TOWARD THE DEVELOPMENT OF AN OPTIMAL STRATIFICATION PARADIGM FOR THE SURVEY OF CONSUMER FINANCES

Martin Frankel
National Opinion Research Center and Baruch College

Arthur Kennickell
Board of Governors of the Federal Reserve System

Revised October 1995

Prepared for the 1995 Annual Meetings of the American Statistical Association, Orlando, FL

KEY WORDS: Model-based sampling, Administrative records

The authors would like to thank James Faulkner for outstanding research assistance. Louise Woodburn has been deeply involved in the development of sampling methods for the SCF and has played an important role in this paper; only confidentiality considerations kept her from being an author of this paper. Thanks also to Barry Johnson whose help was invaluable in obtaining permission to use the data that underlie an important part of this analysis. Thanks to Dan Skelley, the director of SOI, for his invaluable and continuing assistance. We are particularly grateful to Fritz Scheuren whose insights and encouragement have guided the SCF in innumerable ways. The authors alone are responsible for all errors and opinions in this paper. Address correspondence to: Arthur Kennickell, Federal Reserve Board, Mail Stop 180, Washington, DC 20551, or email m1abk00@frb.gov .

The survey of Consumer Finances (SCF) is sponsored by the Board of Governors of the Federal Reserve System (FRB) in cooperation with Statistics of Income at the Internal Revenue Service (SOI). Data for the survey are collected by the National Opinion Research Center at the University of Chicago (NORC). The mission of the SCF is to collect detailed information on the finances of U.S. households for use in research and policy analysis. For these purposes,, it is important to have adequate representation of the distribution of financial variables that are broadly distributed in the population (such as credit card ownership), and those ones that are relatively narrowly distributed (such as direct holdings of corporate stock). To this end, the SCF employs a dual-frame sample design: an area-probability sample to give good coverage of broadly distributed variables, and a list sample which is intended to over-sample households that are more likely to be wealthy.[1]

This paper focuses on two problems with the list sample that were raised by Kennickell and McManus [1993] (K&M). For reasons of economy, the list sample is not selected independently of the area sample. Eligibility for the list sample is restricted to households in the PSUs selected for the area sample. As K&M noted, the evidence suggests that the assumptions underlying this decision may be inefficient. Here we bring additional evidence to bear on this question. A second point raised by K&M is the adequacy of the model-based algorithm used to stratify the list sample. As noted in more detail below, in the past the list sample has used an index number developed as a proxy for net worth as a stratifier. As K&M noted, this index number turns out to have a low correlation with net worth, and they highlight the potential usefulness of validating the index by merging selected survey and frame data. Here we present such a validation exercise performed for the purpose of sample selection for the 1995 SCF.

The plan of this paper is as follows. First, we give an overview of the SCF list sample. Next, we examine the geographic distribution of the list sample and discuss the implications for the current sampling procedures. Third, we discuss the use of frame data to model net wealth for the purpose of creating a more efficient stratifier for the 1995 list sample. Finally, we summarize our findings and point in the direction of additional research.

I. List Sample Design

The SCF list sample is drawn from the Individual Tax File (ITF), a sample of individual income tax returns selected and maintained by SOI.[2] This file is largely used in modeling responses to changes in the tax code and a version of this file, blurred in significant ways, is made available to private researchers. The ITF is stratified by several types of income, including business, farm, and other types of income, and the design oversamples taxpayers who have high income or other unusual characteristics. Although the ITF is itself a sample, for very high incomes the sampling rate is quite high. The 1990 ITF, the basis for the 1992 SCF sample, contains about 120,000 tax records, mostly returns for tax year 1990.[3]

The list sample is selected in two stages. At the first stage, it is assumed that the geographic distribution of list cases is the same as that of the general population of households (largely with the goal of controlling interviewer costs). Reflecting this assumption, the entire ITF is subsetted to include only filers with addresses in the PSUs selected for the area-probability sample, and the ITF measure of size (the weight) of each case in the selected PSUs is inflated by the inverse of the probability of selection of the PSU.[4] The effects of this assumption on the efficiency of the sample is discussed in more detail in the next section.

