DISCLOSURE REVIEW AND THE 2001 SURVEY OF CONSUMER FINANCES[1]

Gerhard Fries, Federal Reserve Board
Gerhard Fries, FRB, Mail Stop 153, Washington, DC 20551; gfries@frb.gov

Key Words: Confidentiality, Imputation

Surveys that collect sensitive data and/or sample wealthy families need to be particularly concerned about protecting the confidentiality and privacy of the survey respondents. Every effort should be made to review and filter the data to minimize the chance that someone could be identified, as it should feel 'safe' to be a respondent. While trying to protect the data from potential disclosures, it is also imperative to maintain the usefulness and integrity of the data for policy makers and researchers.

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, I describe the disclosure procedures used in preparing the 2001 SCF data for public release. Including this introduction, there are six sections. In the following section, I 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 2001 SCF and the fourth section investigates the effects of the disclosure adjustments on particular analyses performed. The next section illustrates a potential downside of the public extract version of the SCF containing summary variables. I summarize the 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 of the U.S. Department of the Treasury. Data are collected on household balance sheets, use of financial services, pensions, labor force participation, demographic characteristics, and income. Since 1992, data for the SCF have been collected by NORC, a social science and survey research organization at the University of Chicago, generally between May and December of each survey year. The median length interview for the 2001 SCF required about 80 minutes, although some complicated cases took substantially longer. A high percentage of interviews was obtained in-person, but telephone interviews allowed for the convenience of respondents. All interviews are administered using a program running on laptop computers.

Data are collected on items that are highly concentrated in relatively small segments of the population (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 design including 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 statistical records derived from tax returns. The list sample is stratified on an estimated wealth index, with observations having higher index values selected at a higher sampling rate. The sample records 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 disproportionately include wealthy families, but excludes people described by *Forbes* magazine as being among the 400 wealthiest people in the U.S.

Of the 4,449 completed interviews in the 2001 survey, 2,917 families came from the area-probability sample and 1,532 from the list sample. The response rate for the area-probability sample was about 70 percent. The overall response rate for the list sample was about 30 percent, and for the part of the list sample containing the wealthiest families the rate was only about 10 percent.

Nonresponse is an important issue to address for the SCF. Weighting adjustments compensate for both the complexities of the sample design and for differential patterns of nonresponse. 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]).

Disclosure Adjustments

Under the agreement with the Statistics of Income Division (SOI) of the IRS that allows the SCF to use statistical records derived from tax returns for sampling, the public version of the SCF data set is required to meet disclosure avoidance standards similar to those set for the public version of the files of individual tax returns that SOI creates for research purposes . The particular standards for data release are jointly determined by staff from SOI and the Federal Reserve Board (FRB). The rules developed are implemented by FRB staff, and the resulting data, with explicitly identifying data removed, are jointly reviewed by SOI and FRB staff. Often such review indicates that changes or additions should be made to earlier rules to accommodate unforseen changes in the institutional structure of household finances or classes of events previously unobserved in the SCF data. Usually there are several iterations of this process.

The review of the 2001 SCF data closely paralleled that for the 1998 data (see Fries and Johnson [2000]). The most notable difference was a decision not to include any geographic information whatsoever in the public version of the 2001 data set. Previously (in the surveys for 1992, 1995, and 1998), respondents' locations had been identified to the level of the nine Census divisions.

Because of the very high value of location information to a data intruder, protecting geographical identifiers has always been a major concern. Variables identifying the Primary Sampling Unit (PSU) are never released. Even the regional information released earlier had been systematically blurred. The decision to withhold the limited geographic data previously available was reached after discussions with a broad spectrum of SCF users from academia, government and the private sector who were assembled for a one-day SCF workshop. In addition, SCF users who registered on the project web site (*www.federalreserve.gov/pubs/oss/oss2/scfindex.html*) were contacted to ascertain any potential negative results of withholding the regional data.

