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 large sample of 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 Human Resources and Development ministry 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 behavioural 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. The SPSD/M provides 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. Individual-level or family level analysis is possible.

A key challenge in the construction of the database portion of the SPSD/M has 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.

- **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 (1988).

DATA SOURCES

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

- **The Survey of Labour and Income Dynamics (SLID):** 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. It replaces the Survey of Consumer Finances (SCF) which was last conducted in 1997. The SPSD starts with the public use microdata file (PUMF);

- Personal income tax return data: a sample of over 400,000 personal income tax (T1) returns used as the basis of Canada Revenue Agency's annual **Income Statistics** (also known as the Greenbook and formerly known as Taxation Statistics) 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). The SPSD starts with the public use microdata file (PUMF).

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.

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, 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).

- **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 SLID, the Greenbook) 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.
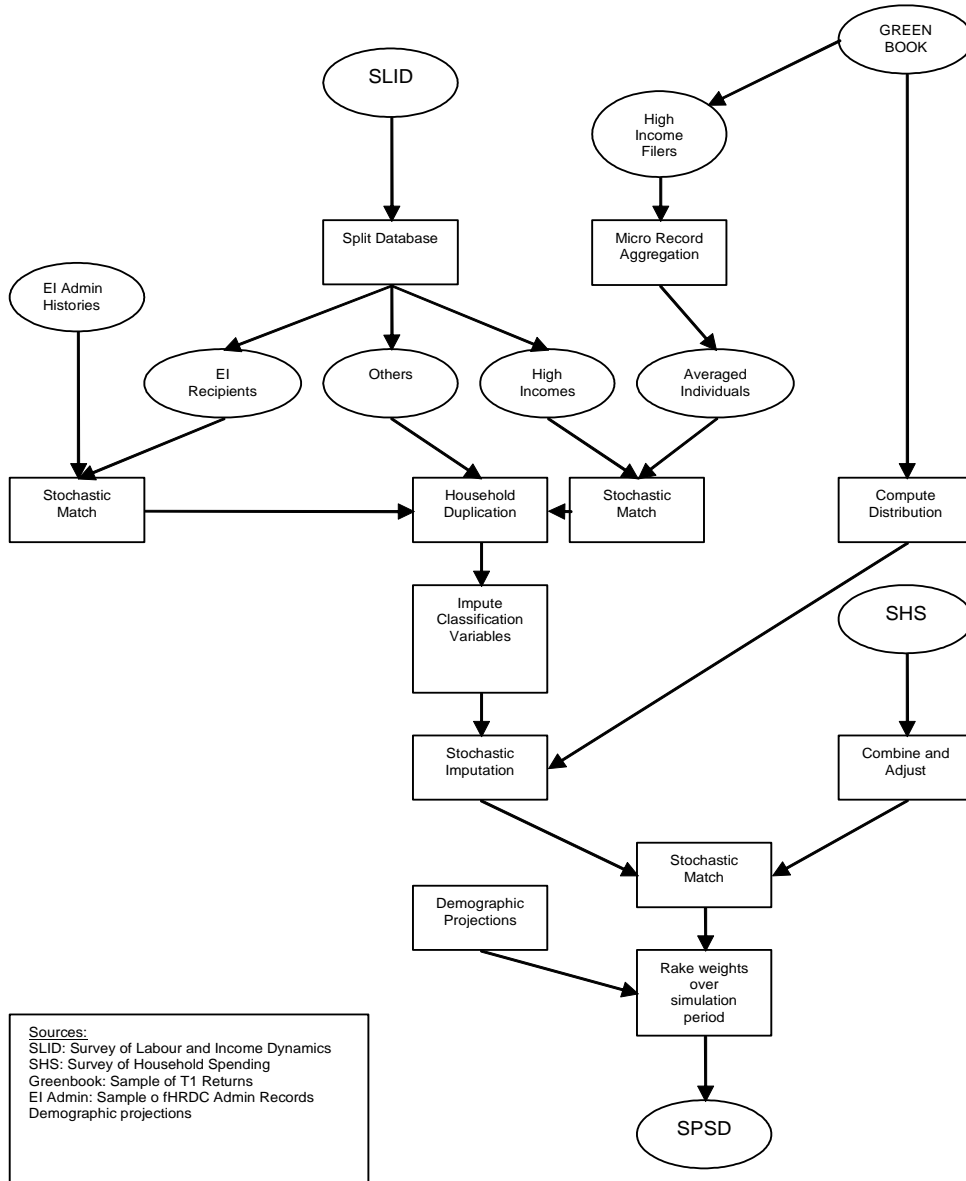

Figure 1: SPSD Database Creation Process

The Host Data

The target or "host" dataset is derived from Statistics Canada's **Survey of Labour and Income Dynamics** (SLID) in the base year. SLID is an annual survey administered to selected households drawn from the survey frame of the Labour Force Survey (LFS). It is a longitudinal survey with respondents staying in the sample for 6 years. In January, they are asked about their labour market experiences in the previous year as well as any educational or family based changes. In May, the data relating to the previous year's income is collected. Two methods are used to collect income information. A large portion of respondents (about 80%) gave Statistics Canada permission to use their T1 tax information. The rest are interviewed in the month of May which allows them to use their tax forms as a guide. The SPSD starts with the SLID public use microdata file.