At the second stage of selection, this subset of cases is separated into strata defined in terms of a "wealth index," which is intended as a proxy for the net worth of the tax filer. This index is based on a capitalization of income flows assuming an average rate of return.[5] The exact form of the index used in 1992 is given by

WINDEX = Home Equity + ABS(taxable interest income)/.1165 + ABS(nontaxable interest income)/.067 + ABS(dividends)/.057 + ABS(rents and royalties)/.115 + (ABS(S-Corp. income) + ABS(estate and trust income))/.230 + (ABS(Schedule C gross) + ABS(Schedule F gross profit) + ABS(other farm income))/.172 + ABS(long-term capital gains) + ABS(short-term capital gains),

where ABS represents the absolute value function.[6] All list cases are assigned a value for home equity, which is estimated separately by the original ITF strata using values estimated from earlier SCFs. The rates of return were determined from aggregate data and are assumed to be uniform for all taxpayers.[7]

Using this wealth index and the PSU-probability-adjusted ITF weight as a measure of size, in 1992 cases were divided into the 8 strata shown in table 1. Stratum 8--filers with a wealth index of more than 250 million--were not sampled at all.[8] Using PPS, strata 2 through 7 were over-sampled at progressively higher rates, and stratum 1 was under-sampled. One might question the efficiency of including stratum 1 cases in this sample given that such units are likely to be generously covered by the area-probability sample. These cases were included for two reasons: first, for weighting (see Kennickell, McManus and Woodburn [1995]) it is important to have an overlap in the two samples; second, as an extra precaution in protecting the privacy of taxpayers, including these cases removes the certainty that list cases are wealthy.

Table 1: Definition of List Strata, 1992 SCF

Stratum number	*Units of index*
1	Less than 100,000
2	100,001 to 500,000
3	500,001 to 1,000,000
4	1,000,001 to 2,500,000
5	2,500,001 to 10,000,000
6	10,000,001 to 100,000,000
7	100,000,001 to 250,000,00
8	More than 250,000,000

As a part of the agreement with SOI, a special approach is taken to interviewing the list sample cases. Before these cases are approached by an interviewer, they are mailed a package containing a description of the survey, and letters from NORC and from the Chairman of the Federal Reserve Board requesting cooperation with the survey. Also enclosed is a postpaid postcard to be returned if the individual does *not* wish to be interviewed.[9] Interviews are attempted with all taxpayers not returning the postcard.

Not surprisingly, response rates are not high for the higher-stratum cases (see Kennickell, McManus and Woodburn [1995]). However, rather than being a singular defect of the survey, this knowledge is actually a strength. Presumably other surveys also have latent differential nonresponse by wealth groups that is lost in the aggregate response rates that are typically reported. The advantage of the SCF is that there is actually frame information to *identify* the problem, and to be used to make systematic adjustments.

II. Geographic Distribution of High-Strata ITF Cases

As discussed above, the list sample implicitly accepts the proposition that the distributions of the cases in the various list strata are the same as that for the general population. For units in strata 1 and 2, this condition holds strongly because these groups comprise the great majority of the population--78.1 percent of filers were in stratum 1 and 17.3 percent in stratum 2. At the opposite

end of the wealth distribution, the question is not a priori obvious, and earlier evidence presented by K&M suggests that high-wealth-index strata cases may cluster much more strongly than the general population.

Using more comprehensive information than K&M, we find compelling evidence of clustering. Figure 1a shows a smoothed estimate of the population density over the PSUs eligible for selection at the first stage of the area-probability sample using 1990 Census data.[10]

Figure 1b shows a comparable smoothed geographic distribution of an estimate of the ratio of the number of cases in stratum 5 or higher to the total population. For this figure, we used the census data underlying figure 1a and all filers at U.S. addresses in the 1990 ITF. Because the ITF is not a universe sample, some precautions were necessary to obtain robust estimates.[11] If cases in stratum 5 and above were distributed across the country like the general population, the figure would be flat. Two points are clear. First, high-index cases cluster strongly in the largest MSAs, which are sampled with probability one in the area-probability sample.[12] About 50.0 percent of such filers are in the 19 self-representing PSUs of the area-probability sample, compared with 36.2 percent of all households. Second, there are a few areas with relatively low general population density that contain a disproportionate number of high-strata cases.

