DISCLOSURE REVIEW AND THE 1998 SURVEY OF CONSUMER FINANCES[1]

Gerhard Fries, Federal Reserve Board; Barry Johnson, Internal Revenue Service
Gerhard Fries, FRB, Mail Stop 153, Washington, DC 20551; gfries@frb.gov

Key Words: Confidentiality, Imputation

Protecting the confidentiality and privacy of survey respondents should be a major concern for all survey practitioners. Participation in surveys is critical, and every effort should be made to keep an individual's data anonymous. It should be considered 'safe' to be a respondent, even in today's world where public records from credit bureaus and real estate files, for example, are readily available for potentially unscrupulous use in trying to identify a respondent. It is also important, on the other hand, to provide as much useful data as possible to policy makers and researchers. Adjustments made to the data in order to protect a respondent's identity could severely restrict the usefulness of the data. Thus, it is imperative to take measures to keep the integrity of the data intact.

This paper is based on our experiences with the Survey of Consumer Finances (SCF), a triennial household survey that includes data on finances, employment and demographics. In this paper, we describe the disclosure procedures used in preparing the 1998 SCF data for public release. Including this introduction, there are five sections. In the following section, we provide a brief summary of the SCF, covering the sample design, data collected, and issues involving nonresponse and variance estimates. The next section details the disclosure strategy used for the 1998 SCF and the fourth section investigates the effects of the disclosure adjustments on particular analyses performed. We summarize our results and discuss their implications for future surveys in the last section.

Background on the SCF

The SCF is a triennial household survey sponsored by the Board of Governors of the Federal Reserve System with cooperation from the Statistics of Income Division (SOI) of the Internal Revenue Service. Data are collected on household finances, income, assets, debts, employment, demographics, and businesses. Interviews for the 1998 SCF were conducted via computer-assisted personal interviewing (CAPI) by the National Opinion Research Center at the University of Chicago (NORC) between June and December of 1998. The median length interview required about 75 minutes, although some complicated cases took substantially longer. A high percentage of interviews were obtained in-person, with some telephone interviews allowed for the convenience of respondents.

Data are collected on items that are not widely distributed (e.g. non-corporate businesses, or tax-exempt bonds). To provide adequate coverage of such variables and to provide good coverage of broadly distributed characteristics in the population (e.g. home ownership) the SCF combines two techniques for random sampling. The sample is selected from a dual frame that is composed of a standard, multistage area-probability (AP) sample and a list sample (see Kennickell and McManus [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 observations having higher index values selected at a higher sampling rate. The SOI data are made available for this purpose under strict confidentiality rules governing the use of those data as well as the data collected from the sample in the SCF interviews. The list sample is designed to oversample relatively wealthy families, but excludes people described by *Forbes* magazine as being among the 400 wealthiest people in the U.S.

Of the 4,309 completed interviews in the 1998 survey, 2,813 families came from the area-probability sample and 1,496 from the list sample. The response rate for the area-probability sample was about 66 percent. The overall response rate for the list sample was about 29 percent, and for the part of the list sample containing the wealthiest families the rate was only about 8 percent.

Both unit and item nonresponse are important issues for the SCF. Weighting adjustments compensate for nonrespondent households. The adjustments include post-stratification to known external control totals for age, location, and home ownership. For the list sample, frame data on financial income and the wealth index are also used (see Kennickell and Woodburn [1999]). Multiple imputation deals with missing data (see Kennickell [1998]).

Both sampling error and imputation error are measurable for the SCF. Estimates of the variance due to imputation are computed using five imputation replicates ("implicates"). Estimates of the variance due to sampling are computed using replication methods where samples are drawn from actual respondent records in such a way that the important dimensions of the original sample design are incorporated. These estimates can then be combined to yield standard errors for analysis (see Kennickell and Woodburn [1999]).

Disclosure Adjustments

The major goal of the disclosure review is to protect the identity of respondents while preserving data integrity (see Fries and Woodburn, [1994]). In light of this goal, some data items were not included in the final public release for the 1998 SCF. These included variables

relating to sample design and weight design, as well as, variables pertaining to the marital history of respondents and their spouses or partners. Most such variables are not related to the main purpose of the survey, but cause disclosure concerns. For example, variables giving information on the PSU (Primary Sampling Unit) or whether the observation was a list case could be quite damaging if revealed. Sophisticated data analysts could use such data to try to identify a respondent in the public file.

