MOVING FROM DESCRIPTIVE STATISTICS TO INFERENCE

Karen Vlasek O'Conor, Internal Revenue Service
B. K. Atrostic and Robert Gillette, Office of Tax Analysis

1. INTRODUCTION

1.1 Quality Concerns

Over the past few years, the quality revolution -- led by Deming (1986) and Juran (1988) -- has had an enormous impact on the production of Federal statistics. In the U. S., as in Canada, major improvements are being seen in all areas of statistical methodology. This paper describes four major U. S. Federal surveys and the efforts they are making to improve the quality and usability of their public-use files, particularly for tax policy research.

Each program discussed here has addressed quality concerns in a different way. For example, the Internal Revenue Service's Statistics of Income Division (SOI) recently established an interactive process of customers and data suppliers for redesigning its SOI Individual File of sampled income tax returns. Along another vein, the Current Population Survey (CPS), produced by the U. S. Census Bureau, has calculated and released replicate weights, which can be used to produce sampling errors for virtually any type of estimate. In another Census survey, the Survey of Income and Program Participation (SIPP) has created an on-line relational database to be more responsive to users' multiple needs. Finally, the Survey of Consumer Finances (SCF), produced by the U. S. Board of Governors of the Federal Reserve, introduced stochastic relaxation to adjust for item nonresponse.

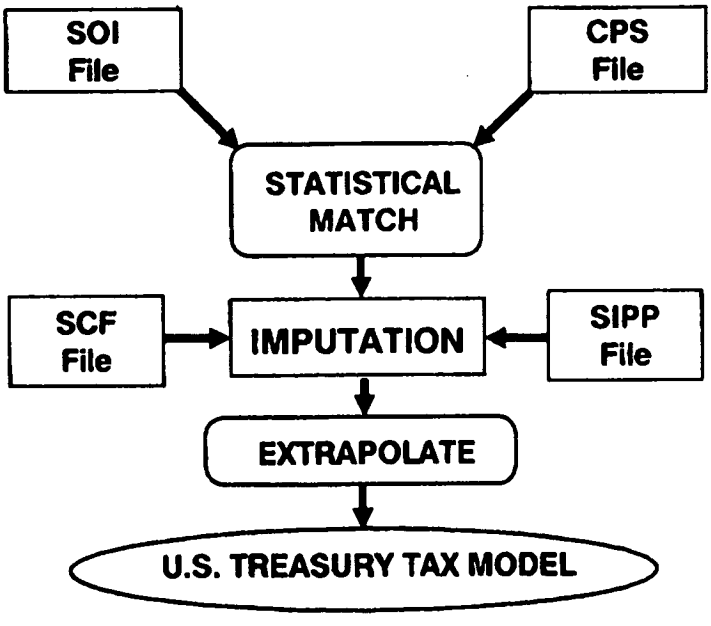
These surveys are just a few examples of work underway in the U. S. Federal sector. They were chosen because they are used by the authors in developing one of the microsimulation models maintained by the U. S. Treasury Department for tax policy research (the U. S. Treasury Tax Model). (See Figure 1.) The following description of that model provides a brief overview of an intense use of public data files and classic insight into the needs of modelers. Later sections describe both the four component surveys' samples and their efforts to improve the quality and usability of their public-use files, so as to achieve better tax policy models.

1.2 Brief Description of the U. S. Treasury Tax Model

In Canada and the United States, tax policy is developed with the aid of microsimulation models that combine the data available from tax administration sources with survey data and other information. The combined database is then adjusted to fit economic and demographic projections for future years, in order to calculate the revenue effects and distributional consequences of proposed tax law changes. The precision of these estimates varies according to the accuracy and availability of the supplemental data. The challenge is to raise the breadth and accuracy of data provided to policymakers, so that they can make more informed decisions. This requires better data, better use of those data, and better understanding by data producers of how users actually make decisions.

The development of the U. S. Treasury Tax Model has two phases: first, a match and imputation phase and, second, an extrapolation phase. The former is actually the matching or linking of statistically "similar" individuals (i.e., statistical matching) -- rather than exact record linkage of identical individuals -- using age and income as determining variables. The match is constrained so that the marginal distributions of the two populations are maintained (Barr and Turner, 1987). Then, the CPS, the Survey of Income and Program Participation (SIPP), and the Survey of Consumer Finances (SCF) are used in the imputation phase. After matching and imputation, the extrapolation process forecasts to achieve the Administration's economic forecast. An overview of the U. S. Treasury Tax Model is given in

Figure 1.–U.S. Department of the Treasury's Tax Model



Cilke and Wyscarver, 1987, and detailed documentation, in Cilke and Wyscarver, 1990.

