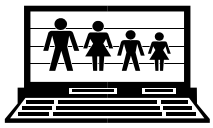
SPSD/M 

Database Creation Guide

The Database Creation Guide describes the techniques used in constructing the SPSD.

Statistics Canada Statistique Canada

Canada

Table of Contents

Summary

This guide describes the construction of the database provided with the **Social Policy Simulation Database/Model (SPSD/M)**. This database was explicitly designed to support the analysis of personal income and sales tax and income transfer policies. These policies increasingly require integrated analysis that cuts across traditional jurisdictional and program lines. The SPSD/M database was constructed to support micro-analytic modelling by combining individual administrative data from personal income tax returns and Employment Insurance claimant histories with survey data on family incomes and expenditure patterns.

Additional aggregate administrative data has been used in the creation of both the database and model portions of the SPSD/M. Input-output data were also applied in modelling sales taxes and duties as they relate to personal consumption. The techniques used to create the database and avoid confidential data disclosure include various forms of categorical matching and stochastic imputation.

Introduction

In Canada, a small number of federal government ministries have had a virtual monopoly on the ability to do detailed analyses of the impacts of tax and transfer policy changes. There is keen public interest in which groups of families or individuals will gain or lose on account of a particular policy proposal. Interested parties outside the particular ministries (including other federal ministries and provincial governments) have had no way to assess the published estimates of such distributional impacts of policy proposals, no way to explore the impacts in greater detail, and no way to develop comparable figures for their own proposals. This situation is unlike that in the United States where various independent agencies such as the Urban Institute and Mathematica Policy Inc. have sophisticated microsimulation capabilities. It is also unlike the situation in the area of macro-economic policy where many agencies in both countries regularly provide independent analyses and forecasts.

With the **Social Policy Simulation Database/Model (SPSD/M)** from Statistics Canada, anyone, with sufficient effort, can perform microsimulation impact analyses of tax and transfer program changes on their own personal computer (PC). The level of sophistication approaches, and in some cases exceeds, that of federal government ministries.

The SPSD/M represents a different philosophy from the traditional products of a national statistical agency - typically print publications with many tables of numbers. The SPSD/M project started with the objective of making available to the public a capacity for performing policy relevant tax/transfer program analysis. Given this objective, a specially designed database has been constructed along with a retrieval and analytical software package.

The database was explicitly tailored to the software and analytical applications,

unlike the more common situation where the analysis is constrained by the data already available. As further development constraints, the database had to be non-confidential within the meaning of the Statistics Act, and the database and software package had to be portable across a range of computing environments, especially PCs. These constraints are necessary for the SPSD/M to meet the objective of broad public accessibility.

Policy relevant analysis in the case of tax and transfer programs can only be conducted effectively with microsimulation. To estimate the likely impact of a change in income tax exemptions for different types of families by income range, for example, the federal Ministry of Finance employs a microsimulation model that recomputes income tax liabilities for a sample of about 400,000 taxpayers, based on their actual tax returns for a recent year. Essentially, the software steps through a representative sample of tax returns one at a time, and for each of these returns calculates tax under some alternative policy scenario. Similarly, the Ministry of Employment and Immigration has their own microsimulation model for the Employment Insurance system based on a sample of their own internal administrative data files.

In virtually all cases in Canada, these are only (but not necessarily simply) accounting calculations; no behavioral response is assumed. The SPSD/M is similar in this regard - the modelling software only does accounting calculations.

A significant and unique aspect of the SPSD/M is the provision of an integrated framework for tax/transfer analysis. At present, there are two federal ministries with major microsimulation capabilities: Finance for personal income tax, Human Resources Development for Employment Insurance, transfers related to children and OAS. Historically, these models have developed independently and are substantially non-overlapping in their capabilities.

The lack of integration in these departmental policy models is proving to be an increasing problem in the Canadian policy context as more attention is focused on the interfaces between major groups of programs and the often complex interactions among them. (We include as "programs" the various tax expenditure provisions in the income tax system.)

For example, there are concerns about how the unemployed move between Employment Insurance and welfare, and about the interaction between income tax provisions and transfer programs directed toward children. The SPSD/M addresses this problem by providing in one package, integrated at the microdata level, sufficient data to model personal income tax, Employment Insurance, major transfer programs (except earnings related pensions and welfare), and commodity taxes.

A key challenge in the construction of the database portion of the SPSD/M has thus been to assemble and merge a number of microdata sets. It is essential that most of the richness of detail in each of the donor microdata sets is preserved. The merger of these microdata sets also has to result in joint or merged microdata records --

each one of which is realistic or plausible, even if it turns out to be synthetic and artificial. On the other hand, the resulting microdata set has to comply with the Statistics Act and not allow any real individuals to be identified.

This guide describes the way in which the Social Policy Simulation Database has been constructed. We start with the general objectives of the SPSD and the character of the source data. Then, in the main part of the paper, the many steps in the assembly of the SPSD are described.

We strongly recommend that users review this manual in some depth. The validity of analysis conducted with the SPSD/M will be dependent on the user's understanding of the microdata on which the model is based.

In this guide, base year is the year on which all the databases used to build SPSD are based.

Objectives, Data Sources, and Techniques

OBJECTIVES

In developing the SPSD, every attempt has been made to maintain the variety and utility of the original source data while ensuring its non-confidentiality so that the resultant database and model can be publicly released. Four central objectives guided the selection of techniques, data sources and variables, and process:

- **Public Accessibility/Non-Confidentiality**

  The first objective has been to ensure that no actual individual represented in any of the databases could be identified through either explicit or residual disclosure. This is a prerequisite for the SPSD/M to be released to the public. Also related to public accessibility is the requirement that the database and model be capable of executing on a moderately priced PC.

- **Aggregate and Distributional Accuracy**

  The SPSD/M has been designed to reproduce as closely as possible "known" aggregates such as total number of Employment Insurance beneficiaries. Furthermore, particular efforts have been made to represent accurately the distribution of aggregates across several classifications key to public policy analysis in Canada such as province, age, income, family type, and sex. Finally, it is important that at the microdata level, the shapes of the distributions of specific variables are well represented.

- **Completeness and Detail of Data**

  The selection and aggregation of variables from the main data sources has attempted to foresee likely policy options as well as serve the needs of the current tax/transfer models. For example childcare costs are included in the

database yet are not currently used in any of the models.

- **Micro-Record Consistency**

  For confidentiality reasons, stochastic rather than exact matching techniques have been used. In turn, it has been necessary to give consideration to avoiding the creation of unrealistic individual microdata records - for example an elderly childless couple with a full child care expense tax deduction.