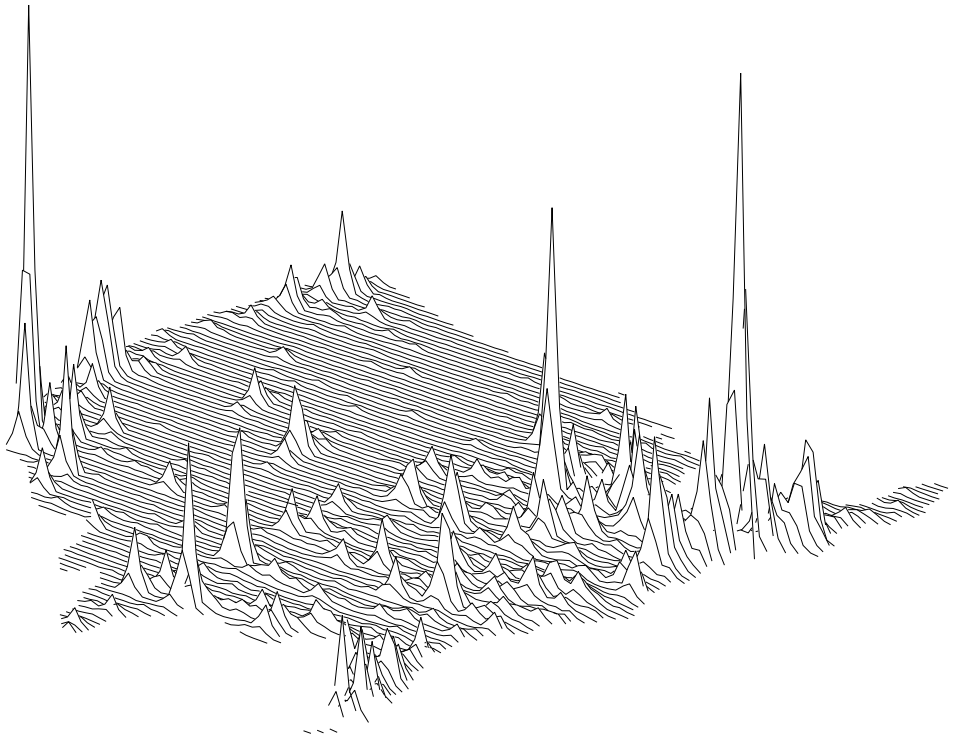

Figure 1A: Smoothed Distribution of All U.S. Households, by PSU, 1990

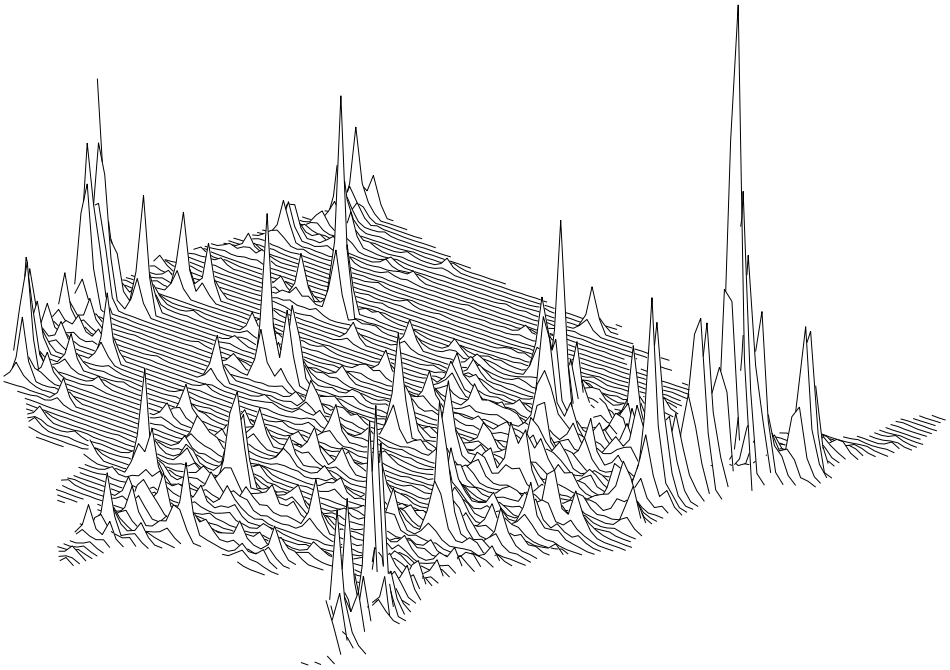

Figure 1B: Smoothed Distribution of Ratio of Stratum 5-7 Filers to All U.S. Households, by PSU, 1990

Figure 1: Distribution Over Continental U.S. of All HHs and High-Stratum Filers, 1990

Another way to examine the effects of the concentration of the list population is look at how the set of PSUs would change if we redrew the sample of PSUs using probability proportional to the number of units in stratum 5 and above in each PSU, rather than the total number of households. Although it is quite difficult for us to replicate the drawing of the entire sample, it is straightforward to determine which areas would be considered self-representing in a PPS design based on the high-strata tax filers. The assumption that we draw 100 PSUs as was done for the area-probability sample determines a sampling interval of about 4200, and fifteen areas (MSAs, CMSAs, and counties) have a size larger than this interval: New York; Los Angeles; San Francisco; Chicago; Boston; Philadelphia; Dallas-Fort Worth; Houston; Washington, DC; West Palm Beach, FL; Detroit; Seattle; Atlanta; San Diego; and St. Louis. The largest two areas account for about 20 percent of the total, and this group together contain about 74 percent of the total. Recomputing the sampling interval after removing the first set of self-representing areas implies a second, larger set of 22 PSUs (containing about 8 percent of all cases), and a third recalculation adds another 4 areas (containing about 1 percent of all cases). Altogether the 41 self-representing PSUs account for about 83 percent of the total high-strata cases.

All of the 19 areas that are considered self-representing in the area-probability sample are also self-representing in this hypothetical list design. Necessarily, the converse is not true. West Palm Beach, the 10th largest of the self-representing list PSUs, is a PSU in the area-probability sample being used for the 1995 SCF (it was not even in the sample for the 1992 SCF), but