In Canada, models like the Wolfson model (Wolfson et al., 1989) do more exact and statistical matching than the U. S. Treasury Tax Model, in part because the structure of the Canadian statistical system facilitates such approaches. For one thing, they are able to exchange data among Federal agencies more freely.

Comparability of income measures among the surveys and reliability of their income measures are important considerations in choosing the surveys to use in building the Treasury Tax Model. Because the SOI Individual File contains little information other than (selected) income items, income is one of the key variables in any statistical match or imputation between the SOI Individual File and another survey. While income definitions vary among the CPS, SIPP, and SCF, there is a high degree of overlap.

In fact, much of the new work we wish to focus on in this paper has led to improvements in comparability of income. The sections that follow provide a brief description of the component systems used in the Treasury Tax Model and discussion of some of the new and creative projects that each of the selected surveys is undertaking to improve the usability of its data.

2. STATISTICS OF INCOME INDIVIDUAL FILE

2.1 The SOI Individual File Sample

The U. S. Internal Revenue Service began the Statistics of Income (SOI) Individual File in 1918 and released its first public-use file in 1960. The current SOI Individual File sample size alternates between 80,000 and 120,000 returns. In the current design, data from Forms 1040, 1040A, and 1040EZ are stratified by the larger of total net income or total net loss and the size of business income plus farm receipts. In addition, the strata are based on the presence or absence of Foreign Earned Income (Form 2555), Foreign Tax Credit (Form 1116), Profit or Loss from Business or Profession (Schedule C) and Farm Income and Expenses (Schedule F). Beginning in June of 1991, data will be collected using a new sample design, briefly described in section 2.2.

In both the current and the new designs, returns are selected in each strata using two methods. The first approach uses certain ending digits of the social security number; the second method uses ending digits of random numbers generated from transformations of the social security numbers. This two-stage process was instituted in order to build in an overlap with the Social Security Administration's

Continuous Work History Sample, one of the longest running longitudinal panels in the world, which is used to collect earnings data for social security covered employees. The overlap accounts for approximately 10,000 returns when the sample size is 80,000 (or 20,000 for the 120,000 sample size). It will account for about 20,000 returns in the new design (Smith, 1989).

2.2 Redesign of the Cross-Sectional Sample

2.2.1 Background

As mentioned, the Statistics of Income (SOI) Division of the Internal Revenue Service is redesigning its sample of individual income tax returns to provide better data for modelers to estimate the effects of proposed changes in tax policy. The redesign has three major components:

1. a longitudinal panel of about 83,000 returns, with 1987 as the base year;

2. inclusion of the returns of the dependents of selected families, so that a "tax family" can be constructed; and

3. development of a new design for the annual cross-sectional sample.

All of these components will be selected and processed annually. The first component, the longitudinal panel, will allow the measurement of actual changes in the components of income reported on individual taxpayer's tax returns over time, rather than depending on the repeated cross-section comparisons now available. This will provide a base for quickly drawing specialized panels. The second component, the definition of tax families, will permit the measurement of the incomes of actual families, instead of the synthetic family units previously formed by statistical matches of the SOI Individual File. This will, also, make it easier to reconcile the SOI Individual File with other data sources. The third component, the new design for the annual cross-sectional sample, will address the twin goals of strengthening the sample of income components which are the subject of tax policy and of obtaining better coverage for certain demographic groups (Hostetter, 1990; Czajka and Schirm, 1990).

This redesign has been a very long task. Planning for the effort began two years ago and full implementation is expected in 1991. The longitudinal panel already has data for the years 1987, 1988, and 1989. Definitions are being constructed and revised so that preliminary family data will be produced for our primary cases in late 1991. The system for selecting the new cross-sectional sample is being programmed and will begin selecting the sample in June 1991 (Bates, 1991). The rest of this section will focus on the interactive process which made the redesign effort unique.

2.2.2 Participants and Roles

In 1987 a committee of the data users and producers met to define the objectives of the overall redesign, to lay out the components and needs of the effort, and to build a framework of understanding. The committee consisted of representatives from the Treasury's Office of Tax Analysis (the principal data users), the IRS computer systems area, the IRS service centers (responsible for editing and coding SOI data), independent experts, and the SOI Division (which is charged with managing the survey). By far the most important accomplishment of this committee was the development of a framework of understanding and the agreement on goals. As with many data users and producers, neither had taken the time to understand each others needs. (For more information on how this process came about, see the paper by Susan Hostetter and Karen O'Conor, presented at the 1991 Joint Statistical Meetings in Atlanta.)

