THE CHALLENGES OF PREPARING SENSITIVE DATA FOR PUBLIC RELEASE

Gerhard Fries, Federal Reserve Board, and R. Louise Woodburn, Internal Revenue Service
Gerhard Fries, FRB, Mail Stop 180 Washington, DC 20551, m1gxf00@frb.gov

Key Words: Disclosure, Confidentiality, Survey of Consumer Finances

Survey practitioners are challenged to meet the ever rising demand for microdata files while protecting the confidentiality of the individual data provider. Data processing tools are becoming quite sophisticated which is an aid to survey practitioners, but also a potential tool to data snoops. Additionally, there is a perception that participation in surveys is declining world wide; partially due to a rising concern about confidentiality. Thus, survey practitioners must thoroughly protect the identity of the individual respondents. However, it is also important to retain the usefulness of the original data and for inferences made from masked data to be no stronger than those made from the original data. This paper details the preparation of the public release data file for the Federal Reserve Board's Survey of Consumer Finances (SCF), a triennial household survey that includes data on finances, employment, and demographics. We detail our experiences from the 1989 and 1992 surveys. Including this introduction, there are six sections. A brief summary of the literature on disclosure methodology is provided in the second section. Next, we describe the SCF, the sample design, data collected, and disclosure issues. In the fourth 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 experience and discuss future plans in the last section.

General Disclosure Methodology

Before data can be released publicly, either in tables or in a microdata file, the data must be reviewed for potential disclosure risk. Most government agencies and survey vendors have specific disclosure review policies. Several efforts by the statistical community have recently been completed that report on the issues faced in preparing data for public release. The OMB Statistical Policy Working Paper 22, *Report on Statistical Disclosure Limitation Methodology*, (1994) details techniques for controlling disclosure for tabular data and microdata. The Journal of Official Statistics volume, *Confidentiality and Data Access* (1993), co-sponsored by the Panel on Confidentiality and Data Access of the Committee of National Statistics and the Social Science Research Council, provides a recent summary of the issues of confidentiality, methods to use to measure and minimize disclosure risk, and techniques to analyze data subject to such methods. A good review of the policies of most of the U.S. Statistical agencies is given in this volume by Jabine (1993). Many disclosure avoidance efforts focus on tabular data. For the 1989 and 1992 SCF's, we focussed on releasing a microdata file.

There are many techniques that have been used to minimize disclosure for public use microdata files. The priority of these techniques has been to protect the identity of individual respondents. It is also necessary, however, to retain the integrity and usefulness of the original data and to insure that inferences made from the masked data neither contradict, nor be significantly weaker or stronger than those made from the original data. We compare the original data to the masked data, after disclosure adjustments, to measure the effect of these adjustments.

Potential masking procedures include top/bottom coding, adding random noise, swapping, blurring, and blank and impute, as discussed in the OMB Working Paper 22 (1994). Another suggested method is to only release imputed data for ALL variables (Rubin, 1993). Top/bottom coding truncates a variable at a designated level to hide the original, potentially very different, value. Adding random noise is a procedure that systematically adds a random error to the original value, retaining the first and second moments of the masked variable (Fuller, 1993). Data swapping involves exchanging values of a chosen variable between two cases that match on a set of selected variables. With blurring,the data for a group of selected records, say the top 10, are replaced by the average of that group. In the blank and impute method, sensitive variables are identified; values are then deleted and replaced by some sort of imputation method as if they were originally missing. Rounding is also used, usually to simplify the data and to reflect the appropriate level of accuracy. Rounding is also a disclosure avoidance method, by preventing release of the original data.

For the SCF, our disclosure review incorporated several of these techniques. For discrete variables, decisions ranged from collapsing categories to the complete omission of particular variables. For continuous variables, we used the blank and impute method, as well as rounding. The specifics of the review are given after the description of the survey.

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 (SOI) of the Internal Revenue Service (see 1989 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 frame sample is stratified on an estimated wealth index with the higher indices selected at a higher sampling rate. The 1989 sample was additionally complicated by the inclusion of a panel follow-up from 1983, a portion of which is also appropriately included in the 1989 cross section data set (see Heeringa, S. et al. [1994] for a description of the 1989 sample design). The 1992 study consists of only a cross section sample.