These central objectives are highly interdependent and compromises have been made among them. The process of making trade-offs included consultation with an *ad hoc* working group composed of staff from four federal ministries with an interest in the resulting SPSD/M as well as previous experience with their own microsimulation models. The final product thus represents a compromise among methodological, informational, technological, departmental and public policy concerns.

In addition to these objectives, one further objective can be added from hindsight. In the field of National Accounting, there has been a growing strand of concern about the lack of microdata foundations for macro-economic aggregates, for example in writings by the Ruggles. While this was not the original intention, it turns out that the SPSD can also be seen as the micro foundation for the Canadian household sector, as described explicitly in Adler and Wolfson (1987).

DATA SOURCES

The SPSD has been constructed from five major sources of microdata.

- **The Survey of Consumer Finances (SCF):** Statistics Canada's main source of data on the distribution of income amongst individuals and families served as the host dataset. It is rich in data on family structure and income sources; but it lacks detailed information on unemployment history, tax deductions and consumer expenditures.

- Personal income tax return data: the three percent sample of personal income tax (T1) returns used as the basis of Revenue Canada's annual **Taxation Statistics** (Green Book) publication;

- Employment Insurance (EI) claim histories: a 10% sample of histories from Human Resources Development administrative system; and

- The **Survey of Household Spending** (SHS): Statistics Canada's periodic survey of very detailed data on Canadian income and expenditure patterns at the household level including information on net changes in assets and liabilities (annual savings).

- **Survey of Labour and Income Dynamics** (SLID): a longitudinal survey on household labour market experience and incomes.

These original data sources from which the SPSD has been constructed are confidential. Until now, data from these microdata sets have been disseminated either as public-use samples in which some records and a fair number of variables are suppressed (SCF and SHS), or in the form of summary tables (Taxation Statistics), or not at all (EI claim histories).

For purposes of the Social Policy Simulation Database (SPSD), these four data sources have been transformed into a single non-confidential public use microdata set. In addition, these microdata have been augmented by reference to various aggregate data which served mainly to provide benchmarks or control totals. These aggregate data were drawn from Canada Assistance Plan (welfare) administrative reports, Statistics Canada's census (the closest to the base year), Vital Statistics, and HRDC summary reports.

TECHNIQUES

The joining together of the four initial microdatasets, addition of new information and the replacement or adjustment of biased measures were largely dependent on five techniques employed extensively in the creation of the SPSD: conversion, regression raking, stochastic imputation, micro-record aggregation, and categorical matching.

- **Conversion** is a method for adjusting microdata to deal with the problem of item non-response. It involves identifying appropriate individuals who reported no payment from a particular program (i.e., EI benefits) and imputing a payment to them (i.e., they are 'converted' to respondents).

- **Regression Raking** refers to a technique for reduction of bias by forcing agreement between data and known control totals. The household weights are constrained (raked) to agree with known, individual level, control totals. The control totals employed are population by age, sex and province, number of households by province, and census family status (i.e., child, spouse, single parent or non-family person) by sex and province.

- **Stochastic Imputation** is the generation of synthetic data values for individuals on a host data set by randomly drawing from distributions or density functions derived from a source data set.

- **Micro Record Aggregation** is the process of creating synthetic micro-records by clustering similar records. For example, micro records from high-income taxpayers are clustered into groups of five according to policy-relevant criteria. Within each group of five, values of relevant variables (e.g. capital gains) are (weighted) averaged to create non-identifiable records which resemble microdata but are actually synthetic.

- **Categorical Matching** involves first classifying records on both a host and donor dataset based upon policy-relevant criteria common to both datasets (e.g., dwelling tenure, employment status, income class). The information on donor

records thus classified may then be attributed to records with similar characteristics on the host dataset without the possibility of adding to their identifiability.

Figure 1 provides an overview of the SPSD creation process. The ellipses represent data files (e.g., the SCF, the Green Book) and the rectangles represent processes. The next section of this guide describes each step in the construction of the SPSD, as shown in Figure 1.