The next stage of planning, which began in the fall of 1989, was spent listening to and quantifying ideas for the new sample design of the annual cross-sectional sample and defining the processing of the panel and family data. The research on the former required the most interactions and, to make that happen, a new team was formed. This team was much more focused than the 1987 committee and consisted of three or four data users from the Office of Tax Analysis (OTA), two sampling experts from Mathematica Policy Research (MPR) and a mix of economists and mathematical statisticians from the SOI Division.

2.2.3 Interactive Redesign Process

The Redesign Committee had many issues to decide in order to meet several competing goals. Perhaps the most difficult issue was reaching a definition of returns with complex income structures -- returns most likely to be the focus of proposed tax policy. Complex returns are relatively rare and, therefore, likely to be undersampled. Certain relatively rare filing characteristics, however, are of policy interest and, hence, needed to be identified, so that sufficient numbers were sampled.

After many discussions, an initial consensus was reached. As shown below, the presence of any of 10 income components was taken as a possible indicator of complexity, and 3 filing status indicators that might be undersampled were identified:

Complex income indicators

- capital gain or loss;
- partnership or small corporation income or loss;
- itemized deductions (Schedule A);
- deduction for home mortgage interest;
- social security income;
- pension or annuity income;
- child care credit;
- unemployment compensation;
- alimony income; and
- alternative minimum tax.

Filing status indicators

- aged exemption;
- unmarried head of household; and
- exemption for dependent child living at home or dependent parents.

Next, in order to examine the frequency of occurrence of all possible combinations of related items, MPR constructed indices. The lower positive income strata were the focus of this investigation because a high proportion of returns with income greater than $250,000 were already included in the sample design. A series of tabulations showed that complex returns with income less than $250,000 were dispersed relatively evenly among the remaining indicators.

Two other objectives conflicted with the "complex return" goal. The first was to retain sufficient numbers of noncomplex returns to provide coverage for modelling and descriptive statistics. The second was to maintain as much simplicity as possible, with just two levels of stratification: income and form type. This consideration is important because the series of statistical matches, imputations, and extrapolations brings a high degree of complexity to the construction of the Treasury Tax Model. The modelers also wanted the sample design to be as simple as possible, with just two levels of stratification: income and form type. Test sample designs balanced these conflicting objectives.

The team also determined that it was important to identify noncomplex returns on the basis of income and other characteristics that were unlikely to be the focus of proposed tax law changes. Several iterations were needed to classify returns in a way that both satisfied tax policy needs and maintained the reliability of published SOI estimates.

An initial proposal identified returns that could be undersampled as those where a substantial proportion of the total positive income (75 percent for returns with positive income of $60 to $250,000 and 90 percent for returns with positive income of $0 to $60,000) came from wage and salary or retirement income. This definition turned out to be too narrow, retaining too few returns in the 13 primary categories. An expanded definition added three types of noncomplex returns: returns with alternative minimum tax preference items but zero alternative minimum tax; returns identified as complex only because of substantial tax exempt interest income; and returns which had predominately Schedule C (sole proprietorship) income, and interest and dividend income.

The expanded definition solved one problem only to raise another. The total number of sole proprietors in the sample would have been reduced by approximately 2,300 returns, concentrated in the $0-$30,000 and $30,000-60,000 strata. The resulting sample would not have provided sufficient stability in the estimates of total sole proprietors' net or gross income. Since the majority of the total net or gross income in the population comes from the low income positive strata that would have been undersampled, such a

change would have been unacceptable to our users. The Bureau of Economic Analysis (BEA) uses these data to produce estimates of the National Income and Product Accounts.

As a result, a third condition was added to the first two definitions: a noncomplex return was reclassified as complex if total negative income exceeded 40 percent of its total positive income. By so doing, an estimated 520 sole proprietors returned to the sample in the $0-$60,000 income range (Czajka, 1988; Hostetter, et al., 1990).