Overall, there were almost 2,600 variables to consider for disclosure adjustments including over 500 monetary variables. The review of the monetary variables initially involved graphical analysis, especially the use of scatter plots of variables by sampling strata indicators (see Fries and Woodburn [1994]). These were useful for identifying 'unique' responses, both overall and for population subgroups. For ordinal and discrete variables, frequency tables were used to check the number of responses in determined cell categories and to estimate disclosure potential.

For the discrete variables, small (less than 3 responses) or unusual cells were collapsed. These responses were combined with responses from related categories. Some variables that posed extreme disclosure concern, such as occupation and industry codes, were collapsed even further. Over 300 such variables were treated in this fashion and almost 200 were top or bottom coded. Most of these variables related to a date, number of years or number of items owned, etc. Other discrete variables, most pertaining to ages, were rounded as well.

A set of more than 350 observations containing very unusual responses and a random selection of cases was subjected to more stringent treatment. For these observations, all originally reported dollar values, as well as geography (4-level Census region and 9-level Census division) were simulated using the same multiple imputation technology developed for imputing missing data in the SCF (see Kennickell [1998]). For this purpose, the simulations for the dollar variables were constrained to lie in a neighborhood of the originally reported values. In addition, about 50 observations were chosen in which only geography was simulated.

A similar approach was used in the 1995 SCF: for the cases selected for special treatment, reported dollar values were simulated, originally missing dollar values were reimputed, and the same geographic variables were altered systematically (see Fries, Johnson, and Woodburn [1997]). For the geographic variables, records containing similar characteristics on key variables were used for swapping across cases. This data swapping technique is attractive in that it preserves univariate statistics for the overall data and for important subsets of the population. For the 1992 SCF, key variables, previously not imputed, that were regarded as being very unusual or extreme were simulated (see Fries, Johnson, and Woodburn [1996]). Data swapping for geography was also done, using a technique similar to the 1995 procedure.

Thus, going from 1992 to 1998, the major differences in the disclosure adjustments were: 1) for 1998, in the cases selected for special treatment, data simulation involved simulating only originally reported dollar values and not all dollar values or values deemed as key for disclosure reasons as in earlier surveys and 2) for 1998, geographical variables for those special cases were simulated and not subjected to data swapping. In each of the surveys, all of the dollar variables were subjected to rounding (see individual SCF codebooks for details). Large (absolute) negative values were bounded at -$1,000,000. A relatively small number of cases were subjected to data blurring by other unspecified means.

As previously mentioned, the SCF sample excludes people included by Forbes in their list of the 400 wealthiest people in the U.S. In the 1998 SCF, however, there were four observations (after normal data processing and imputation for missing values) that had net worth greater than the minimum needed to get into the Forbes list. Because of concern about the potential identifiability of these cases, it was decided to remove them from the public dataset.

It is worth noting that users of the public dataset will not be able to tell for certain which data items have been altered for disclosure purposes or which cases were selected for special treatment. Users who find that their analysis is unduly hindered by the constraints in the public dataset are encouraged to communicate their concerns. In many instances it has been possible to address these concerns either by revising disclosure protections, or by running analyses on the internal data.

Analysis of Disclosure Adjustments

This section concentrates on comparing estimates derived from the public version of the 1998 SCF with those results obtained using the unaltered final SCF file (internal dataset). Ideally, these results would not show major differences, since a high priority in the design of these adjustments is to avoid changing the underlying integrity and usefulness of the data. This discussion focuses on household wealth (net worth), although stock equity (direct and indirect holdings), income, total assets and debts were reviewed, as well, with very similar results. The analyses performed include overall distributional comparisons, mean comparisons by Census region, and a robust regression analysis.

Table 1 shows estimates of aggregate holdings and the percent of total aggregate holdings of the net worth of groups defined in terms of percentile groups of the net worth distribution as measured by the public and internal datasets. Standard errors with respect to imputation and sampling are also shown[2]. All of the estimates, as well as, their corresponding standard errors from both datasets are virtually identical. The largest difference

Table 1. Proportion of Total Net Worth Held by Different Percentile Groups: 1998 SCF, Internal and Public Use Datasets. All dollar values given in billions of 1998 dollars.