it is not self-representing (it is the 35th largest such PSU). In the second and third tranches of self-representing list areas (after recomputing the sampling interval), the areas that emerge are a mixture of older industrial cities such as Pittsburgh, Cleveland, Milwaukee, Rochester, etc., other large cities such as Miami, Denver, Minneapolis, etc., and some exurban counties similar to West Palm Beach (mainly in areas associated with natural resources or retirement). About 4 percent of all the high-strata cases are estimated to be in the 10 areas in the second and third tranches of hypothetically self-representing PSUs that are not included in the actual sample.

After selecting the hypothetical self-representing areas, about 17 percent of the population remains in the 2448 unselected areas, from which an additional 59 PSUs would be selected. The unselected areas contain 53.1 percent of all households. Nearly 2/3s of the remaining areas are estimated to contain one or no high-strata cases and these areas contain 13.2 percent of all

households. Although some of the zeroes are not "true" zeroes, it is still likely that the high-strata cases are very thin in such areas. Figure 2 shows a plot of the rank in terms of number of households against the rank in terms of number of high-strata cases for the 843 areas that are estimated to contain more than one high-strata case. There is much variation between these rankings, particularly outside the the self-representing PSUs.

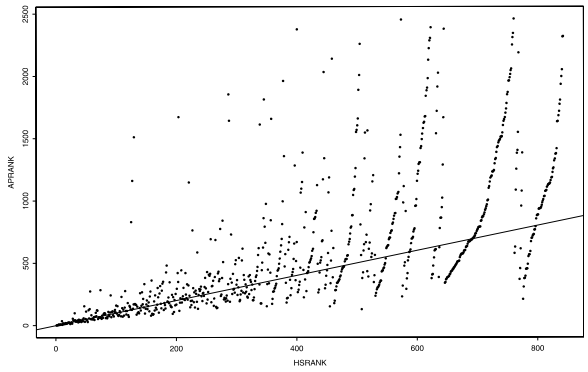

Figure 2: Household Rank vs. High-Strata Rank

Basing the list sample on the areas selected for the area-probability sample does deviate from PPS sampling for the high-strata cases. Only cases in the self-representing areas in both the area-probability and hypothetical samples have the "correct" size at the second stage of selection of the list sample. Cases in other areas included in the sample have a second-stage size measure that is "too large." Effectively, wealthy people in less densely populated areas are more likely to be selected than would be the case under true PPS sampling.