2.3 Benefits

Test tabulations were run to observe the impact on the sample statistics of each of the income and form type classification constraints imposed. These provided insight for subsequent decisions. Furthermore, since several different sets of interests had to be met by the final sample, the simulation studies helped provide a more factual basis for compromise. Clearly, the give and take process which followed led to greater understanding of all of the participants' needs. It helped the team explore facts rather than criticize opinions. As the committee worked to achieve an operational definition of complexity, there were differences of opinion on the value of various tabulations because some were quite costly. The high cost of such calculations usually prevents large numbers of them from being calculated. But, all the discussed tabulations were completed; and they were worth every penny because the resulting data kept the decisionmaking on a factual level. The final result is a sample design which satisfies the needs of both the tax modelers (OTA) and the users of descriptive statistics (BEA and IRS).

3. CURRENT POPULATION SURVEY

3.1 The Current Population Sample

Treasury uses CPS and other sources to fill in the demographic and some financial information unavailable from the SOI Individual File. Begun in 1942 when the Survey of Unemployment was transferred to the Bureau of the Census, the CPS is perhaps the most widely known of the component files discussed here. The first CPS public-use file was distributed in 1968.

The CPS, a monthly household survey conducted by the U. S. Census Bureau for the Bureau of Labor Statistics, collects data on a wide range of topics. While initially designed to produce labor force and demographic data, CPS now also collects (primarily through a supplement to the annual March survey) information on topics such as hours worked, occupation, industry, periodic, personal and family income, migration, educational attainment, etc. It is the March CPS that is linked to the tax model at Treasury.

The Current Population Survey covers the civilian noninstitutional population of the United States. The sample size is 60,000 households or about 113,000 persons 16 or older. The CPS is a cluster sample of housing units stratified geographically. The sample is rotated, such that each household is interviewed for four months, not interviewed for eight months, interviewed for another four months, and then retired from the CPS sample (Bureau of the Census, 1978). Of particular interest, here, are the replicate weights recently developed for the CPS, which allow researchers to calculate variances for a wide range of estimates.

3.2 Generalized Variance Functions

In the past, the CPS has used generalized variance functions to produce sampling errors. This had several drawbacks. For example, the Bureau of Labor Statistics (BLS), which is particularly interested in using the CPS to estimate the unemployment rate for each state, had to use this approximation to compute a measure of reliability for each state estimate. In so doing, they encountered two serious problems:

1) the national within-primary-sampling-unit design effect is assumed constant across states, with an adjustment for sample units determined to be out-of-scope of the survey; and

2) the ratio of between-primary-sampling-unit variance to total variance is assumed constant across time (Lent, 1991).

The BLS is now working with a file of replicate weights developed by the Census Bureau to compute variance and correlation estimates by state.

Another problem with the old procedure was that, because of confidentiality, most CPS users could not get the stratification (geographic) information needed to calculate their own variance estimates. They were forced to rely on the generalized variance tables which are only useful for percentages and totals. For very complex estimators, it was often impossible to calculate a variance.

The new generalized replication technique attempts to address both these concerns. This advance has made it possible to provide BLS with a file of replicate weights to compute variance and correlation estimates by state, while maintaining flexibility for meeting smaller users' needs -- all without sacrificing confidentiality. The Census Bureau will continue to calculate the generalized variance function and provide the corresponding tables for those who prefer to use them.

3.3 Overview of Generalized Replication Theory

For the new approach, the CPS is using the generalized replication theory developed by Robert Fay to produce the replicate weights (Fay, 1984 and 1989; Wolter, 1985). This technique is very similar to Balanced Repeated Replication, a procedure for repeatedly constructing estimates from half the sample, such that each stratum has half of its observations included in each estimate. Any observation is in half of the samples.

In generalized replication theory, all of the observations are used in each replicate. Replicate factors are used to weight the contribution of each half sample in the replicate weight. For example, with classical balanced repeated replication, the replicate factors for a half sample in a stratum are either 0 or 2. Using the generalized replicate theory, the replicate factors could be one half or one and a half. Since CPS has only one primary sampling unit per stratum, pseudo strata are then created by collapsing sampling strata. Some of the pseudo strata in the CPS replication system are portions of self-representing strata rather than groups of collapsed strata. The generalized replication theory uses a method to assign factors which differ by the pseudo strata.

Simulation tests have been run to evaluate Fay's generalized variance estimator. They conclude that the technique is useful when variance estimates are needed for both smooth and nonsmooth statistics or when there are very few degrees of freedom available for variance estimation (Judkins, 1990). These are the types of estimates many users of CPS data are making.

3.4 Calculation of the Uncertainty Using Replicate Weights

The resulting replicate weights have made it easy to calculate a measure of uncertainty. Each record has 48 replicate weights. Estimates are constructed by using the replicate weight in the calculation of the estimate, instead of the household weight, which previously had been the only source available. An estimate is calculated for each replicate. The variance is calculated using the standard sum of the differences squared variance formula multiplied by a factor of 4 to account for the use of the whole sample in each estimate.