	Percentiles of the net worth distribution									
	All Families		*0 to 89.9*		*90 to 99*		*99 to 99.5*		*99.5 to 100*	
Dataset	$	% of total	$	% of total	$	% of total	$	% of total	$	% of total
Internal	28,928.9	100.0	9,042.1	31.3	10,045.5	34.7	2,362.1	8.2	7,464.3	25.8
	1,684.8	*0.0*	*622.6*	*1.7*	*931.8*	*1.7*	*204.1*	*0.5*	*590.1*	*1.8*
Public	28,908.7	100.0	9,044.9	31.3	10,042.6	34.7	2,362.8	8.2	7,445.8	25.8
	1,684.7	*0.0*	*622.2*	*1.7*	*929.3*	*1.7*	*205.2*	*0.5*	*591.3*	*1.8*

Standard errors due to imputation and sampling are given in italics.

(18.5 Billion for aggregate holdings of the top ½ wealthiest families) is statistically insignificant. Thus, for this analysis, little measurable error was introduced into the public dataset by implementation of the disclosure adjustments.

Figure 1 shows a Q-Q plot of net worth estimated from the internal data versus net worth estimated from the public data. The inverse hyperbolic sine transformation with a scale parameter of .0001 was used. This transformation eliminates exaggerations near zero and compresses large spreads in the tails of the distribution. To avoid graphical aberrations caused by the very few values of negative net worth (and to avoid disclosure of additional information about cases with large [absolute] values of negative net worth), all negative values were set to zero. Also shown are lines corresponding to the 90th (solid line), 95th (dashed line), and 99th (rightmost dashed line) percentiles of the net worth distributions. A Q-Q plot lying on the 45 degree line would indicate that the distributions are identical.

An investigation of the figure reveals only very small distortions, and those well past the vertical line corresponding to the 99th percentile where data can be more sparse. This plot is consistent with the findings from Table1.

The SCF is often used in economic modeling. Economists often try to understand what variables (e.g. asset holdings, income, demographic characteristics, and environmental factors) influence behavior. It is important that changes introduced into the data during the disclosure review process preserve the interrelationships between variables. Robust regressions predicting total family income using a set of dependant variables were performed using both the internal dataset and the public dataset[3]. These variables included log of head of household age, whether there was a checking account, logged amount in all checking accounts, and similar variables for IRA's, money market accounts, CD's, savings accounts, mutual funds, savings bonds, regular bonds, stocks, cash value of life insurance, and face value of life insurance. The results of these regressions showed that both the magnitude of the parameter estimates and their associated signs where nearly identical. T-statistics testing whether a variable was significantly different from zero were also nearly the same, in magnitude and sign for both sets of estimates. Of course, certain variables in the public dataset that were collapsed or top/bottom coded might cause limitations to a researcher's modeling efforts, but the results from these analyses are encouraging nevertheless.

In order to review the effects of simulating the geography variables instead of applying data swapping, Q-Q plots (Figure 2 - Figure 5) were constructed for net worth , debt, equity (direct and indirect stock) holdings, and total family income, all by Census region. The plots shown here are for Census region 1 (Northeast), but plots for the other three Census regions were similar. What is apparent is that the lines for all four variables show a good fit to the 45 degree line for almost the entire distribution. Small distortions are evident for net worth and larger ones for the other three variables of interest,

Table 2: Mean value of net worth in thousands of 1998 dollars; by Census region

Census region	*Internal* data	*Rounding* only	*All adj.* ex. region	*Public* data
1	302.4	302.4	302.3	319.4
2	252.1	252.1	252.1	261.5
3	267.5	267.5	267.5	253.7
4	328.1	328.1	327.6	325.5

but not until well into the upper tail past the 99th percentile. Note that the Q-Q plot (Figure 1) for aggregate net worth showed very minimal distortions even in the upper tail region. It is clear that simulating geography did add some noise to these distributions in the upper tail once we checked by Census region.

Table 2 shows mean values of net worth by Census region for particular versions of the 1998 SCF dataset, including an interim version where the only disclosure adjustments made were rounding, and another version that included all disclosure adjustments except simulation of geography. It is easy to see that both interim datasets produced mean values very close to the mean values from the internal dataset. However, after including the geography simulations (final public dataset), the means do vary slightly by region. This result is consistent with Figure 2. The results are similar for debt, equity holdings and total family income. Note that the largest difference (17.2) is for Census region 1, but it is statistically insignificant since the standard error for the mean of net worth for Census region 1 is 26.1.

Conclusions and Future Plans

This paper provides some encouraging results regarding the disclosure adjustments used in creation of the 1998 SCF public use dataset. As in 1995 (see Fries, Johnson, and Woodburn [1997]), controlled simulation of reported monetary variables and geography seemed to have little effect on univariate distributions aggregated to the U.S. population for a given set of important variables. This is important, since there was a slight modification from the 1995 disclosure review with respect to the controlled simulation strategy. For the given analyses presented in this paper, there were no significant differences between results produced using the public data and those using the internal data. Again, it is worth noting that restrictions on the amount of detail (e.g. categorical collapsing of 3-digit Census occupation codes and 3-digit Census industry codes) could limit certain types of analyses, but such tradeoffs are necessary in order to protect the privacy of individual respondents.