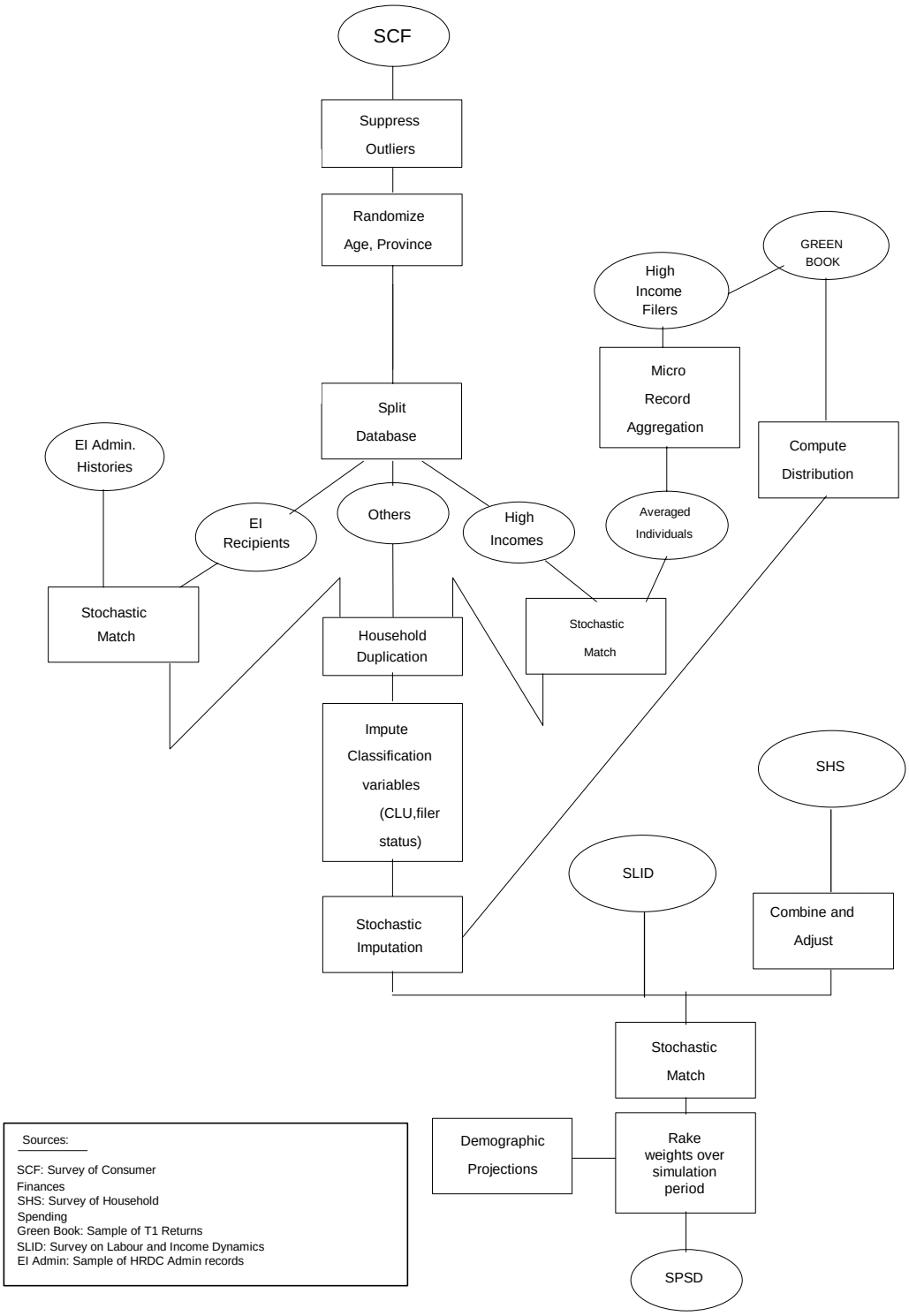


Figure 1: SPSD Database Creation Process

The Host Data

The target or "host" dataset is derived from Statistics Canada **Survey of Consumer Finances** (SCF) in the base year. SCF is an annual survey administered to selected households drawn from the survey frame of the Labour Force Survey (LFS). Four different forms are collected from each sampled household. The Household Record Docket contains demographic information on each individual in the household, as well as family structure information. The LFS form contains information on the labour force status for individuals aged 15 and over in the household. The SCF form has the income, by source, for each member of the household aged 15 and over. The original weights in SCF are producing 6% more wages and salaries than national account value. To solve the issue we reweighted SCF.

The Household Income Facilities and Equipment (HIFE) form details the characteristics of the dwelling, and certain kinds of equipment contained in it. Associated with each household in the sample is a Record Docket and a HIFE form, and associated with each individual in the household aged 15 and over is an LFS form and an SCF form. Because of the great wealth of already linked information that results, this combined hierarchical database forms the starting point for the SPSD creation process.

It may be noted that even though these diverse data are fully integrated at the microdata level in the early production phases of the survey, the public so far has never had access to this rich multivariate information. The survey results emanate from Statistics Canada as distinct public use sample tapes or print publications on individual incomes, economic family incomes, census family incomes, HIFE and the labour force survey. This traditional and fragmented view of the utility of microdata sets is one that is overcome by the SPSD. We have provided a fully hierarchical database including individuals, census families, economic families and households.

The information from the EI, Greenbook and SHS files was then "added" to the SCF. In order to exploit the full variety of this information being imputed from other sources, many original SCF records were cloned or duplicated. For example, records representing unemployed individuals were duplicated until the number conformed to the sample size of the EI file. Records representing high-income filers (those with an income of over $112,500) were duplicated to correspond to the number of high-income records derived via micro-record aggregation from the Revenue Canada sample. To maintain the family structure and overall sum of weights, the records of all other persons in households containing either unemployed or high income individuals were similarly duplicated. The weight assigned to a record was reduced to account for the number of times it was duplicated.

ORIGINAL SCF REWEIGHTING

SCF wages and salaries produce total wages up to 6% higher than national account. A comparison with T4 file shows an over representation of population in the median group and an under-representation of the low wages population. The T4 file is used to recalibrate the wage distribution.

The count of population within six income classes, by province, from T4 file is used as target value in the recalibration of SCF. The population used to define those classes is all the wage earners in T4 file with income higher or equal to $1,500. The classes correspond to 25, 50, 65 and 75% of population in T4 with wages and salaries. The last two classes depend on the province: in Newfoundland and PEI it is 95 and 100%; in Quebec and Ontario it is 99 and 100%; in all the other provinces it is 98 and 100%.

Because most of the people with self-employment income also have small amounts of wages within a year, the wage distribution reweighting worsen the distribution of farm and non-farm self-employment incomes. The solution was to calibrate the distribution of farm and non-farm self-employment income in SCF, based on T1 distribution. For non-farm self employment income, two classes by province are used corresponding to 50% of the population with self-employment income larger or equal to $500. Negative values were excluded because there are not enough observations in SCF to support calibration.

The same definition of income classes applies to farm self-employment income. Because ther are not enough observations in Newfoundland, Prince-Edward-Island, Nova-Scotia, New-Brunswick and British-Columbia, those provinces were excluded from calibration. They count for less than 6% of all people with farm self-employment income larger than $500.

RANDOMIZATION

A guarantee of the non-confidentiality of the constructed database (SPSD) is provided if each input microdataset is itself non-confidential, and if data "merging" does not involve exact matching. This is the strategy that has been adopted, and begins with screening the SCF file.

Public release versions of the host (SCF) data are pre-screened for potentially sensitive cases. For example, households with more than nine members have the province of residence blanked out. In the SPSD, these household entries are not suppressed but are randomized for geographic location. The same treatment is applied to census families with more than four EI recipients or more than 6 earners.

Similarly, the geographical location of other unusual household types are changed by randomly reassigning their province and urban size class codes. Unusual household types are defined as households containing more than eight individuals, more than 2 census families, more than one economic family, or individuals with special income or tax characteristics (e.g. females with income above $80,000, or male or female with income below $150,000 and income tax greater than $150,000).

Further protection against release of identifiable households is provided by age-sex and regional randomization, or "controlled blurring". At the same time, if this "blurring" is suitably structured, it need not adversely affect the utility of the database from the point of view of the policy simulations for which it has been designed.

In addition to the above measures, certain SCF recodes were performed. These involved, for example, merging certain geographic areas (e.g., Brandon with Winnipeg) or recoding as unknown the occupation codes for spouses of high-income individuals.

CONVERSION

External evidence suggests that under-reporting of EI or welfare benefits and of CPP/QPP payments are likely to be item non-response. The problem is not that the recipients are under-represented in the sample, rather they forget or neglect to report the payments.

The conversion technique attempts to deal with the problem of item non-response by identifying appropriate individuals who reported no payment and imputing a payment to them (i.e., they are 'converted' to respondents). This step of database adjustment is undertaken, like regression raking (see next section), to ensure that the database balances to known control controls for the items raked. This conversion is not undertaken in the preparation of the SPSD microdata but during the execution of the SPSM, as noted below. We describe it here because, like the other aspects of database creation, it affects the nature of the SPSD microdata and may be of interest in interpreting the results of analysis.