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. For the item nonresponse, missing values are multiply imputed using a Gibbs sampling approach as described in Kennickell (1992). For the SCF, the respondent has three options for a given question, he can: 1) give a particular value, 2) refuse to answer, or 3) choose an interval from a range card provided by the interviewer. In the imputation procedure, both refusals and range card values are imputed. The imputations for the range card responses are constrained by the range interval boundaries. 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 dataset that preserves the distributions and relations found in the non-imputed 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 given. The imputation machinery is used in the disclosure avoidance preparation of the public use file as described below.

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.

SCF Data Release Strategy

Although the results described in this paper derive from both the 1989 and 1992 surveys, we detail the steps for release for the 1989. For the release of the 1992 survey, we followed a similar strategy as in 1989. In preparing the data for public release, several factors must be considered. The disclosure avoidance strategy is the main topic of this paper. Other tasks, however, affect the completion of the disclosure review, such as the computation of sampling weights and the imputation for item nonresponse.

Due to persistent demand and pressure from government agencies, university researchers, and the private sector, preliminary datasets for the 1989 SCF were released. The overall strategy was to release more detailed data over time. This was accomplished in several ways -- by the omission of cases, the suppression of variables, and also by truncating continuous variables and collapsing discrete variables. This progressive release pattern reflected our uncertainty at the time of each release of the risk of disclosure. Indeed, the release of more data is still under consideration. For example, some users are interested in geography, a variable completely omitted from the 1989 SCF so far.

In September, 1991, the FRB released the first preliminary public version of the 1989 SCF cross-section dataset. Missing value imputations were the result of the first iteration of the Gibbs sampling model. The dataset was a subset of the complete dataset, both in variables included and in the sample included. To minimize disclosure risk, only a representative part of the area-probability sample interviewed was included (all list cases were suppressed). Many variables were set to missing and all dollar amounts were truncated at the 95th percentile (unweighted). Limiting the data available in this manner was necessary because the detailed disclosure review was not yet finished, nor were the analysis weights finalized. For the user, this dataset was only useful for developing programs or perhaps examining some median behavior.

In March, 1992, the FRB released the second preliminary version of the 1989 SCF cross-section dataset. Again, the dataset was a subset of the complete dataset. For this release, cross section cases from both the area probability and list samples were included. However, 300 cases were omitted completely. About 200 of these cases were chosen to be omitted due to sensitive data; the remaining cases were chosen at

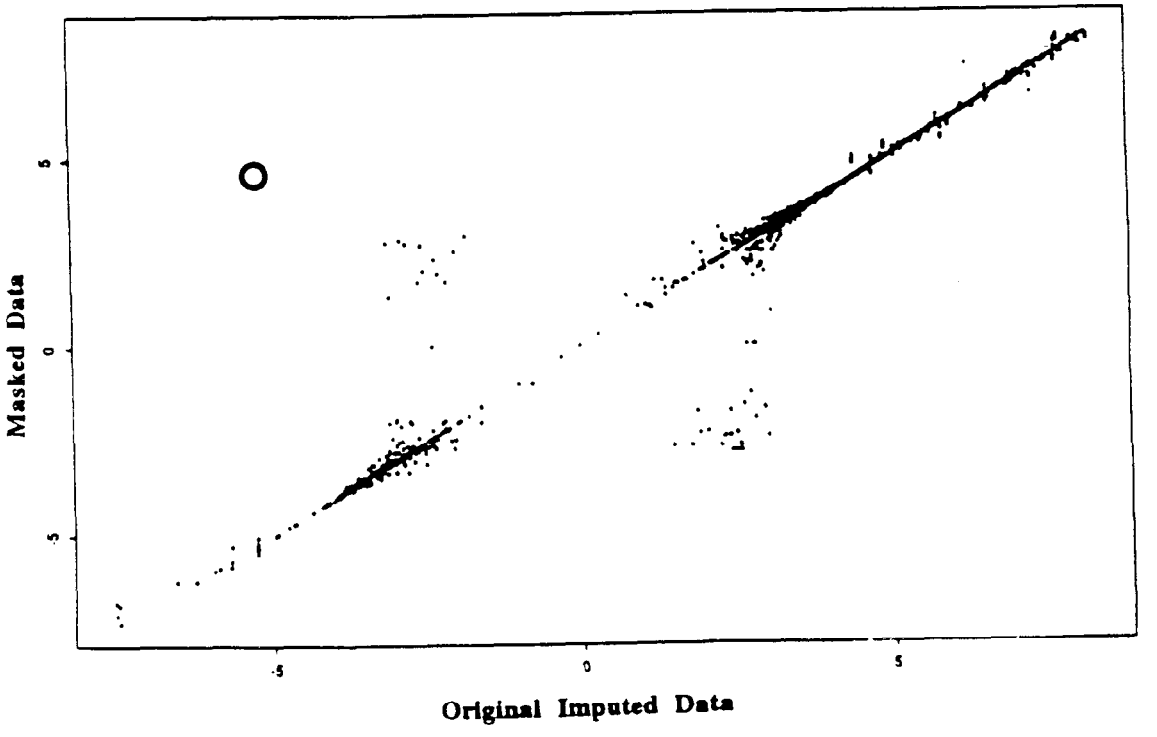
random. The omitted cases included both area probability and list cases. Again, variables that might compromise disclosure were not released, such as geography and make/model of car. The item imputations were the result of the third iteration of the Gibbs sampling model. Additionally, rounding, collapsing and bounding schemes were established that were to be used for the final public release. All dollar amounts were rounded. Large negative values were bounded at -$1,000,000. Negative values for certain income variables were selectively bounded, as well. Rounding was also done for some non-dollar amount variables, e.g. the year cash settlements were received was rounded to a multiple of 5. Many non-dollar amount variables were bounded, such as, the year a loan was taken out, the model year of owned cars, and the number of companies in which stock is owned. Collapsing of cells for discrete variables was done for several variables including race, the type of inheritance, and 1980 occupation and industry codes. Several sets of analysis weights were included, as well as a set of bootstrap replicate weights corresponding to a model-based weight and their respective multiplicity factors from which estimates of sampling variances could be derived. However, any analyses from this data are limited due to the omission of the 300 cases.