Although these findings came too late to alter the selection of the 1995 SCF sample, they will have an effect on the weighting of the sample cases and the selection of future list samples. One possibility may be the following. Because such a large fraction of PSUs have a very small number of high-strata cases, the number of "pseudo-PSUs" for such cases may be too large. If we choose, say, 75 PSUs to represent the high-strata cases, we would have 17 PSUs that are self-representing in this sense--all in the AP sample as well, but not all self-representing in that framework. Applying Kefitz sampling to the remaining PSUs in the AP sample, it appears, based on a visual inspection of the data, that we would be able to select most of the remaining list pseudo-PSUs from among the remaining 82 AP PSUs that have non-zero high-strata cases.

III. Modeling Net Worth

The wealth index described above has always been seen as an ad hoc approximation to net worth. As shown by the cross-plot in figure 3 of PNW3, the logarithm of the index (linearly adjusted by OLS), against the logarithm of net worth in 1992, the relationship is noisy: the Spearman correlation is only .76.

Ever since this device was used in the design of the 1989 SCF, efforts have been made to obtain permission to validate the index in a way that had no chance of violating confidentiality pledges made to respondents to the survey or important ethical principles. Negotiations involved outside advisors, the confidentiality committee at NORC, staff at SOI, and the authors. Ultimately, it was agreed that for the limited purposes of this analysis, a special linked file could be created from selected items in the 1990 ITF, and the 1992 SCF.[13] This file contained no identifiers after merging of the data, and all work took place on an isolated file system at the Federal Reserve accessible only to Kennickell. No name and address information is available to the Federal Reserve. No information from the linked file other than some model estimates was available to either SOI, NORC, or anyone else.

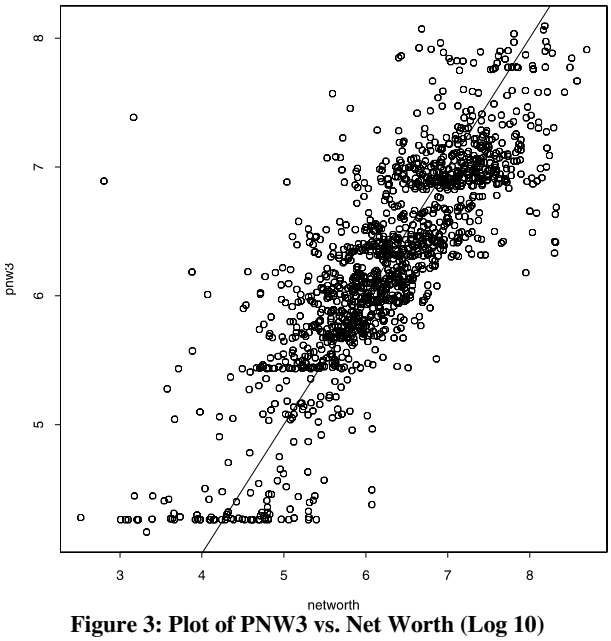

Figure 3: Plot of PNW3 vs. Net Worth (Log 10)