There may be a pattern to occurrences of non-response. For example, EI non-respondents may include those whose claims ended in the first few weeks of the calendar year. In that case, any attempt to identify actual non-respondents should include an examination of individuals who may have had a few weeks of unemployment in the year.

In the absence of auxiliary information on non-response patterns, an attempt to identify and convert actual non- respondents might introduce distortions on the database. As in the example above, non-respondents may have quite different characteristics from respondents.

The conversion strategy that has been adopted is designed to introduce as little distortion as possible. The first step involves computing a logistic regression on response status (i.e., respondent/non-respondent) in order to assign a response probability to each individual. In effect, this permits ranking non-respondents in terms of similarity to respondents.

The identification of those who will be converted has been carried out by Rank Method. The Rank Method ensures that control totals are satisfied and converts only those who are similar to respondents.

> Rank Method: - within classes determined by control totals convert the highest ranking non-respondents until control totals are satisfied.

REGRESSION RAKING

Given that the SPSD database includes complete household and family structures, it

is essential to associate a single weight with each household that will guarantee consistency in tabulations at the household, family and individual levels.

The household weights are constrained (raked) to agree exactly with known, individual level, control totals. The control totals employed in this round are population by age, sex and province, number of households by province and census family status (i.e., child, spouse, single parent or non-family person) by sex and province. A full discussion of the technique may be found in Dufour and Lemaître (1987).

While regression raking is a linear approximation, this does not mean that agreement with control totals is approximate. The algorithm provides weights which satisfy the control totals exactly.

SPLITTING DATABASE

Splitting refers to a mechanical data preparation step that partitions the SCF (after suppression of outliers, randomization and regression raking) into three mutually exclusive subsets: high-income individuals, EI recipients, and all others. To simplify subsequent steps in the database creation, this split is done in such a way that no households containing high-income individuals also contain EI recipients. There are, in fact, a handful of such cases but EI recipients in these households are treated as though they received no EI. High income individuals are those whom are defined as high income filers while EI recipients are those who reported receiving some benefit in the SCF survey (or were converted to being recipients as a result of imputed item non-response).

Categorical Matching

Categorical matching involves creating 'fused' composite records from two micro-data databases. Consider two databases, a host database **A** and a donor database **B**. There are a variety of methods that can be used to attribute some or all of the information on a record from database **B** onto any given record from database **A**. All are based on the idea that we wish to find a record from database **B** which is in some sense similar to the given record from database ***A***. The determination of similarity is based upon variables common to both databases and is affected by the intended use of the 'fused' records. Various 'nearest-neighbour' algorithms, which use methods similar to those of cluster analysis, can be used to determine a mathematically 'optimal' match, given a particular method of determining distance in N-dimensional space. Complications arise in practice due to limitations on the size of the set of 'donor' records (database **B** in our example) and the desire to use non-continuous variables (e.g. discrete or categorical).

In the SPSD a different, more heuristic, technique is used. It involves partitioning the two databases into identically-defined 'bins' of records, which are then sorted based upon one of the continuous variables common to the two databases (usually total income in SPSD). Records in a given bin are then matched one-for-one across the two databases (i.e. record **n** in bin **m** of database **A** is matched with record **n** of bin

m in database **B**). Complications arise because the number of records in a given bin is generally not equal in the two databases, and also as a result of the presence of record weights on one or both databases. These problems are solved by selectively duplicating records from one or both databases.

The SPSD uses categorical matching for adding EI data, and Green Book income data for high-income recipients, and for SHS data to all households. The technique allows the preservation of inter-item correlations from the donor record. Each of the matching procedures is described more fully in the following sections.

High Income Adjustment

The SCF has known reporting and sampling biases which result in a lower number of high-income individuals and fewer dollars of income per high-income individual than is indicated by personal income tax records. In the creation of the SPSD, both under-reporting and non-reporting of several income and deduction items are dealt with. Figure 2 provides an overview of this high-income adjustment process.

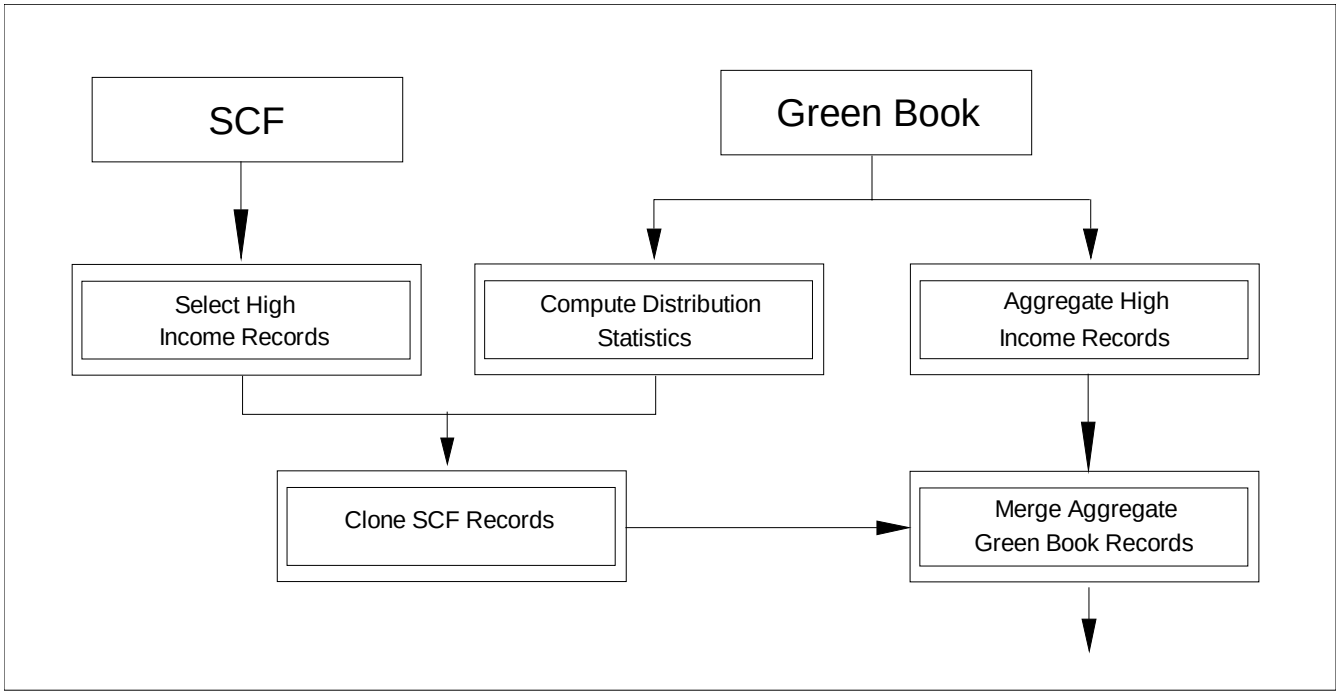


Figure 2: High Income Adjustment Process

MICRO-RECORD AGGREGATION

Non-reporting by high-income individuals in the SCF is ameliorated by using the Green Book counts for high-income filers. The weights of each high-income record on the SCF are adjusted so that the sum of the weights corresponds to the Green Book.

