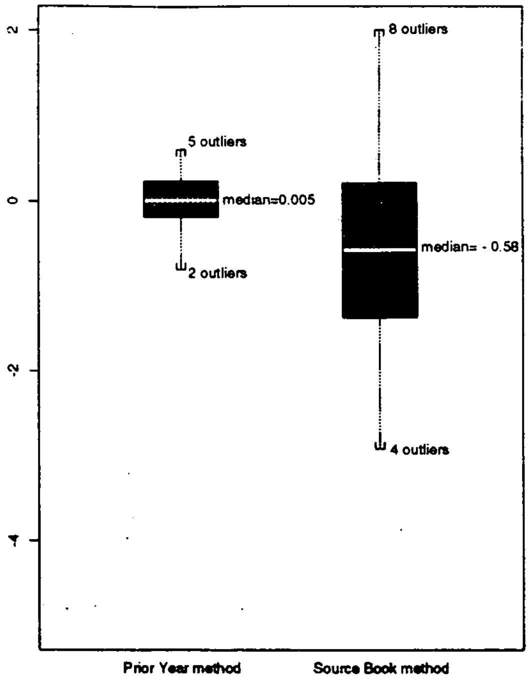


Figure 7.--Relative Errors for Retained Earnings (Unappropriated)



■ Conclusion

Our results for Life Insurance companies seem to support our intuition: by using prior year results from a specific return's balance sheet to impute the current year's balance sheet, we are able to estimate the missing items much better than if we simply use aggregate data from previous years for the entire industry to which the return belongs. As a next step, it would be interesting to compare aggregate estimates for an entire industry (Life Insurance or other) for both methods, as well as try to mimic the reporting behavior of a given industry by assuming the incomplete returns are randomly distributed and appear in proportions similar to those actually observed during processing.

For the future, it will be interesting to see if the *Source Book* imputation method fares better for the 1993 program, Beginning in 1993, *Source Book* aggregation is performed at the minor industry level, which is a much finer aggregation than major industry. It also will be interesting to examine the results of a simulation for which all items, including Total Assets, are missing, and another for which some items other than Total Assets are present. The imputation of the Total Assets field stands apart from that of the other fields, and it seems necessary to study how well this imputation is carried out, because this field plays a crucial role in the SOI corporate sampling program.

■ Acknowledgments

The author would like to thank the many members of the Corporation Branch and the Mathematical Statistics Team for their invaluable contributions. A special thanks is extended to Kathleen Walker, who programmed the imputation and helped design the simulation study.

■ Bibliography

Hinkins, Susan (1982). Imputation of Missing Items on Corporate Balance Sheets, *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Internal Revenue Service (1994). *Statistics of Income -- 1991 Corporation Income Tax Returns* (Publication 16).

Internal Revenue Service (1994). *Statistics of Income -- 1992 Corporation Income Tax Returns Documentation Guide* (Document 6930).

Internal Revenue Service (1993). *1990 Corporation Source Book* (Publication 1053). ■