DISCLOSURE REVIEW AND ITS IMPLICATIONS FOR THE 1992 SURVEY OF CONSUMER FINANCES

Gerhard Fries, Federal Reserve Board, R. Louise Woodburn and Barry Johnson, Internal Revenue Service
Gerhard Fries, FRB, Mail Stop 180 Washington, DC 20551, mlgxfOO@frbgov

Key Words: Confidentiality, Imputation

A principal concern among survey practitioners is protecting the confidentiality of the survey respondent This is important, not only for the direct consequence of keeping an individual's data anomynous, but also for the more global perception that it is 'safe' to participate in surveys. On the other hand, it is important to provide as much useful data as possible to policyrnakers and researchers. Steps taken to protect a respondent's identity often compromise the usefulness of the data. Thus, it is important to keep the integrity of the data intact. That is, inferences from the public data should be no stronger and not significantly weaker than those using the internal data. This paper is based on our experiences with the Federal Reserve Board's Survey of Consumer Finances (SCF), a triennial household survey that includes data on finances, employment and demographics. A major objective of the 1992 SCF was to release geographic region data that had been omitted from the 1989 SCF public-use data set This omission prompted numerous requests for geographical information from SCF data users. In this paper, we further the initial research in Fries and Woodburn [1994]. We detail the disclosure procedures used that allowed us to release the nine Census regions and we examine the effects that these procedures have on analyses using the public data. Including this introduction, there are five sections. In the next section, we provide a brief summary of the SCF, covering the sample design, data collected, and disclosure issues. In the third section, we detail the disclosure strategy currently used in the SCF. The effects of the disclosure adjustments on selected estimates are presented next. We summarize our results and discuss their implications for future surveys in the last section.

The Survey of Consumer Finances

The SCF is a triennial household survey sponsored by the Federal Reserve Board with cooperation from the Statistics of Income Division (SOI) of the Internal Revenue Service (see 1992 SCF Codebook for details). Data are collected on household finances, income, assets, debts, employment demographics, and businesses. The interview averages about 75 minutes, but interviews of households with more complicated finances sometimes last several hours. An important objective of the SCF effort is to collect representative data to measure wealth. In order to accomplish this, the sample is selected from a dual frame that is composed of an area probability frame and a list frame (see Kennickell, A. B. and McManus, D. A., [1993] for details on the strengths and limitations of the sample design). The list frame is based on administrative records maintained by SOI. The list sample is stratified on an estimated wealth index with higher indexes selected at a higher sampling rate.

Due to the sensitive nature of the financial questions, both unit and item nonresponse are concerns in the SCF. The complex sample design and the use of frame information for estimation helps to address the unit nonresponse concern. Sampling weights are computed that account for differential nonresponse in the list sample. The final weights are constrained by control totals computed using frame data driven by the SCF data (Kennickell, McManus, and Woodburn [1996]).

Data also drive the process to account for item nonresponse, with missing values multiply-imputed using a Gibbs sampling approach (Kennickell[1992]). For the SCF, the respondent has three options for a given question, he can: 1) give a specific value, 2) decline to answer (refuse or reply "don't know"), or 3) choose an interval from a range card provided by the interviewer. In the imputation procedure, both (2) and (3) are imputed. The imputations for the range card responses are constrained by the range interval boundaries; in fact, all amputations are drawn from truncated distributions. The Gibbs sampling approach involves iteratively estimating a sequence of large randomized regression models to predict the missing values based on variables that are available for a given respondent. The result is an imputed data set that preserves the distributions and relations found in the nonimputed data. A shadow variable is included that indicates the status of the original data, such as, whether or not the value is imputed, and what the range card interval was, if appropriate. The imputation machinery is used in the disclosure avoidance preparation of the public use file as described below.