As noted earlier, underlying the wealth index is a notion that wealth can be modeled in terms of income flows. In the original wealth index, rates of return for each income type have been approximated using market data. If we take this model and estimate the coefficients from the data via OLS, we obtain apparently reasonably sensible implied rates of return for a few items: for taxable interest, 7.0 percent; for non-taxable interest, 9.4 percent, and for dividends, 17.5 percent.[14] However, other terms are either implausible in size or of an "incorrect" sign. Several factors are probably large contributors to the poor fit. (1) The income data are for tax year 1990, but the wealth data are for 1992, and people may have substantially rearranged their portfolios over that time. In the future, we would like to match 1992 income data with the survey data to test this proposition. (2) Rates of return are unlikely to be constant across individuals (see K&M for some evidence on this). Various factors in the model likely proxy for such variation. (3) Because the data are very highly skewed in many dimensions, it is likely that the fit of OLS on such data is poor. One way of dealing with this issue might be to use some type of robust estimation. However, time was very limited if we hoped to use the results of this exercise for selecting the 1995 list sample. Because the

ability to search over classes of models is important for this exercise and because our existing programs are based on OLS, we simply used a logarithmic data transformation to lessen the likelihood of our models' being affected by outlying values.

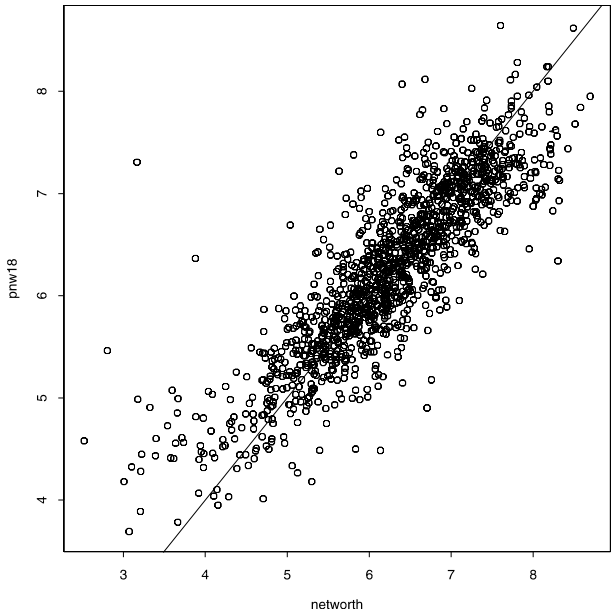

Figure 4: Plot of PNW18 vs. Net Worth (Log 10)

The final model was selected using a forward search routine. Variables available for selection included up to the second power of the logarithms of all of the variables in the original index, in addition to wage and salary income, pension income, deductions, real estate taxes paid, filing status, and age of the principal filer. After the search routine, a model was constructed retaining all powers of a variable lower than the highest selected on (e.g., if the model selected the second power of the logarithm of pension income, the first power and the dummy variable indicating the presence of the income type were also included regardless of whether they were selected by the search routine). The fitted values of this model (PNW18) are plotted against net worth in figure 4. The adjusted R^2 of this estimation was .730, an increase of .073 over the model in figure 3.

The final model appears to represent a substantial improvement over the original wealth index in terms of the ability of the 1990 ITF data to predict 1992 net worth. The form is flexible enough to pick up sources and variation in wealth that cannot be captured by the wealth index. However, there is still some risks in using this model to develop strata for the 1995 SCF. Whether recognized explicitly or not, rates of return are important to the predictive power of the model. If the generating process of income changes over time, the meaning of income changes in the model (equivalently, the model coefficients may be time-varying). Over the period 1990 to 1993 (the dates of the tax data used for the 1992 and 1995 SCFs, respectively), some rates of return changed substantially: for example, 6-month CD rates fell from about 8 percent to about 4 percent. It is not possible to alter the fitted model to account for such changes without some very strong assumptions. The original wealth index is easy to alter, but it misses some important indicators of wealth. Using income data from the 1993 ITF, we computed a compromise stratifier for the 1995 SCF list sample combining

information from an updated version of the original wealth index and a predicted value of net worth using the coefficients from the final fitted model. Because the two distributions differ, we standardized them to have the same mean and standard error and took a simple average. To keep the stratum sizes comparable to those in 1992, we defined the stratum boundaries in terms of percentile breaks comparable to those implied by the wealth index strata in the 1992 list sample.

In the future we hope to refine this process in several ways. First, we would like to reestimate the model with concurrent income and wealth data. This is probably technically feasible, it may not be possible for other more complicated reasons, including the possibility that this would be seen as too much of an invasion of respondents' privacy. Although it is very unlikely that such information could ever be available for sampling, it could provide a useful gauge of the misclassification due to the use of dated data. Second, it would be very useful to investigate more fully the differences in classification under various models. Finally, it may be useful to incorporate formally the probability of misclassification under various models in sample selection.