The weight adjustment process leaves them with very high weights (on the order of 200-500). These records are used as the "hosts" for accepting the more precise

information from the Green Book. This in turn provides the basis for an adjustment of income items for the high-income group.

Even with a scaling up of the weights for high-income records on the SCF, there is still a substantial under-reporting of income in this group. As a second step, under-reporting bias is corrected by replacing the income components on these records with plausible but non-identifiable sets of income items from the Green Book.

SCF Income Items Replaced for High Income Individuals

- **Employment Related**

  Earnings from Employment
  Farming Net Income
  Other Allowable Employment Expenses
  Self-employed Income - Non-farming

- **Investment Related**

  Allowable Other Years Capital Loss
  Allowable Prior Years Non-capital Loss
  Carrying Charges
  Capital Loss on Disposition of CCPC Equities
  Interest Income
  Net Rental Income
  Other Investment Income
  Taxable Capital Gain/Loss For Year
  Taxable Amount of Canadian Dividends

- **Other**

  Other Taxable Income
  Imputed Total Income - Sum of Components

Records from the Green Book are grouped into sets of at least 5 records. These grouped records are considered to be a non-confidential table although they retain many of the characteristics of micro records. The groups represent individuals of similar age, employment income, investment income, dividend income and capital gains. For these groups an average is calculated for the items listed above. Once grouped, the records are considered non-confidential since they represent 5 or more individuals. This is equivalent to publishing a table in which each cell contains no less than 5 individuals.

The resultant aggregate contains several thousand pseudo microdata records representing several tens of thousands of Green Book Records, in turn representing more than one hundred thousand high-income filers. These aggregate records, derived from otherwise confidential microdata, are now able to become part of a public use data set with little loss of information.

CATEGORICAL MATCH

The original SPSD records are duplicated to match the number of aggregated Green Book high-income records. These SPSD records do not provide a sufficient basis for the demographic characteristics of the high-income filer population. Thus a detailed match by age, sex, province and total income would not be feasible. Instead, the duplicated SPSD records were imputed a new value of total income based on a very simple age break (2 groups), sex and region using the same procedure described in a subsequent section (**Stochastic Imputation of Income Tax Information**). This new imputed value of total income was used as a key to sort the SPSD records before merging the similarly sorted, aggregate Green Book pseudo microdata records.

To improve the match with regard to age, sex, province, total income and tax status, a much larger original SCF sample would be required.

Employment Insurance History Imputation

Employment Insurance (EI) is a complex insurance and temporary income maintenance program, the administration of which requires monitoring claimants` weekly labour market activities. The administrative data collected under the program serves to (i) track the weekly benefits and claim activity of EI recipients, (ii) establish eligibility and entitlements by monitoring previous program participation in the event of repeat or re-entrant claims, and (iii) monitor past employment patterns through "Records of Employment".

EI benefits are an important component of both disposable and taxable income. Reported and simulated EI benefits serve to indicate program costs, client population, and gainers and losers under alternative program structures. For consistent analysis as well as input to the income tax module, benefit payments are needed on a calendar year rather than a claim basis. Thus, the initial task in constructing this component of the database required simultaneous development of a EI simulation module and identification of a limited set of "program relevant" EI variables (Table 2) that could serve as input to the EI simulation module.

EI DONOR DATASET

The EI administrative histories imputed to SPSD were based on a 10% sample of administrative records from the population with some EI claim activity within the base calendar year.

The sample consists of about 221,000 individuals and represents about 240,000 claims. The information selected from the file insures data confidentiality, and is rich enough to capture the labour force history relevant to application of EI program regulations. The following list shows a set of variables employed as input to the EI model.

EI History Variables

Claim Sequence Number (1st. or 2nd in current year)

Repeater Flag
Initial Benefit Type
Type Change Flag
Weeks of Benefits (current claim)
Weeks of Benefits (in previous 52 weeks)
Hours of Work (prior to current claim)
Minimum divisor weeks
Average Weekly Earnings (prior to claim)
Week Claim Established
Benefits Paid in Calendar Year (1 or 2 claims)
Weeks of Benefits Paid in Calendar Year
Weeks of EI benefits in each of the last five years previous to claim (1 or 2 claims)

Each SCF records which had some reported EI income in the calendar year was categorically matched to four beneficiaries selected from the 10% sample of EI beneficiaries. The matching keys are claimant age, benefit type, total benefits in the base year, province and sex.

Claim types are an important element in the match, since there are currently major differences in eligibility rules and in entitlements between these types. A claim type classification was constructed on the SCF dataset by (i) identifying EI recipients aged 65+ (retirement benefits), (ii) identifying EI recipients with occupation coded as "Hunting, Fishing, Trapping" (fishing benefits), and (iii) identifying female EI recipients with a child aged 0-1 (maternity benefits). No distinction could be made between sickness and regular benefit types on the SCF dataset.

CATEGORICAL MATCHING

Matching was carried out by first partitioning the donor administrative (EI) and host (SCF) datasets on the basis of age group, province, sex, and claim type. Duplication of records within cells was carried out to ensure that corresponding cells of the EI and SCF datasets had equal numbers of records. If in any given cell the number of host records exceeded the EI records, then the EI records were uniformly duplicated (EI data were a simple random sample). Correspondingly, if the number of EI records exceeded host records, then host records were duplicated in proportion to their weights (recall that the host data were based on a stratified sample). The latter case was the more frequent condition (in 170 out of 218 cells), but the former also occurred (a consequence of stratified survey design). Duplicates of SCF dataset records had weights adjusted in proportion to the number of times that they had been duplicated.

The outcome of the cell match and duplication steps was an increase in the number of records representing the EI claimant population. Initially, the SCF dataset contained about 10,000 such records, while after duplication there were approximately 40,000 records. This expansion of the dataset was intended to ensure full use of the EI histories available from the 10% sample.