In order to estimate the total error in the estimates, both sampling and imputation error are included. Estimates of the variance due to sampling are computed using the bootstrap method with 999 bootstrap replicates. (A thorough reference for the Bootstrap method is Shao and Tu [1995]). Estimates of the variance due to imputation are computed using five imputation implicates. (Rubin [1987] develops multiple imputation for the purpose of enabling the user to measure the error due to imputation).

SCF Data Release Strategy

The release of microdata from the SCF is complicated both by the nature of the sample design and also by the type of data collected. Due to the use of the SOI administrative data in the sample design, disclosure review of the data must satisfy the same conditions that guide SOI data release. Additionally, several processing factors must be considered: construction of sampling weights, imputation for item nonresponse, and development of variance computation tools. For the 1992 SCF, as with the 1989 SCF, preliminary data were released prior to the final data release. The disclosure procedures used for the data releases for the 1989 SCF are detailed in Fries and Woodburn [1994]. For the 1992 SCF, the review process for continuous and cardinal variables was improved with the use of graphical tools (Fries and Woodburn [1995]). For the continuous variables, plots that show influence in the cumulative distribution, either in weight or weighted value, are reviewed. Also, for 1992, some demographic data were swapped for selected cases. The data release strategy for the 1992 SCF was similar to that for the 1989 SCF with the first data release being somewhat limited in detail. For the 1992 SCF, there were only two data releases -- the preliminary data release in 11/94 and the final data release in 4/96.

There were two main objectives for the preliminary data release. First, it was important to release as many variables with as much detail as possible. Second, it was necessary to limit the amount of detail both to satisfy disclosure concerns, and also because the data, as well as the sampling weights, were preliminary. In order to address both of these objectives, all continuous variables were top and bottom coded as shown in Table 1. This made it possible to release all variables with the exception of investment real estate which was set to missing. Also, selected variables for specific cases were imputed or set to missing. Most cardinal values, such as year of birth were rounded to the nearest five and top/bottom coded as necessary. For discrete variables the strategy included omitting completely, collapsing, or assigning a value of "other". Industry and occupation codes were collapsed to the 1-digit level. The variables completely omitted from this release included geography, make and model of car, and sampling weight components. It was decided that of these previously omitted variables, only geographic region would be in the final public release.

The disclosure review of the final release of the 1992 SCF took into account not only the data that were previously released, but also the desire to release even more data with less interference in the form of excessive rounding, suppression and top coding. It was important to avoid creating a data set where all data for the wealthiest respondents appear to be imputed. This final criteria was due both to user and to internal concerns. Both concerns stem from the fact that the wealthy respondents in the SCF account for a large portion of the estimates of certain skewed variables. Since users typically presume that respondent data have more integrity than imputed data, it was desired to keep as much respondent data intact as possible. (This is in contrast to the opinion that no respondent data be released publicly, that is, all data are imputed - see Rubin [1993].) The internal concerns revolved around the imputation variance which increases as more data are imputed. (Although this is true, it turns out that the sampling variance dwarfs the imputation variance as shown in Kennickell, McManus and Woodburn [1996].)

TABLE 1 - Rounding Scheme for Continuous Variables - Preliminary Release

Data Range	Rounded to Nearest
x > 25 mill.	set = 25 mill.
I mill.<x<=25 mill.	100,000
100,000<=x<1 min	10,000
10,000<=x<100,000	1,000
1,000<=x<10,000	100
5<=x<1,000	10
0<=x<5	set= 1
-4<=x<0	set = to original value
-1,000<x<=-5	10
-10,000<x<= - 1,000	100
-100,000<x<= -10,000	1,000
-1 mm.<x<= -100,000	10,000
x <= -1 mill.	set= -1 mill.

For the final 1992 data release, it was important to release geography at the levels of the four and nine Census regions. Also, as a result of internal data requests at the FRB, it was decided to provide two additional industrial classifications. For this release, the severe rounding and top[bottom coding performed for the preliminary release were relaxed. The rounding strategy used is shown in Table 2. Rounding of cardinal variables was removed, although some variables were still top/bottom coded.