Simulating geography did seem to increase noise in the very upper tails of individual distributions when examined by Census region, and this finding will be investigated in future work.

Acknowledgments

The views presented in this paper are those of the authors alone and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System, or the Internal Revenue Service. The authors would like to express their gratitude to all of the SCF staff for their support with the disclosure review implementation, and especially Annelise Li and Amber Lytle for outstanding research assistance. A special thanks to Arthur Kennickell for invaluable guidance and comments.

Endnotes

1. The full version of the paper will be available on the Internet at:
www.federalreserve.gov/pubs/oss/oss2/method.html

2. The standard error for statistic X is estimated as $SX_{tot} = \{(6/5)*SX^2_{imp} + SX^2_{samp}\}^{1/2}$, where the imputation variance SX^2_{imp} is given by $SX^2_{imp} = (1/4) * \Sigma_{i=1 \text{ to } 5}(X_i - \text{mean}(X))^2$
and the sampling variance SX^2_{samp} is given by
$SX^2_{samp} = (1/999) * \Sigma_{r=1 \text{ to } 999}(X_r - \text{mean}(X))^2$.

3. This model is not intended as a structural description of household income. Its purpose is only to look at the sensitivity of the estimated partial correlations to the disclosure adjustments.

References

Fries, G., B. Johnson, and R.L. Woodburn [1996] "Disclosure Review and its Implications for the 1992 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1996 Annual Meeting of the American Statistical Association, Chicago, IL.

Fries, G.B. Johnson, and R.L. Woodburn [1997] "Analyzing the Disclosure Review Procedures for the 1995 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1997 Annual Meeting of the American Statistical Association, Anaheim, CA.

Fries, G., and R.L. Woodburn [1994] "The Challenges of Preparing Sensitive Data for Public Release," *Proceedings of the Section on Survey Research Methods*, 1994 Annual Meeting of the American Statistical Association, Toronto, Canada.

Fries, G., and R.L. Woodburn [1995] "Using Graphical Analyses to Improve all Aspects of the Survey of Consumer Finances," *Proceedings on the Section of Survey Research Methods*, 1995 Annual Meeting of the American Statistical Association, Orlando, FL.

Kennickell, A.B. [1998] "Multiple Imputation in the Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1998 Annual Meeting of the American Statistical Association, Dallas, TX.

Kennickell, A.B., and D.A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," Proceedings of the Section of Survey Research Methods, 1993 Annual Meeting of the American Statistical Association, San Francisco, CA.

Kennickell, A.B., and R.L. Woodburn [1999] "Consistent Weight Design for the 1989,1992, and 1995 SCFs, and the Distribution of Wealth," *Review of Income and Wealth* (Series 45, number 2), June 1999, pp. 193-215.