Within cells, matching host and EI records were identified as the records with corresponding rank in the two datasets. The records were ranked on the EI benefits received (in dollars).

A second independent matching procedure is applied to records identified as potential beneficiaries. This add some extra 20,000 to 25,000 new records. Those records are added to provide a tool to increase EI participation without getting into complicated reweighting procedures. Potential beneficiaries were identified at the same time the correction for under-reported EI benefits is implemented.

Household Duplication

There are three conditions under which duplicates of SCF household records are created. These are: (1) in the imputation of taxation data to high-income earners, (2) in the categorical matching of EI data, and (3) in the creation of a synthetic group of institutionalized elderly. This latter group has been "created" because the underlying sample frame of the host dataset, the SCF, excludes the institutionalized population, and because the elderly are the largest and most policy relevant portion of this excluded population.

In the case of taxation or EI data, the motivation for household duplication is to utilize as much of the richness and variety in the donor administrative microdata sets as is possible. Duplication or cloning of host SCF records provides the basis for fully absorbing this variety in the donor datasets. Note that in both of these cases, duplicates of individuals are formed first. Then the other individuals in their household are also duplicated. In the event that more than one member of the same household is duplicated (e.g. if more than one household member received EI benefits), then additional duplication is necessary to ensure that each individual is properly represented. Duplication, rather than changing individual weights, is necessary if the weights of all the members of the household are to remain the same.

Finally, a pseudo sample of the institutionalized elderly has been created. This was done simply by duplicating the records of the non-institutionalized unattached elderly (aged 65+) who are not labour force participants. The motivation for selecting this donor population is that these individuals are most likely to resemble the institutional population. The weights on these records are adjusted to reflect the census counts of the institutional population by age, sex and province. When the base year is not a census year, the closest census is used and shares of institutionalized in the census are applied to SCF data.

Stochastic Imputation of Income Tax Information

This section will describe stochastic imputation, the method used to attribute personal income tax information to the SPSD records. The information in this case differs from the match used to improve the representation of high-income recipients. In that former case, the information being added was principally incomes by source.

In this case, the information being added is mainly various itemized deductions, exemptions and tax credits reqEIred for the calculation of income tax liability. The following list of items was imputed from the Green Book onto the SPSD. These are items which are not well represented on the SCF (e.g., capital gains), entirely absent (such as carrying charges) or not easily modeled (e.g., disability deduction).

1. Other Allowable Employment Expenses
2. Carrying Charges
3. Child Care Expenses Allowable
4. Charitable Donations and Gifts
5. Allowable Other Years Capital Loss
6. Disability Deduction
7. Union and Professional Dues
8. Education Deduction for Student
9. Other Federal Tax Credits
10. Federal Political Contribution Tax Credit
11. Taxable Capital Gains
12. Capital Loss on Disposition of CCPC Equities
13. Federal Investment Tax Credit
14. Net Medical Calculated Amount
15. Allowable Prior Years' Non-capital Loss
16. Other Deductions from Net Income
17. Other Dependent Exemptions
18. Provincial Tax Credits
19. Total RPP + RRSP Contributions
20. Proportion of RRSPs in (RRSP + RPP)
21. Tuition Fees

These items, in combination with other provisions which can be readily computed from available data (e.g., personal exemptions) allow a complete calculation of taxable income and tax payable.

THE DONOR DATA

The source data for the imputation were derived from a Revenue Canada sample of Individual Tax Returns in the base year. The sample is stratified by source of income, urban geographic area, rural geographic area, tax status (taxable and non-taxable), and income range.

The information in this sample contains most of the information submitted in the base year T1 Federal and Provincial Individual Income Tax Return and accompanying schedules. This sample has no explicit family structure (i.e., the returns of the head, spouse and dependents cannot be analyzed together in an identifiable family unit).

DATA TRANSFORMATIONS

To join these Green Book income tax data with the SCF-based host sample, a set of common classification characteristics were defined. The following attributes were

chosen as much for their degree of policy relevance as for their availability and similarity of definition on both datasets:

1. Taxing province
2. Age group
3. Sex
4. Marital status as taxed
5. Total Income class (excluding Capital Gains)
6. Employment Income class
7. Children claimed for the Child Care Expense Deduction (on SCF, number of children eligible for claiming).

Sub-samples defined by the cross-classification of these items are assumed to have sufficiently different distributions to merit retaining the uniqueness of these distributions. A comparison of charitable donations between the same groups is provided in Figure 3.

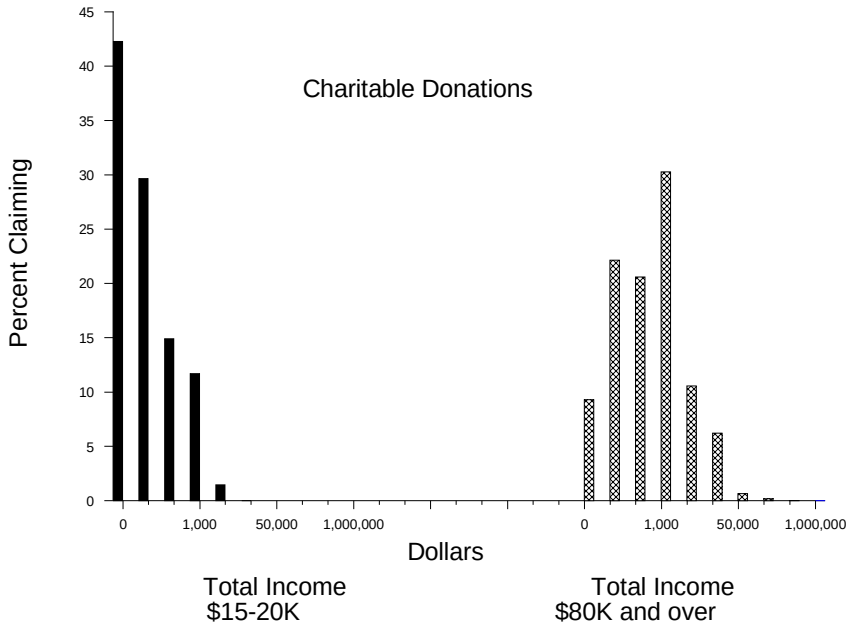


Figure 3. Green Book Distribution of Charitable Donations

Prior to imputation, the host dataset was prepared by identifying potential tax filers, establishing eligibility for certain targeted items (Education, Tuition and Child Care Expense Deductions), and creating a parallel classification scheme on both the host SPSD and donor Green Book datasets.