TABLE 2 - Rounding Scheme for Continuous Variables - FINAL Release

Data Range	Rounded to Nearest
x >= I mill	10,000
10,000<=x< 1 mill.	1,000
1,000<=x<10,000	100
5<=x<1,000	10
I<=x< 5	set= 1
-4<=x<1	set = to original value
-1,000<x<=-5	10
-10,000<x<=-1,000	100
-1 mill.<x<=-10,000	1,000
x <= -1 mm.	set = -I mill.

Examples include the number of businesses owned which was top coded at 25, and the model year of cars which was bottom coded at 1940. Some of the collapsing of the discrete variables was kept intact for this release. For example, for other asset types, rare books, antiques, oriental rugs and furniture were collapsed together.

For the continuous variables, the disclosure review for the final data release dictated that respondent data that were deemed unique be imputed using the imputation procedure subject to range value constraints. This allowed for data for all continuous variables to be included in the public release including investment real estate, which was completely omitted from the preliminary release.

Analysis of Disclosure Adjustments

Our evaluation to date of the disclosure adjustments concentrates on the integrity of the public data. Since geographic region is now included in the public release, it is important to evaluate the effects of the adjustments to estimates by region. Additionally, since an important aspect of the SCF is to be able to measure wealth, we reviewed wealth, asset and debt estimates derived from the internal data and the public data. There are four steps in the evaluation of the effects of the disclosure adjustments. First, we look at point estimates by the four and nine Census regions. Second, we look at how the point estimate changes through the various imputation steps. Third we review the effect on variance estimates. Finally, regression estimates are computed. All analyses are based on the 1992 SCF data.

First, point estimates of average and median wealth were reviewed by the four Census regions. In Table 3, the average and median wealth for the Northeast and South Census regions are shown. We reviewed the data for all of the Census regions with similar conclusions, but only a few are presented due to disclosure concerns. At this level, the estimates do not change significantly.

Table 3 - Average and Median Household Wealth

Estimate	--- Northeast ----		------ South ----	
('92 S)	Internal	Public	Internal	Public
Average	208.9	209.8	139.0	140.5
Median	63.7	63.4	34.0	34.0

We also reviewed estimates of average wealth for the nine Census regions released, broken down by different demographic variables. Specifically, we reviewed average wealth, average debt and average assets by income, education status, and age cohorts. Overall, the estimates did not change greatly. In order to compare the tables quickly, we graphed the table cells for the internal estimates vs the table cells for the public estimates. Figure 1 shows the results for the average wealth by income categories. The income categories are shown in Table 4 which includes the estimates for New England for the five income categories. The point circled in Figure 1 corresponds to the >= 125,000 income category in Table 4.

Table 4 - Average Wealth by Household Income New England Region

Household Income Category ('92 $)	Public	Internal
<35,000	65.8	66.1
35,000 < 50,000	120.4	117.1
50,000 < 75,000	265.7	267.4
75,000 < 125,000	472.5	468.2
>=125,000	2577.6	2484.8

It is useful to investigate how the estimates change through the different disclosure protection steps. For the data we reviewed, neither the rounding nor the imputation step had any measurable, effect on the estimate. The effects of the different disclosure steps on estimates of average net worth for households in the South Census region with income greater than 125,000 are shown in Table 5.

Table 5 - The Effects of Disclosure Adjustments

Disclosure Adjustments	% Change from Internal
--Rounding Only	-0.03%
--Rounding & Imputation Only	-0.04%
--Public Data (All Adjustments)	4.04%

It is comforting that estimates of averages do not differ greatly from the internal to the public data. It is important, however, also to consider variance estimates especially since part of the disclosure adjustment process involved replacing respondent data values with imputed data. In Table 6, the sampling, imputation, and overall variance are shown for the estimate of wealth for the 99.5 to 100 percentile of the wealth distribution. Both the imputation and the sampling variance increased in the public data, but neither increase was large.