In September, 1992, the FRB released the full 1989 SCF cross-section dataset. The dataset included all cross-section cases and all important dollar variables. Again, geography and other sensitive variables were not released. The penultimate step in the disclosure avoidance strategy was to blank and impute selected variables for the 300 cases omitted from the prior release. The imputations for these variables were constrained as if the response had been a range value. As a final precaution, the boundaries of the ranges used for these imputations were different from those used in the survey. Once a value was imputed, the shadow variable was assigned a value that indicated the data was originally missing. Thus, these values are indistinguishable from a true missing response. The procedures for the rounding of dollar amount and other non-discrete variables, for bounding certain variables, and for the collapsing of cells were nearly identical to the procedures used to produce the March, 1992 release. The item imputations were the result of the sixth iteration of the Gibbs sampling model. Other unspecified minor adjustments were made to add uncertainty to the original status of the data.

Analysis of Disclosure Adjustments

Our main concerns with the disclosure avoidance strategy focus on protecting the respondent's identity while preserving the usefulness and integrity of the microdata. In order to measure how effectively we protect the respondent's identity, we would have to develop a measure of each respondent's uniqueness in both the sample and population (Greenberg [1990]).

Figure 1. Scatterplot of log(wealth)



We have not yet attempted this task. By releasing as much of the data as possible, we have retained the maximum amount of usefulness. Of course, users may measure usefulness in a variety of ways. Our evaluation to date of the disclosure adjustments focusses on the integrity of the masked data. Preliminary results are detailed below.

Since an important aspect of the SCF is to be able to measure wealth, we concentrate on wealth estimates derived from the original data compared to the masked data. The scatterplot of log(wealth), original data vs. masked data, is shown in Figure 1. The plot reveals no major differences in the two data sets for any given point. The data points off the 45 degree line around the origin of the graph represent values near zero that changed slightly. The aberrations around zero are exaggerated by the use of the log transform. For example, the circle represents a case where the original imputed wealth differs from the masked wealth by approximately $200,000. This difference arises from blanking and imputing several items in the loan sequence for other vehicles owned by the household. As a result, the total amount still owed on the loan decreases by approximately $200,000 and thus total wealth increases. Many of the other differences are due to rounding alone.

Next, we look at the estimates of the wealth distribution. In Figure 2, a qq plot of the original wealth distribution vs. the masked wealth distribution is shown. In a qq plot, the percentiles of one distribution are graphed against the percentiles of the other. A 45 degree line represents the case where the two distributions are identical. A description of the use of qq plots can be found in Hoaglin et. al. [1985]. If the two distributional shapes are not identical, then the plot will not be a straight line. The plot in Figure 2 conforms to the 45 degree line very well.