A model of the personal income tax system (the same one subsequently used for policy analysis) was initially employed to identify likely tax filers and to impute marital status as taxed. For example, a married person eligible to claim his or her spouse as a dependent, would be designated married-taxed-married. This imputation was essential to restrict the imputation to a similar universe as the donor dataset.

Three of the deduction items were treated specially in that the eligibility for these

items could be identified on the host dataset. From information available on the SCF, one is able to determine if the individual is eligible for the Education Deduction (self or dependent is attending a post-secondary educational institution), Tuition Deduction (self is attending a post-secondary institution) and the Child Care Expense Deduction (for lower income spouse with children under 15 present). Targeting the imputation to individuals eligible for these deductions ensures some degree of internal consistency in the synthetic records. For example, only persons with children will be imputed the Childcare Expense deduction. Unfortunately it is not as simple to determine eligibility for all deductions and income items imputed.

The joint distribution of RPP (Registered Pension Plan) and RRSP (Registered Retirement Savings Plan) contributions posed a problem in that the tax law restricts the total of the two to be below a certain limit. Imputing the two separately would not ensure that this threshold is not exceeded. To overcome this, we imputed the sum of the tax filer's RPP and RRSP contributions, and then RRSP contributions alone as a proportion of this sum.

DERIVING DISTRIBUTIONAL STATISTICS

One objective of this imputation process is to ensure that average amounts of various deductions, exemptions and credits claimed on the SPSD accurately reflect the actual (e.g. published) averages for sub-groups defined, for example, by province, age, income range. etc. A further and more stringent objective is for the SPSD to reproduce the distribution of these items as found in the Green Book file. This requires a method of representing a arbitrary density functions. For example, the method should equally well represent bimodal, truncated and long-tailed distributions.

Another factor in the choice of method was its computational intensity. Since the source dataset contains almost 400,000 records, the algorithms to generate these representations had to be reasonably efficient.

The method eventually chosen was first to disaggregate the overall population hierarchically using the classification variables listed above. Then within each of these hierarchically defined subgroups, the univariate distributions of particular items was represented first by the proportion in any given sub-group with a non-zero value for the item. Then, for the sub-sub-group with non-zero values, the density function was represented by the decile cut-off points, with special treatment of the tails of the distributions.

A constraint was imposed on the hierarchical disaggregation procedure in order to assure non-confidentiality of the resulting statistics. This constraint was to require a minimum number of observations in each of the sub- or sub-sub-groups. To make the fullest possible use of the data, the disaggregation process was applied independently for the percentage reporting and distribution (i.e. decile) statistics. The percentage reporting statistics could be based on a much smaller number of observations than the decile cut points, so that information from a finer level of disaggregation could be used.

The percentage reporting statistic was kept if the sum of weights for the cell exceeded 400 or the number of records representing a non-zero value exceeded 20. If these criteria were not met, the statistics for a higher level of aggregation was substituted.

The criteria for the distribution statistics had to be more rigorous. The minimum cell size was 100 records, i.e. if a cell did not contain at least 100 non-zero records, statistics for that cell were not computed. Instead, the distribution statistics were computed from a higher level of aggregation.

For each item to be imputed (all those listed at the beginning of this section), the nearly 400,000 income tax return records were classified into relevant cells (e.g., income group by age by marital status by sex by province).

For each of these groups, given a sufficient sample, the following statistics were computed:

- values for decile cut-points 1 through 9,
- the mean of the bottom and top deciles,
- the mean of the highest 5 values and the mean of the lowest 5 values, and
- the percentage within the cell reporting a non-zero value for the item.

These statistics are well suited for representing an arbitrary distribution and they are simple to calculate.

For confidentiality reasons, the actual maximum and minimum values in a cell could not be used. The mean of the highest five values and the mean of the lowest five values in the cell were used as substitutes.

The same statistics were then generated for aggregations of cells, in this case, for income group by age by marital status by sex by region. Collapsing the 10 provinces into 5 regions increases the level of aggregation and therefore increases the number of individuals within a cell. More cells will then meet the minimum size criterion for computing the sets of distributional statistics.

Ideally, all values would be imputed from the lowest level of aggregation. However, due to the sparseness of many of the data items this is rarely possible. For example, Other Allowable Employment Expenses are concentrated in the higher income groups and cells in this region would be well represented. For the lower income groups, the cells are sparser and often empty.

To fill in these sparse and empty cells, statistics from higher levels of aggregation are substituted. If, for instance, the cell representing the following classification:

Income Group	$35,000 to $39,999
Age Group	25 to 35
Marital Status	Single, Taxed Married
Sex	Female
Province	Quebec

were empty or rejected on the size criterion, statistics would be substituted from the next level of aggregation:

Income Group	$35,000 to $39,999
Age Group	25 to 35
Marital Status	Single, Taxed Married
Sex	Female

representing this income group, age group, marital status and sex for all of Canada. If this cell were also sparse or empty, statistics would be substituted from the next higher level of aggregation. In the worst case, the statistics for a cell would be derived from the entire sample, i.e., all income groups, all age groups, all marital statuses, both sexes and all provinces.

The resultant distribution and percentage reporting statistics are non-confidential since they never reveal raw data values. The extreme values are synthesized by calculating the mean of the highest 5 values and the mean of the lowest five values. Thus, each statistic is based on at least five observations, the rule of thumb adopted for assuring non-confidentiality.

IMPUTATION

Using this complex set of distributional statistics generated from the Green Book file of income tax returns, it is possible to recreate the same distribution of values on the host dataset. For each eligible individual on the host dataset, a synthetic value is drawn from a distribution representing the tax returns of a similar group of people.

Values for the middle eight deciles are generated assuming a uniform distribution between decile cut-off points. (More complex density functions were tried within these deciles. However, tests suggested that the gain in accuracy was marginal, especially in light of the much increased computational costs.)

The top and bottom deciles are treated specially so that both the shape and the size of the tails are accurately represented. Preservation of the tail of the distribution is essential to maintaining overall means and totals, especially for items with long-tailed distributions such as capital gains or business losses.

In imputing the upper and lower deciles, values are drawn assuming a Pareto distribution to generate the appropriately shaped tail. The specific Pareto distribution used in each case is such that the mean of the decile is maintained. Extreme values are truncated at the mean of the highest or lowest 5 values in the group.

Survey of Household Spending Data Imputations

SPSD is based on SHS in the year closest to the base year. When SHS survey year is different from the base year, SHS data are adjusted to reflect SCF household compositions in the base year.