Table 6 - Imputation, Sampling & Overall Variance for Total Net Worth - 99.5 to 100 Percentile

Estimate	Internal	Public
Total Net Worth	4,026.4	4,024.1
Total Variance	$(371.4)^2$	$(373.1)^2$
-Imputation Var.	$(191.0)^2$	$(192.9)^2$
-Sampling Var	$(318.6)^2$	$(319.3)^2$
-%due to Imp	26.5%	26.7%

The final step in the investigation involved the stability of the relationships between the variables on the internal and the public data. A simple least squares regression was computed using the logs of different types of assets and indicator variables as to whether or not an asset was present (a total of 22 independent variables) to predict the log of total income. The results of the two regressions are remarkably similar. Both models computed the same independent variables to be significant. Only the parameter estimate of the very insignificant *cash value of life insurance* variable changed (from -0.000032 to +0.00043, significance level of .98). The R-square value of the model using the internal data was .5509, for the public data it was .5504.

Conclusions and Future Plans

The disclosure strategy that has been developed for the SCF has both strengths and limitations. The blank and impute method for the continuous variables is straightforward to implement using the existing imputation software. However, the decisions on which values to blank, and for discrete variables, which values to collapse, require an intensive review of the data. The use of graphical tools for the 1992 SCF disclosure review improved the ability to review thoroughly the data in a flexible manner.

The preliminary results presented here indicate that the integrity of the SCF data has been preserved through the disclosure review adjustments. More extensive analysis should be performed to investigate the effects of the adjustments on inferences.

Acknowledgments

The authors would like to express their gratitude to all of the Survey of Consumer Finances staff for their support with the disclosure review implementation. A special thanks to Arthur Kennickell for his guidance and comments

Bibliography

FRIES, G. and WOODBURN, R. L., (1994), 'The Challenges of Preparing Sensitive Data for Public Release," *Proceedings of the Section of Survey Research Methods, ASA.*

FRIES, G. and WOODBURN, R. L., (1995), 'Using Graphical Analyses to Improve all Aspects of the Survey of Consumer Finances," *Proceedings of the Section of Survey Research Methods, ASA.*

INTERNAL REVENUE SERVICE (1990), *Individual Income Tax Returns 1987,* Department of the Treasury, pp. 13-17.

HEERINGA, S., CONNOR, J. and WOODBURN, R. L., (1994), 'The 1989 Survey of Consumer Finances, Sample Design Documentation," Working Paper, ISR, University of Michigan.

JABINE, T. B. (I 993), "Statistical Disclosure Limitations Practices of United States Statistical Agencies,"Journalof *Official Statistics,* Vol, 9. No. 2, pp. 427-454.

Journal of Official Statistics (1993), *Confidentiality and Data Access,* Vol. 9. No. 2.

KENNICKELL, A.B. (1991), 'imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," *Proceedings of the Section of Survey Research Methods, ASA.*

KENNICKELL, A.B., and MCMANUS, D.A., (1993), "Sampling for Household Financial Characteristics Using Frame Information on Past Income,' *Proceedings of the Section of Surve yResearch Methods, ASA.*

KENNICKELL, A.B., and MCMANUS, D.A., and WOODBURN, R.L. (1996), 'Weighting Design for the 1992. Survey of Consumer Finances," *Federal Reserve Board Working Paper.*

Office of Management and Budget (1994), "Report on Statistical Disclosure Limitation Methodology,' Statistical Policy Working Paper 22.

RUBIN, D.B. (I 993), 'Discussion: Statistical Disclosure Limitation,' Journal *of Official Statistics,* Vol 9. No. 2, pp. 461-468.

RUBIN, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys,* John Wiley and Sons, Inc.

WILSON, O., and SMITH, W. J. Jr., (1983), 'Access to Tax Records for Statistical Purposes," *Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 591-601.

Figure 1 - Average Networth by Income Categories and Region: Internal Estimates vs. Public Estimates