The motivation for withholding this information was to allow for the possibility of releasing later other geographically-based data that might be of more value to users of the SCF data. If the division data had been released, release of geographically-based data tied to areas other than the Census divisions would have the effect of revealing a finer level of geography than is allowed for the SCF. At the recent SCF workshop, users expressed strong support for additional variables that would allow for further work on the effects of taxation taking into account variations of taxes over states. Kevin Moore of the SCF staff working with Daniel Feenberg at the National Bureau of Economic Research (NBER) has applied the NBER TAXSIM model (see Feenberg and Coutts [1993]) to simulate federal, state and local taxes and various other tax-related statistics for every observation in all of the SCFs. To a person knowledgeable about state taxation, such information would directly reveal the state of residence for the SCF cases. The proposal discussed at the workshop was to release either indicators for whether cases were in areas of "high", "medium," or "low" state taxes. Alternatively, for each case one might simulate tax variables for all states within the group appropriate to the case and report the average value across those states in the public SCF data.

Aside from the withholding of geographic information, the disclosure adjustments for the 2001 SCF changed the data in a number of ways, some of which may be revealed. All variables relating to the sample design, weight design, and almost all marital history were suppressed. Categorical variables were compared to their respective codeframes and responses reported by a "small" number of respondents and responses that were sufficiently "unusual" were combined with other responses (e.g. among vehicle owners, the category "truck (except pickup)" and "antique/classic/collector vehicle" were combined). Some other discrete variables were top or bottom coded or rounded as needed. An additional set of cases was selected, some reporting unusual assets, liabilities, or income and some at random. For these cases, all originally reported dollar values, except those for car values, were multiply imputed (see Kennickell [1997]). The simulated outcomes were subjected to constraints to ensure that they would lie in a sufficiently "close" neighborhood of the original reported values to minimize distortions to the data for analytical purposes, but be sufficiently variable to provide adequate disclosure protection. Nine cases were removed from the public data set because the net worth of the family exceeded the minimum value necessary to be included in the *Forbes* list of the 400 wealthiest people in the U.S. Finally, some cases were subjected to unspecified data blurring and other unspecified manipulations that should have minimal effect on most analysis of the data but that should introduce significant uncertainty for a data intruder. Importantly, users of the public data set 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.

Analysis of Disclosure Adjustments

This section will compare some estimates derived from the public version of the 2001 SCF with those results obtained using the unaltered SCF file (internal data set). The analysis focuses on distributional comparisons for household wealth (net worth) and before tax family income with some additional income by age cohort comparisons. Disclosure adjustments should not affect the analytical integrity of the data, so it would be desirable for these results to reveal no major differences in the comparisons.

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 data sets. Standard errors with respect to imputation and sampling are also shown[2]. Most of the estimates are fairly similar across the two data sets. The largest difference (232.1 Billion dollars for aggregate holdings of the top one percent of the wealthiest families) is not statistically significant. Part of this gap is explained by the omission of the cases with wealth exceeding the *Forbes* cut-off. A similar table for income showed even more comparability.

Figures 1 and 2 show relative "quantile-difference" (Q-D) plots for both net worth and income. These graph differences in the values of distributions at common percentile points (see Kennickell [1997]). The horizontal axis is labeled in terms of the common percentiles. The vertical axis shows the percent difference in the distributions, where the public data set is taken as the base.

When two distributions are fairly similar, points will not vary "much" from the zero line. The plot for income looks fairly flat with a small spike at the end which could be caused by the omitted cases and/or by disclosure imputation of large values (e.g. a value of \$1,000,000 may be imputed to be \$1,100,00 showing a relative 10 percent difference). Note the huge spikes around the 10th percentile in the plot for net worth. These are caused by small values around zero where even just rounding can cause large percentage changes. In general, these plots show distributions that are very similar and along with the results from Table 1 seem to indicate that the disclosure adjustments did not produce any obvious distributional differences for these variables.