Figure 3. Comparison of Net Worth Estimates

ORIGINAL DATA

NET WORTH	Over 65	Self-employed
Mean	$262,528	$639,269
Std. Error	$24,076	$47,825

MASKED DATA

NET WORTH	Over 65	Self-employed
Mean	$260,302	$643,892
Std. Error	$23,720	$48,996

Figure 2. QQ Plot of log(wealth) Distribution

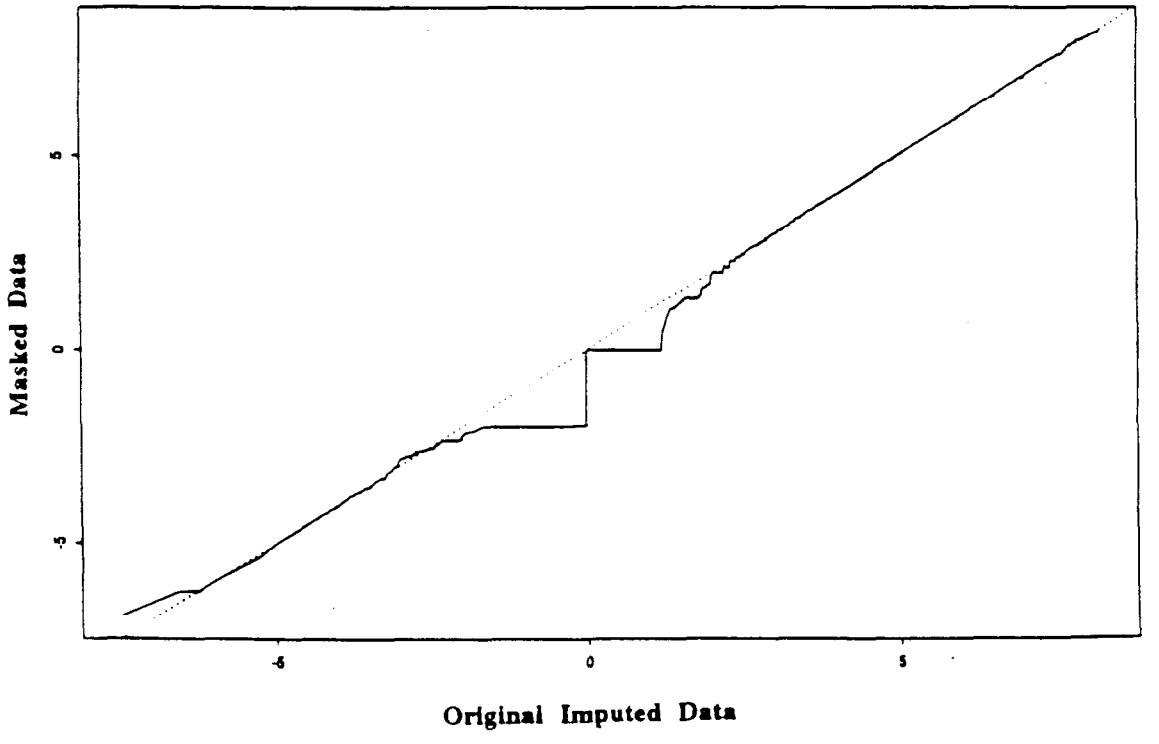
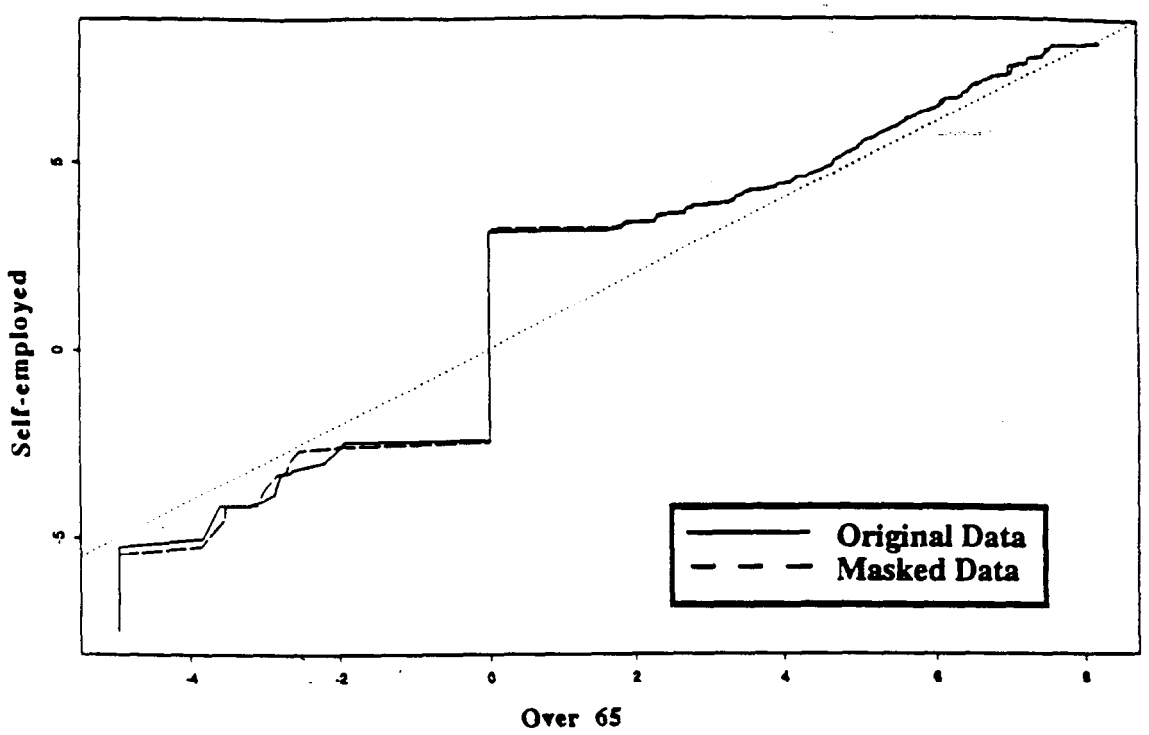