3.5 Benefits

This method permits the researcher to calculate variances for virtually any estimate and subgroup of the population. Although there are problems with the variance estimates for small geographic subsets (the coefficient of variation of the variance for small states is as high as 50 percent), this is a tremendous improvement over generalized variances.

4. SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP)

4.1 The Survey of Income and Program Participation Sample

The Survey of Income and Program Participation was first collected in 1983. SIPP is a multi-panel longitudinal survey of persons aged 15 and over which measures their economic and demographic characteristics over a period of two and a half years. Core questions cover demographic characteristics, labor force participation, program participation,

amounts and types of earned and unearned income received, asset ownership, private health insurance, and pension receipt. The sample size for SIPP has varied between about 11,500 and 37,000 households, due to budget constraints and panel overlap. From February to July there are three panels in the survey; the rest of the year there are two. SIPP data are collected by the Census Bureau (Census, SIPP User's Guide, 1987). The new relational database, which Census is implementing for SIPP users, will vastly improve quality and usability.

4.2 Early SIPP Public Use Files

The Survey of Income and Program Participation represented a major departure from the Census Bureau's approach to current surveys. SIPP was much more complex and ambitious, and resulting data offered great promise to demographers and researchers in the area of tax policy and transfer issues. Unfortunately, the producers of the SIPP initially received very negative comments from file users. The record layout was confusing. There were redundant fields (e.g., age original, age edited, and age imputed). The location of variables shifted among files. "Zero" had different meanings, such as missing, not applicable, and zero. The documentation of the editing and imputation procedures was not available and the release of clean files was not timely (ADPU, 1989).

Drawing on the experience of the user community, the Census Bureau has redesigned the SIPP public-use files. Data from the 1990 panel will be released in a person-month format -- one record for each month a person is in sample. This provides a one-to-one relationship between the person interviewed and the timing of the data reported. The redesign also solves one of the most troubling problems of the public-use files, where a single record recorded data representing several different time spans. In addition, the 1990 data have been reorganized. Redundant data have been eliminated; a single record layout has been developed for all waves of core data; and new recodes have been added to aid the user in working with some of the more complicated concepts in the survey.

With many user concerns ameliorated, Census proceeded to explore other approaches to improve the quality and usefulness of SIPP. A large scale research project at the University of Wisconsin, funded by the National Science Foundation, explored ways of making SIPP data more accessible to the academic community (David, 1989). That project developed a relational database using the 1984 SIPP panel. The database is now supported by and available through the Census Bureau. The Census Bureau is also in the process of adding data from the 1985 panel to the database.

4.3 SIPP Relational Database

A relational database permits the user to develop his or her own aggregate tabulations to meet individual needs. The current system developed at Census consists of the 1984 SIPP panel data, which is loaded into INGRES, a relational database software package. The relational database has 210 tables. These tables can be broken down into three types: wave, stacked and other. The wave tables have long records with many variables for one wave of data. In order to reduce the amount of storage required for the table, the number of rows or records in the tables is restricted. There are separate tables for persons, households, and families. There are also separate tables for different kinds of people (e.g., those earning wages and salaries). The rows consist of the records for persons with valid data for the table. The columns are variables related to the subject matter of the table.

The stacked tables have short records with a few variables for all responding waves. These tables were specifically designed for use when the number of members in the sample is smaller -- e.g., persons with Medicare or persons with self-employment income.

The remaining tables provide additional data, such as:

- link tables, which link people, families and households;
- miscellaneous tables of constants, couples, marital status, retention, and recipiency change; and
- utility tables, with survey, and reference dates.

These other tables are used in links to the stack

and wave tables to find out information, such as how many people from a household get Medicare. The relational nature of the database permits users to link several files to get the data they want.

4.4 Benefits

The increased flexibility and ease of access created by the relational database has many benefits for SIPP users. It should be noted the SIPP has not been used in constructing previous Treasury Tax Models because access to the microdata was too complex before the relational database was developed. Instead, SIPP had been used to supplement analyses of tax provisions when the SIPP contained information unavailable on the CPS or other sources. Even with the relational database, SIPP is still more likely to be used for imputations than for statistical matching, because of the difference in sample size between SIPP and SOI Individual File.