Household expenditure data are intended to support simulations requiring information on shelter costs (e.g. Social Assistance), simulations concerned with child care costs, and simulations of commodity taxes. Due to the limited number of records in SHS (about 16,600), it was decided to perform a consumption structure imputation using as many households categories data can support. The limit is 30 observations by class.

Two main steps were involved for the consumption pattern imputation:

- Multivariate analysis creating the matching variable
- Categorical Matching (Weighted Duplication)

CREATION OF THE MATCHING KEY AND CONSUMPTION PATTERN

A special matching solution was developed for SHS based on complex multivariate analysis. Households in one class are grouped according to the similarity of their consumption patterns, not their consumption levels. One underlying assumption is that pattern of consumption are more similar between provinces within a given class than they are between classes. Variables used to define the consumption patterns are the share of total income for the 47 categories of expenditures, and some extra variables (e.g. income, taxes, savings) which are included in order to complete the basic household accounting identity - income equals expenditure plus saving.

Household type, tenure, sex of household head, income class, age group, urban size, head working full time, spouse working full time, region, presence of teenagers, and presence of kids are used as classification variables. The multivariate analysis evaluates the explanatory power of each classification variable. For a given level of the selected classification variable, the same methodology is applied to the remaining variables – income splits are an exception and can be selected many times. The limit is reached when there are less than 30 observations in the same category in SCF or SHS. The first variable is forced to be household type in order to insure consistency of expenditure patterns: it is expected that a single parent family with 2 kids does not have the same consumption pattern as a household with two adults and one kid.

The procedure to create the matching key starts with the household type. If household type is 0, H0; the next selected variable is tenure. For the tenure type 0, T0; the procedure select for the sequence H0T0 the next variable with the most explanatory power. It appears to be total income median split of the population in the H0T0 class, M0 and M1. For H0T0M0, the next variable selected is the sex of the household head. In another sequence, H4T1M0, the model selects the age group as the variable with most explanatory power. In the 1997 database, the procedure generates 343 categories. Each category is associated with a sequence number: the matching key hdevmv.

The likely analytical uses of the data (e.g. tables by province, income group, or family type) were taken into account in creating the matching key. If a more complex household structure is used, the variable hdevmv could help the user in defining the structure. Because many households share the same hdevmv, the user should make sure that all the households with a given hdevmv fall in the same class within a province, otherwise inconsistent results may be produced.

CATEGORICAL MATCH

The categorical match of records was performed at the household level and required only the duplication of SHS records. In order to make the fullest possible use of SHS data without having to duplicate SCF records, matching bins were created in such a way as to ensure that the SHS bin sample size was always smaller than its SCF counterpart. Because the unduplicated host dataset was approximately four times as large, it was infrequent that a bin would have to be redefined because the SCF bin had fewer observations than its SHS counterpart. The match took the form of a weighted duplication of SHS records, and was designed to force the SHS sample counts within bins to match the corresponding host bin.

The general task in this weighted duplication procedure is to increase the number of SHS observations in any bin, by cloning or duplication, to equal the number of host SCF observations in the bin. The first step is to sort both the host and donor bins in ascending order of total income. On average, there might be about six times as many SCF records as SHS records. However, it would be inappropriate simply to make five clones of each SHS record because this would in effect treat the SHS as a sample rather than as a stratified random sample; no account would be taken of the SHS sample weights. Instead, those SHS records with higher weights are cloned proportionately more than those with smaller weights.

More precisely, a weighted probability of occurrence of SHS household i in bin j is calculated. By multiplying this probability by the desired host bin sample size, an estimate of the number of times a given SHS household should appear in the host dataset is obtained. However, in some cases this number is less than one and in these cases the household would not be matched with any host records. In order to insure no such loss of data from SHS dataset, at least one match is assigned to every SHS record, and then the probability is multiplied by the difference between the sample sizes of the host and donor bins. In other words, every SHS household is given at least one match and the number of duplications still required to hit host bin size are distributed across the SHS records according to their weight. If the probability so determined is simply rounded or truncated to its integer equivalent, rounding error can produce an incorrect total host bin count. To correct for this error a cumulative total of the host cell frequencies D is calculated.

$$D_{ij} = \sum_{k=1}^{i} \left[\left[\frac{Wij}{\sum_{i=1}^{n} Wij} \right] \times \left(N_j^h - N_j^d \right) \right]$$

Where:

i = the i^{th} SHS household

j = the j^{th} matching bin

W = the weight of the SHS donor record

N^h = the sample size of the SPSD host bin

N^d = the sample size of the SHS donor bin

Each SHS record is then duplicated by the rounded value of the cumulative total minus the rounded value of the previous record's cumulative total plus one. In this way the rounding error is distributed throughout the cell, every SHS record is ensured at least one match, and the correct host cell totals are reached.

This procedure serves largely to preserve the weighted distributions of the SHS data, at least until SPSD weights are associated with it. The difference between the SCF and SHS weights can however create distortions in the matched distributions.

Survey of Labour and Income Dynamics Data Imputations

SPSD includes data from the Survey of Labour and Income Dynamics (SLID) in the base year. It was used to impute hours worked to EI claims. This was necessary since the Employment Insurance program uses hours worked as a basis for qualifying for claims, while the Employment Insurance program used weeks worked.

The method of imputation used was categorical matching which was done in a similar way as the SHS final match.

Three main steps were involved for each of the three imputations.

- Selection/Grouping of weekly hours of work for Imputation
- Selection/Construction of Matching Variables
- Categorical Matching (Weighted Duplication)

Since the imputation of weekly hours of work was being done to persons who had EI claims in the base year, the SLID donor sample was chosen to reflect this population. Only those SLID respondents who received EI benefits during the year were used in the match. Furthermore, in deriving the weekly hours of work, only months with no EI receipt were chosen and with average weekly hours of work of 15 hours or more were included since these are the months which will most closely

match the hours worked prior to EI (since weeks with low income and less than 15 hours were not included in the program prior to 1996).

The matching variables used were sex, industry, and province. A categorical match was then done. See the previous SHX section for more details on the methodology of categorical matching.

References

Adler, H.J. and M.C.Wolfson (1987), "A Prototype Micro-Macro Link for the Canadian Household Sector", International Association for Research in Income and Wealth, Rome, 1987, and forthcoming in The Review of Income and Wealth.

Dufour, J. and G. Lemaître (1987), "An Integrated Method for Weighting Persons and Families", Survey Methodology , Vol.13, No. 2, 199-207.

Ruggles, R. and N. Ruggles, "The Integration of Micro and Macro Data for the Household Sector", The Review of Income and Wealth, series 32, no. 2.