Figure 1: Q-Q plot of net worth all families

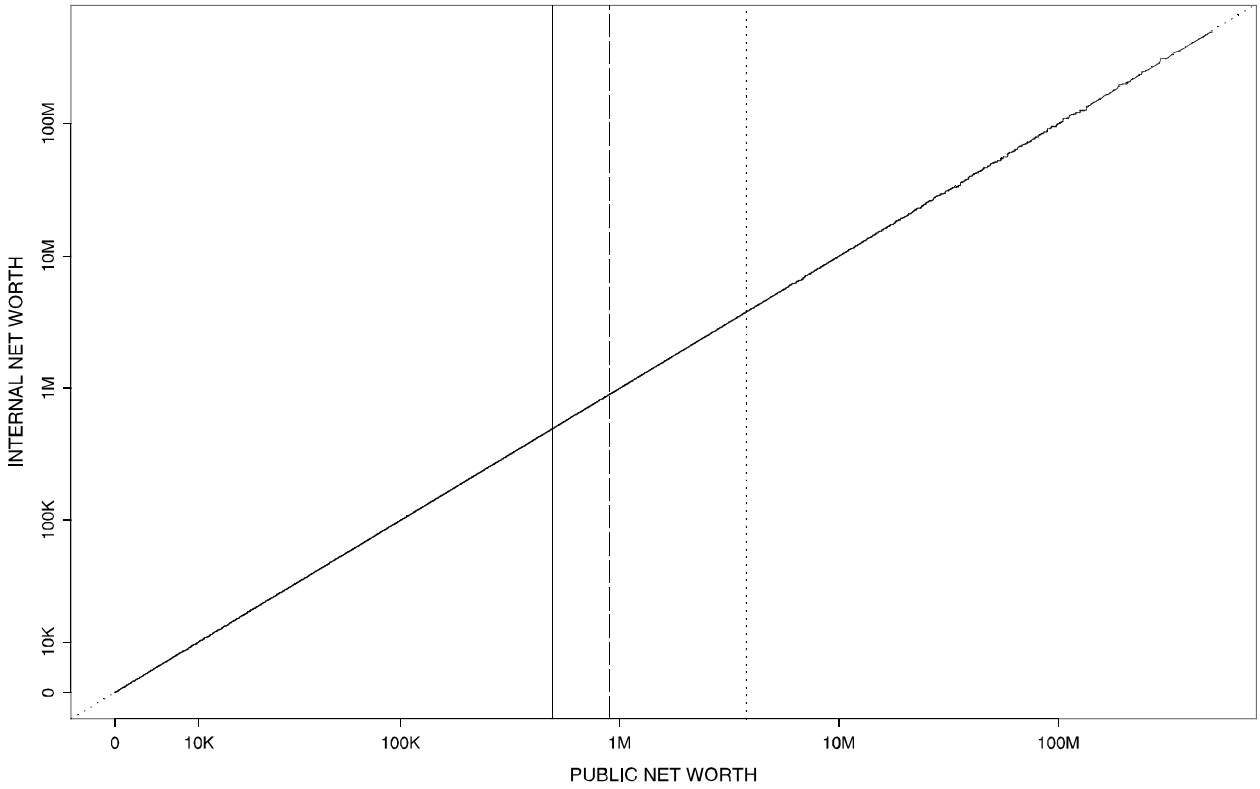

Figure 2: Q-Q plot of net worth region 1

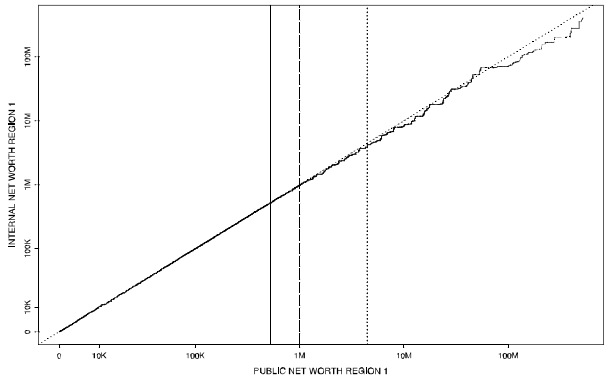

Figure 3: Q-Q plot of equity holdings region 1

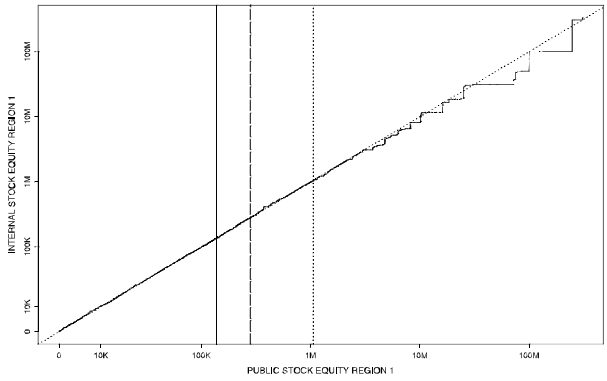

Figure 4: Q-Q plot of debt region 1

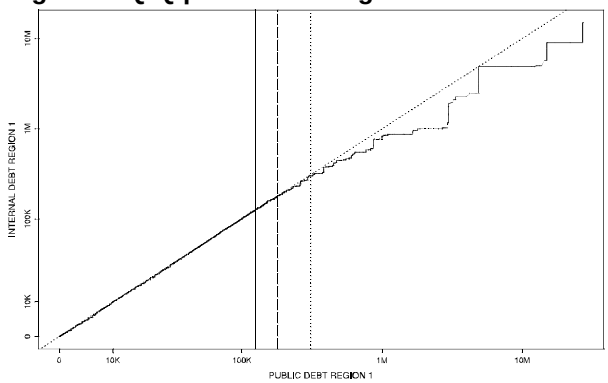

Figure 5: Q-Q plot of income region 1