Figure 4. QQ Plot of log(wealth): Self-employed vs Over 65



Finally, we look at estimates for two different subgroups. The estimates used are those of wealth for the self-employed versus those over 65 years of age. In Figure 3, the table shows the mean and the standard errors of these two groups using the original data and the masked data. The standard errors incorporate both the sampling variance and the variance due to imputation. If all else remained the same, and only the blank and impute process were carried out, then the standard errors of the masked data would be greater. This would reflect the additional uncertainty due to the imputation of sensitive values. However, the masked data are also subjected to rounding and bounding, thus complicating the issue. For our example, there are no large differences for these estimates. As an additional check, the wealth distribution of the self-employed is plotted against that of the over 65 group in Figure 4. This qq plot shows the differences of these distributions for the two groups. Both the original and masked data plots are included. Again, there are no major differences in the distributions.

Conclusions and Future Plans

The disclosure strategy that has been developed for the SCF has both strengths and limitations. The blank and impute method used 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. Although portions of this are automated, a significant amount of manual review is necessary. Although more automation would decrease the manual review time, it is unclear whether or not this is an improvement, since it would mean that the data would not be reviewed as closely by human eyes.

By the nature of the imputation process, the integrity of the continuous data is preserved. The results of our preliminary investigation show that the wealth distribution is not affected by the disclosure changes. However, there are other analyses that should be conducted that investigate the effects of the collapsing of the discrete categories and more sophisticated analyses with the continuous variables such as regression modelling. Also, we need to investigate how inferences are affected by the masking process.

We are confident that the disclosure procedures described here reduce to the practical minimum the risk of a respondent being identified using the public

microdata file. However, we plan to investigate the extension of these procedures to bivariate and multivariate concerns. For example, we will not only look at the univariate distribution of wealth, but also investigate the wealth distribution by age category.

Finally, the procedures and improvements discussed here will soon be applied to the 1983-1989 SCF panel data set. Applying these procedures to longitudinal data will present new challenges.

Acknowledgements

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. Many thanks also to Barry Johnson of SOI who helped to implement the disclosure review and had significant input on the direction of the disclosure strategy. The help of Wendy Alvey was invaluable in the preparation of the poster presentation of this paper at the ASA meetings in Toronto, Ontario Canada.

Bibliography

FULLER, W. A., (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, Vol, 9, No. 2, pp. 383-406.

GREENBERG, B., [1990], "Disclosure Avoidance Research at the Census Bureau," *Proceedings of the 1990 Annual Research Conference*, pp. 144- 166.

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.

HOAGLIN, D. C. et. al. [1985]. *Exploring Data Tables, Trends, and Shape*, John Wiley and Sons, INc., pp. 432-442.

JABINE, T. B. (1993), "Statistical Disclosure Limitation Practices of United States Statistical Agencies," *Journal of 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 Survey Research Methods, ASA.*

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

RUBIN, D.B. [1993], "Discussion: Statistical Disclosure Limitation," *Journal of Official Statistics*, Vol, 9, No. 2, pp. 461-468.

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