5. SURVEY OF CONSUMER FINANCES (SCF)

5.1 The Survey of Consumer Finances Sample

The last component of the Treasury Tax Model that has undergone some important changes in order to better meet user needs is the Survey of Consumer Finances. The Federal Reserve Board has conducted the Survey of Consumer Finances on a triennial basis since 1983. The primary purpose of the SCF is to gather comprehensive data on assets, liabilities, pensions and income, together with descriptive data on employment, marital history, family structure, health, and other demographic variables. The 1989 SCF (like the 1983 SCF, but not the 1986) actually consists of two related surveys: a household survey and a survey of pension providers. It is the household survey what will be discussed here because of the work Arthur Kennickell, of the Board of Governors of the Federal Reserve, is doing in multiple imputation.

The household sample was selected using a dual frame, composed of a list and an area frame. The stratification of the list frame was a wealth proxy index and that of the area frame was geographic. For the 1989 SCF, the number of respondents from the list frame was approximately 870 households; from the area frame, 1,130 households. This household sample is important because of the high number of wealthy households responding -- 100 of the respondents each have an estimated wealth of $10 - $250 million.

5.2 Household Nonresponse

The SCF is the only source of survey data on wealth that has a sufficiently large sample in the upper income strata to permit separate analyses of them. (The SIPP also surveys wealth, periodically, but its design does not provide enough observations for separate analyses of the highest strata of the wealth distribution.) High income taxpayers are very important for modeling changes in tax policy, as well as being of considerable interest to other researchers. In fact, taxpayers in the top 10 percent of the income distribution pay 50 percent of the taxes; those in the top 1 percent pay 26 percent.

Despite the importance of these data, it is generally felt that response rates decline as wealth rises. As Figure 2 illustrates, the SCF response rates corroborate this and point out the difficult problem of collecting data from such "large" taxpayers.

The high income portion of the SCF sample is a population subgroup whose income and wealth are very hard to measure, due, in part, to the household and item nonresponse. A study focusing exclusively on wealthy households as identified through a wealth index would not meet the expected 70 percent response rate required for approval by the U. S. Office of Management and Budget. (Federal surveys must be approved by the Office of Management and Budget before data can be collected.) However, these households are crucial to tax modeling. For the purposes of comparison and to make national estimates, an area frame of primarily lower income households supplemented the list sample. It raised the overall response rate to about 70 percent. Still, all surveys have subgroups with poor response rates. One of the strengths of the SCF is that there are strong variables from the sampling frame to help in modeling the missing data. Hence, it was this area which was the focus of new efforts to improve data quality for the users.

5.3 Item Nonresponse Adjustment

Item nonresponse is also a serious issue in the SCF. Although not nearly as large as the unit nonresponse problem, the adjusted gross income

Figure 2.--Response Rates from a List Frame with a High Number of Wealthy Respondents

Wealth Proxy Index	Response Rate	Number of Respondents
$0 < $100,000	48.4%	45
$100,000 < $500,000	43.3%	116
$500,000 < $1 million	39.6%	158
$1 million < $2.5 million	39.4%	232
$2.5 million < $10 million	30.6%	215
$10 million < $250 million	20.1%	100
Total	34.1%	866

nonresponse problem, the adjusted gross income variable has an item nonresponse rate of 28.6 percent. (See Figure 3.) To deal with this problem, the SCF uses sophisticated regression techniques combined with multiple imputation to adjust for item nonresponse. The method, called stochastic relaxation, is also known as Gibbs sampling and expectations maximization algorithm (Rubin, 1987a, 1987b and 1990b; Geman and German, 1984).

The theory for this approach requires all the variables in the imputation to be continuous and that all the variables can be transformed to normal. These assumptions are approximately satisfied. By using an iterative procedure, a randomized regression model is employed to estimate all the missing values for a particular item -- for example, adjusted gross income (AGI). Once all the records with missing adjusted gross income have had a value imputed, the process moves on to the next variable. The process moves through all the variables in turn, then the cycle starts again.

The actual imputation is done using a randomized regression model. There are four major steps in the imputation of each variable, as follows:

1. A set of conditioning variables are determined. For AGI there are 300 conditioning variables.

2. The variance covariance matrix is generated using cases with at least 75 percent of the conditioning variables responding.

3. For each record with nonresponse in the variable being imputed, the following substeps a and b are carried out:

   a. A covariance matrix is constructed from the matrix in Step 2, based on the responding conditioning variables for that record; and

   b. Five randomized, independent replicates of the missing variable are imputed, using the regression defined by the covariance matrix for that record.