ENDNOTES

1. Heeringa, Connor, and Woodburn [1994] describe the basic design of the SCF sample.

2. The SOI data are describe in *Individual Income Tax Returns, 1990* [1993]. In general, statistical and research uses of SOI data are closely regulated to guarantee that individuals (and other entities) will remain protected against any disclosure of their financial and tax data (e.g., Wilson and Smith [1983]). For the SCF, contractual agreements between the FRB, NORC, and SOI clearly specify the limitations on the use of the administrative data and require that any use of the data must satisfy the strictest standard of protection of the three organizations.

3. The ITF also contains some returns for earlier years, multiple returns for the same taxpayer (initial and revised returns, or multiple years of returns), and returns for taxpayers who do not live in the U.S. For the SCF sample, all foreign addresses are deleted; for filers with multiple returns, only the most recent return is retained.

4. Some addresses may be that of a tax preparer, rather than the filer. Evidence from earlier surveys suggests that this tends to generate significant "gate-keeper" problems, but no significant geographic distortions.

5. For example, if a taxpayer reports $100 in interest income and the assumed interest rate is 10 percent, then the estimated value of the underlying asset is $1,000.

6. The use of the absolute value function here is a little troubling if we believe we are literally computing wealth by grossing up income flows. The reasoning is that there are very few cases with negative income at the level of the components we use and anyone with negative income must have substantial assets to sustain such a flow.

7. For this analysis, a sample couple filing a joint return that had divorced by the time they were contacted were assigned a new value of the wealth index given by WINDEX_D=½(WINDEX-home equity)+home equity. Where filing status was "married filing separately," both spouses are assumed to have filed identical returns and their weight and stratum were adjusted accordingly.

8. The total number of cases in the highest stratum is very small and the probability of obtaining an interview is remote. Even though the top group probably controls a large amount of assets, the fraction of net wealth held by the group is small and might be more precisely estimated from other sources, such as *Forbes*.

9. This design has been in place since the 1989 survey. In 1983, the postcard was to be returned only if the person agreed to be interviewed; the response rate for the list sample was dramatically lower, only about 10%.

10. Alaska and Hawaii are included, but not shown.

11. Briefly, for each PSU, no case included in the estimation was allowed to have a weight larger than the number of high-strata cases in the area. This constraint applies to 375 cases in stratum 5 or above. The most serious such truncation is a weight of 260 that is reduced to 1. The vast majority of the truncated weights are quite small prior to the truncation, and most of the affected areas are small rural areas. Virtually the same pattern emerges in the analysis is restricted to strata 6 and higher, for which the ITF sample is more like a census.

12. The spike above New York indicates that the density is over twice as high as the national average. Although some of the other areas are substantially higher, the plotting algorithm imposes a strong smoothness criterion that tends to flattens more isolated peaks.

13. The version of the SCF used was the first iteration of a multiple imputation routine. The fully multiply-imputed dataset was not completed in time for this analysis.

14. Cases that changed marital status between the time the return was filed and the time of the survey (as determined from the survey data) and 27 cases with zero or negative net worth were deleted from this analysis.

BIBLIOGRAPHY

Heeringa, Steven, Judith Conner, and R. Louise Woodburn [1994] "The 1989 Surveys of Consumer Finances: Sample Design Documentation," working paper, Survey Research Center, University of Michigan.

Statistics of Income Division, Internal Revenue Service [1993] "Individual Income Tax Returns 1990," Washington, DC.

Kennickell, Arthur B. and Douglas A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," *Proceedings of the Section on Survey Research Methods*, 1993 Annual Meetings of the ASA.

_____, _____, and R. Louise Woodburn [1995] "Weighting Design for the 1992 SCF," mimeo Federal Reserve Board.

Wilson, O. and Smith, W.J. Jr. [1983] "Access to Tax Records for Statistical Purposes," *Proceedings of the Section on Survey Research Methods*, 1983 Annual Meetings of the ASA.