Relative Q-D plots (Figures 3a - 3f) are also shown for income by age cohorts, and again there are no indications of distributional changes in the public data set compared to the internal data set. Of course, these are just a few examples, but nevertheless are encouraging.

Note: Disclosure Adjustments and the SCF Aggregate Variable Public Use File (Excel Format Extract)

The project web site provides several versions of the public data set. They include both a full SAS version and an ASCII version where the variables are basically the answers to all of the questions from the survey instrument (i.e. no summary variables (e.g net worth)). Also included is a data extract in Excel format which does contain summary variables and other variables which were used to produce the tables in the January 2003 Federal Reserve Bulletin summary article: "*Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances".* This extract gives SCF users who do not have SAS or other sophisticated statistical software packages an alternative data set. With a tabling utility also provided, such users can easily calculate weighted medians, weighted means, and percent type calculations for the summary variables provided. The documentation includes a program which defines the construction of all of the summary variables.

This program was designed to work as comparably as possible with both the internal and public versions of the data set. However, in some cases codes that fall into different analytical categories in this program have been combined for purposes of disclosure limitation. Normally, the differences in the resulting summary variables are small, but they may be noticeable for some subpopulations.

For example, the variable INSTALL includes a variety of debts customarily treated as types of installment debt. One type of debt included is non-mortgage real estate loans that are not plausibly associated with any of the properties reported in the interview. To make this allocation exactly, it is necessary to know the purpose of a variety of loans collected. However, one consequence of the collapsing of codes is that this information is no longer available in a distinct category, and relying on the collapsed category that contains the code would result in too many other observations being misclassified. For individual cases that are affected by this limitation, differences in the values of INSTALL calculated from the public and internal versions of the data can be substantial. However, at a more aggregated level, INSTALL did not show any major distributional problems. Ideally, analytical distinctions would be taken into account when a disclosure review is conducted, but analytical categories change over time and it is not possible to revise earlier data retrospectively without revealing more refined information about the earlier data. In the SCF data there are other examples of non-nesting codes, but in all cases the distributional consequences are negligible.

Conclusions and Future Plans

As was the case in 1998 (see Fries and Johnson [2000]), controlled simulation of reported dollar figures along with rounding, etc**.** had no obvious distributional effects with respect to the analysis given in this paper. Only net worth and income were examined here. Perhaps, in the future, a more systematic reporting of a variety of other variables can be entertained.

It is important to note that data sets on the SCF web site containing summary type variables calculated from the SCF full version of the public data set can differ substantially at the case level from the SCF internal data. At the aggregate level, this does not seem to be a major concern. In the future, it is expected that the SCF staff will review code collapsing for disclosure as it pertains to affected summary variables in the Excel version of the public data set.

Acknowledgments

The views presented in this paper are those of the author alone and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System. The author would like to express gratitude to all of the SCF staff for their support with the disclosure review implementation and 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
$SX^2_{samp} = (1/999) * \Sigma_{r=1 \text{ to } 999}(X_r\text{-mean}(X))^2$.

References

Feenberg, D. A., and E. Counts [1993] "An Introduction to the TAXSIM Model," Journal of Policy Analysis and Management, Vol. 12, No. 1, Winter 1993, pp. 189-194.

Fries, G., and B. Johnson [2000] "Disclosure Review and the 1998 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods,* 2000 Annual Meeting of the American Statistical Association, Indianapolis, IN.

Kennickell, A.B. [1997] "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," presented at '98, Lisbon, Portugal.

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.

Table 1. Proportion of total net worth held by different percentile groups: 2001 SCF, internal and public data sets. All dollars values given in billions of 2001 dollars.