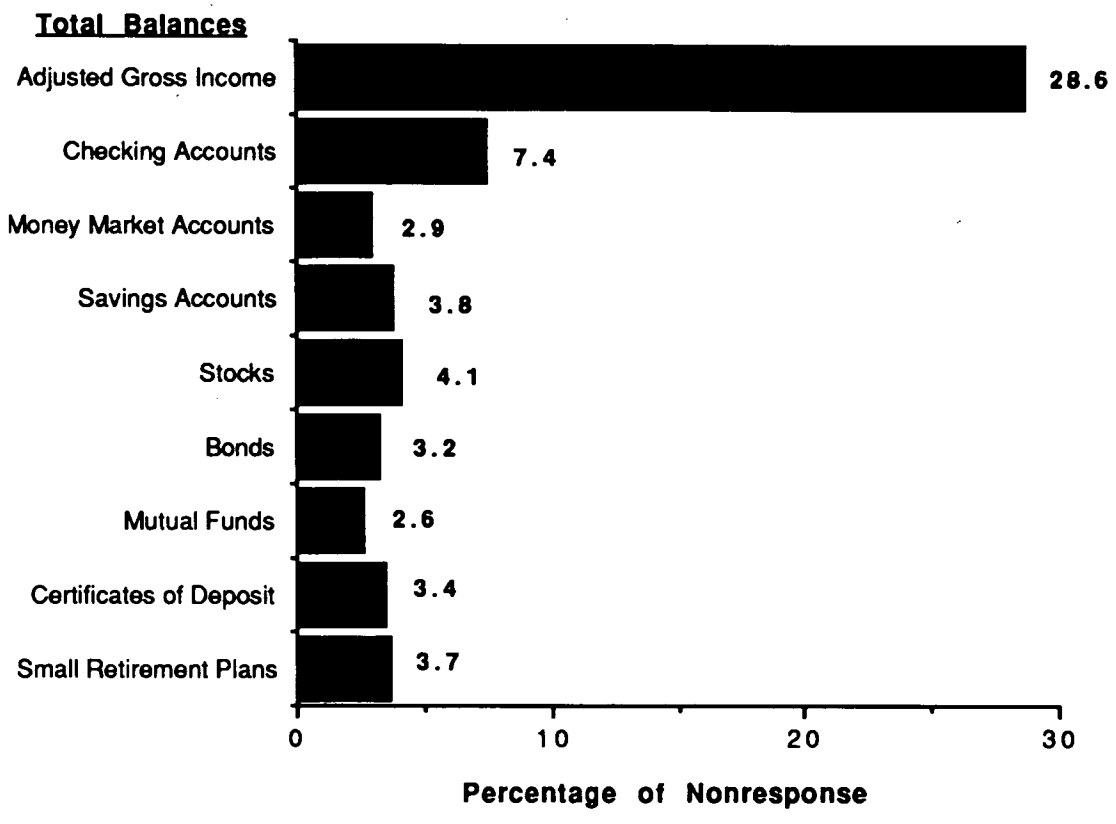
4. The imputed value is used as a conditioning variable in the variance covariance matrix of the next variable.

Once all the variables have been imputed, Step 1 is repeated. In this way, the procedure iterates toward the maximum likelihood estimate of the variance covariance matrix (Kennickell, 1991).

5.4 Benefits

The benefits of stochastic relaxation are that inferences made on the imputed database (using a multiple imputation technique) have the same validity, because the distribution of the imputed data is the same as the underlying distribution. Also, the imputation takes advantage of the relationship between the variables as opposed to independently imputing each variable.

Figure 3.--Nonresponse Rates for the Survey of Consumer Finances



6. CONCLUSION

Data quality has traditionally been measured by looking at issues which concern data producers. A more complete measure of data quality would look at both sides of the question: conformance to standards and fitness for use.

The vast majority of the resources allotted to data collection are spent for the production of descriptive statistics. Increasingly, many users are also interested in the microdata (i.e., public-use files). As the sophistication of data users and their technological capabilities has increased in recent decades, the production and improvement of more flexible products such as public-use files has not kept pace. That is changing now. More importantly these improvements are coming about because data producers are listening to their customers.

As a result of growing quality consciousness in the U. S. Federal statistical agencies, the community of data producers recognized the importance of the user perspective. These four surveys exemplify a wide range of efforts to improve public access to data and the resultant usefulness of that information. The SOI Individual File redesign paid increased attention to user needs, through greater user involvement in the redesign process. The redesigned SOI Individual File should provide more reliable basic data than the previous design. It also should provide sufficient numbers of sample cases for complex returns and other classes of returns that are likely to require separate analysis for tax policy.