	Percentiles of the net worth distribution									
	All Families		*50 to 90*		*90 to 95*		*95 to 99*		*99 to 100*	
Net worth	$	% of total	$	% of total	$	% of total	$	% of total	$	% of total
Internal	42,389.2	100.0	11,603.3	27.4	5,139.9	12.1	10,615.2	25.0	13,855.2	32.7
	712.1	*0.0*	*274.4*	*0.7*	*309.0*	*0.7*	*463.9*	*1.1*	*766.1*	*1.4*
Public	42,153.9	100.0	11,599.1	27.5	5,147.6	12.2	10,608.2	25.2	13,623.1	32.3
	673.3	*0.0*	*272.7*	*0.7*	*310.9*	*0.7*	*464.9*	*1.1*	*741.8*	*1.5*

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

Figure 1: Relative quantile difference plots: Income (internal data set) minus income (public data set) as a percent of income (public data set); by quantiles of the distribution of income.

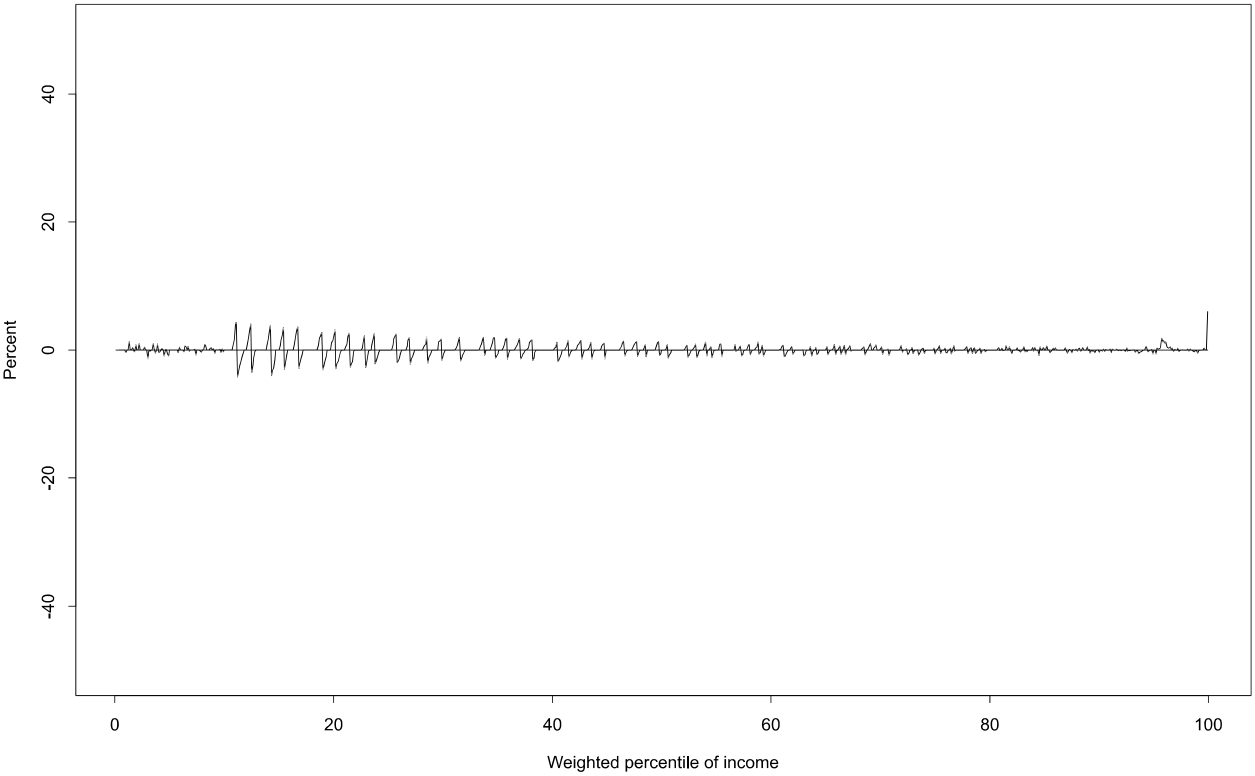
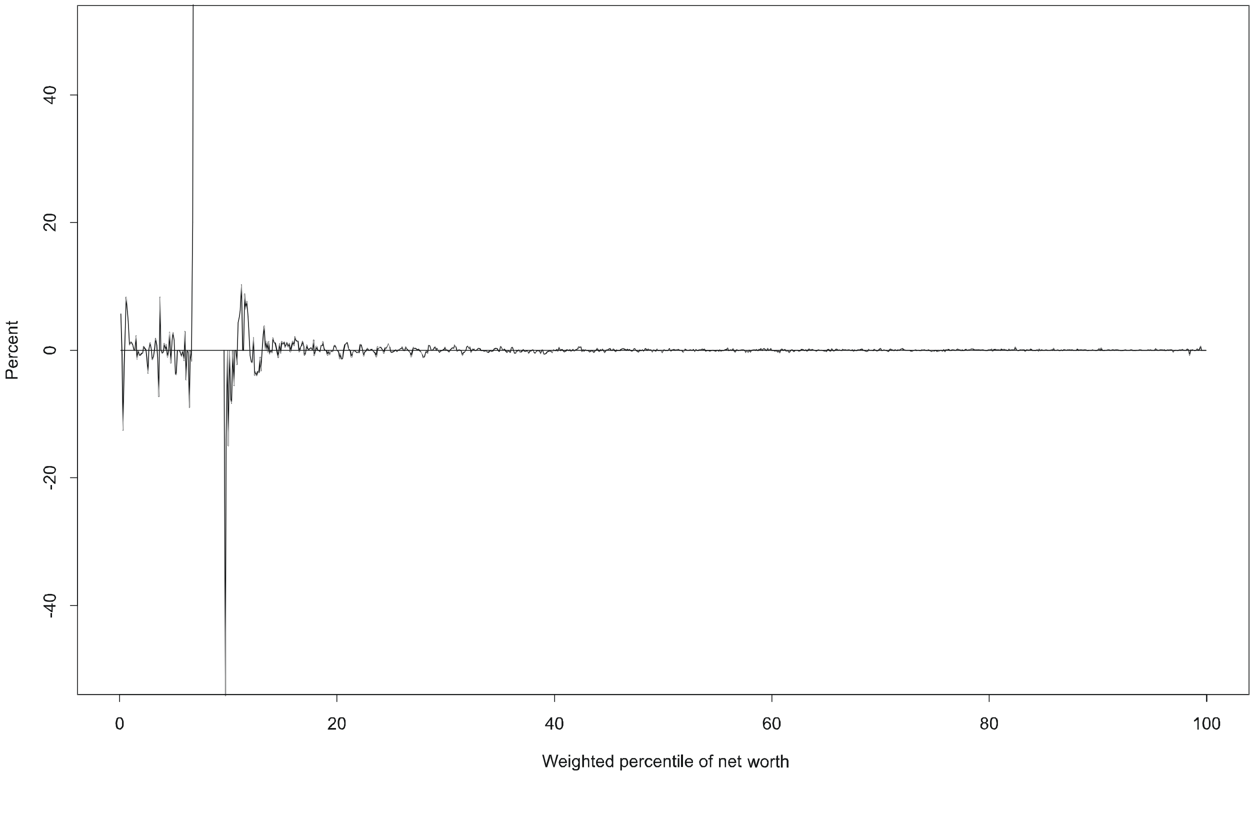
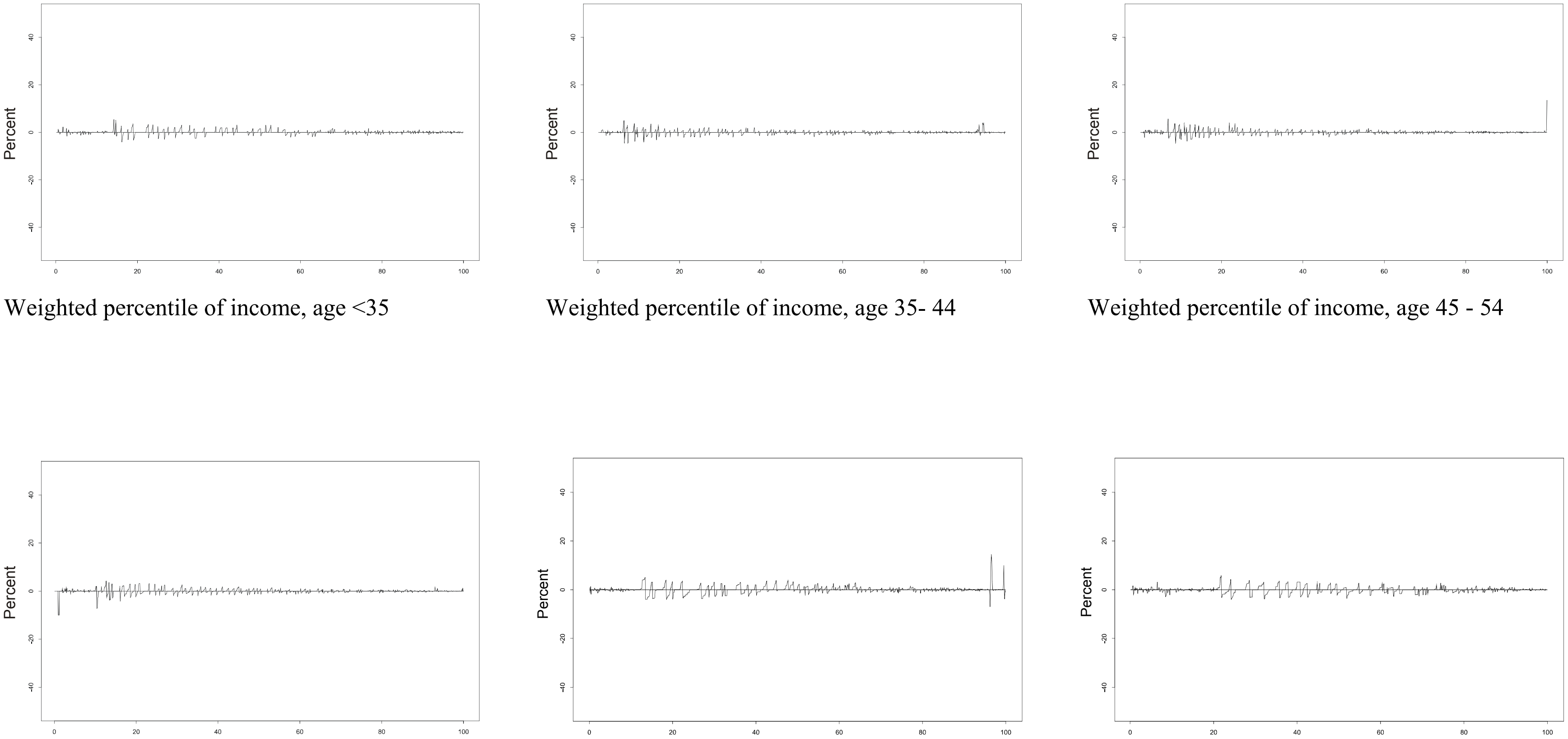


Figure 2: Relative quantile difference plots: net worth (internal data set) minus net worth (public data set) as a percent of net worth (public data set); by quantiles of the distribution of net worth.



Figures 3a-3f: Relative quantile difference plots: Income (internal data set) minus income (public data set) as a percent of income (public data set) by age cohorts; by quantiles of the distribution of income.



Weighted percentile of income, age <35

Weighted percentile of income, age 35- 44

Weighted percentile of income, age 45 - 54

Weighted percentile of income, age 55 - 64

Weighted percentile of income, age 65 -74

Weighted percentile of income, age >=75