Attention to user needs has made the CPS and SCF important data sources for modeling the components of economic income, and for estimating the effects of tax policy. Similar attention in the SIPP program increases the likelihood that its rich array of data can now be used, as well. All of these non-tax sources of data are needed because the basic source of tax information, the SOI Individual File, contains only the variables available on income tax returns. Many policy questions, on the other hand, require additional information; for example, calculating the effects of a

tax change on a family -- rather than on a return -- basis, or modeling the effect of including an income source currently not reported on tax returns. Increased ease of use, enhanced design, and improved public use data sets expand the range of information available for tax policy analysis.

NOTES AND REFERENCES

Association of Public Data Users (1989),SIPP Supplement to the APDU Newsletter, Vol. 1 supplements 1-6 and Vol. 2 supplements 1-2, NJ: Princeton University Computing Center.

Barr, Richard and Turner, Scott J.(1978), "A New Linear Programming Approach To Micro Data File Merging," *1978 Compendium of Tax Research*, Department of the Treasury, pp. 131-149.

Bates, Jeffrey (1991), "Creating a Database for Longitudinal Analysis of Families," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Cilke, James M. and Wyscarver, Roy A. (1987), "The Treasury Individual Income Tax Simulation Model," *Compendium of Tax Research 1987*, Washington, D. C.: Department of the Treasury, Office of Tax Analysis.

Cilke, James M. and Wyscarver, Roy A. (1990), "The Treasury Individual Income Tax Simulation Model," Washington, D. C.: Department of the Treasury, Office of Tax Analysis.

Czajka, John and Schirm, Allen (1990), "Overlapping Membership in Annual Samples of Individual Tax Returns," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Czajka, John (1988), "Development of a New Income Classifier for a Sample of Individual Tax Returns," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

David, Martin and Robbin, Alice (1989), "Database Design for Large-Scale,Complex Data," U.S. Bureau of the Census, SIPP Working Paper 8923.

Deming, W. Edwards (1986), *Out of the Crisis*, Massachusetts Institute of Technology, Center for Advanced Engineering Study: Cambridge, MA.

Fay, Robert E.(1984), "Some Properties of Estimates of Variance Based on Replication Methods," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Fay, Robert E.(1989), "Theory and Application of Replicate Weighting for Variance Calculations," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Geman, Stuart and Geman, Donald (1984), "Stochastic Relaxation, Gibbs Distribution, and the Baysean Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Hostetter, Susan ; O'Conor, Karen; Czajka, John; and Schirm, Allen (1990), "Choosing the Appropriate Income Classifier for Economic Tax Modeling," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Hostetter, Susan and O'Conor, Karen (1991), "Satisfying the Needs of Income Tax Policy Modelers While Preserving the Reliability of Descriptive Statistics," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Judkins, David R. (1990), "Fay's Method for Variance Estimation," *Journal of Official Statistics*, pp 223-239.

Juran, Joseph M. (1988), *Juran on Planning Quality*, The Free Press: New York, NY.

Lent, Janice (1991), "Variance Estimation for Current Population Survey State Labor Force Esti-

mates," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Rubin, Don (1987a), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons: New York, NY.

Rubin, Don (1987b), "EM and Beyond," European Meeting of Biometric Association.

Rubin, Don (1990),"Discussion" of Concurrent Session VIII-A: Imputation, Proceedings of Annual Census Research Conference VI.

Schirm, Allen and Czajka, John (1990), "Intertemporal Stability in Total Income and the Overlap in Annual Samples of Tax Returns," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Smith, Cres (1989), "Social Security Administration Continuous Work History Sample", *Social Security Bulletin.* (The Continuous Work History Sample was begun by the Social Security Administration in the 1930's. Prior to 1980 the CWHS released public-use files; however, due to concerns about the possibility of disclosure, the Social Security Administration no longer releases the data.)

United States Bureau of the Census (1978), "The Current Population Survey: Design and Methodology," Technical Paper 40, p. 2.

United States Bureau of the Census (1987), "Survey of Income and Program Participation: Users' Guide," Chapter 1.

Wolfson, Michael; Gribble, Stephen; Bordt, Michael; Murphy, Brian; and Rowe, Geoff (1989), "The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration," *Survey of Current Business.*

Wolter, Kirk M. (1985), *Introduction to Variance Estimation*, Springer Series in Statistics, New York: Springer-Verlag Inc.