The information from the EI, Greenbook and SHS files was then "added" to the SLID. In order to exploit the full variety of this information being imputed from other sources, many original SLID records were cloned or duplicated. For example, records representing persons who receive EI or who might potentially receive EI were duplicated. Records representing high-income filers (those with an income of over $135,000) were duplicated to correspond to the number of high-income records derived via micro-record aggregation from the Canada Revenue Agency 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.

SLID WEIGHTS

Historically, SLID wages and salaries produced total wages which were greater than corresponding T1 or T4 estimates. The T1 estimates are the estimated wages and salaries as derived from individuals' tax returns. The T4 estimates are the estimates which come from the T4 forms the employer fills out which get sent to the employees. A comparison with T4 file showed an over representation of population in the median group and an under-representation of the low wages population. Since wages represent the largest source of market income in Canada, SLID uses the T4 file as a benchmark for calibrating the wage distribution.

In reference year 2003, SLID underwent a historical revision in which the weights were calibrated based on the wage distribution of the T4 administrative file. The revised weights also incorporated a change from the 1996 Census to the 2006 Census for demographic estimates. The historical revisions went back to 1990 (including the SCF). SPSD/M Version 15.0 was the first release to include SLID's new methodology. Weights of releases prior to Version 15.0 were not revised.

The base year population on the SPSD will not necessarily be the same as the SLID population. SLID excludes people in institutions, population on Indian reserves, those living in military barracks, and residents of territories. The SPSD adds back institutionalized elderly people, but not the other missing population. So while SLID covers 97% of the population, the SPSD covers 98%. This undercoverage is considered when the SPSD/M

weights are developed.

EDITING SLID VARIABLES

The public release microdata file of SLID is the starting point of the SPSD. There were many changes done to variables on this file, and some of the most important changes will be documented here.

Imputation of "Don't Know" responses

There are some variables in SLID which are used in the SPSD which have as an answer "don't know". The variables in question are: province of residence, marital status, industry, occupation, work status last year, education level, education status last year and total paid hours worked last year. Regressions or random draws from distributions on the PUMF were used to impute answers to these individuals.

Guardian children

In SLID, children under the age of 18 who do not live with a parent are not included in census families. In the SPSD these children were considered to be children of the major income earner of the household.

Head of family

The definition of the head of the family was changed in the SPSD from the one used in the SLID. Census families are couples (either married or common-law), single parents, and their children under the age of 25 who live with them. In the SPSD, unattached individuals are considered to be census families of size one. If there is a couple, the man is deemed to be the head of the census family. If not, then the single parent or unattached individual is the head.

Economic families consist of people who live in the same household and who are related to each other by blood, marriage, common-law, or adoption. In the SPSD, unattached individuals are considered to be economic families of size one. The head of the economic family is defined in the SPSD to be the head of the census family which contains the major income earner.

The head of the economic family which contains the major income earner of the household becomes head of the household.

Rounding of income

Income variables on the SLID PUMF are rounded. Since "unrounded" variables are needed in tax modeling, the income variables were all "unrounded". The methodology used in order to "unround" the data was not the same as that used to "round" the data. This was done in order to keep the perturbation techniques confidential. The variables were "unrounded" within intervals while preserving the distribution found in the Greenbook, a sample of T1 tax returns. In order to preserve confidentiality of the Greenbook and ensure that unique values were not found, values found on the Greenbook were averaged in groups of five prior to the distributions being created.

Truncation of age

SLID truncates age at 80.   The distribution of age from the 2006 Census is used to assign more diverse ages to the elderly.   The imputation of age to the elderly also takes into account the Census distributions for those in institutions, marital status and sex.   Note that this is done after the imputation of the institutionalized elderly so that these cloned records may end up with different ages than their original counterparts.

ADDING THE INSTITUTIONALIZED ELDERLY

The SLID frame does not include the institutionalized elderly.   However, because the elderly are a large and policy relevant group, they were added back into the SPSD.   This is done by finding people in SLID who live alone and who did not work in the previous year.   A certain number of these observations are then duplicated and are flagged as institutionalized.   The duplication continues until the SPSD matches census' proportion of elderly who live in an institution by province, age, and sex.

CONVERSION

External evidence suggests that under-reporting of Employment Insurance, Social Assistance, and 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.   This evidence is supported by the fact that the amount of under-reporting of these items in SLID is much less severe than it was in the Survey of Consumer Finances.   A portion of the respondents in SLID had their income variables imputed from their income tax forms which results in less item non-response.

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 to ensure that the database balances to known 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 was 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.

SPLITTING DATABASE

Splitting refers to a mechanical data preparation step that partitions the SLID into four subsets: high-income individuals, EI recipients and potential EI recipients, and all others. Note that a person may belong to both the high income subset and the EI subset. High income individuals are those whom are defined as high income filers, while EI recipients are those who (i) reported receiving some benefit in the SLID survey, (ii) were converted to being recipients as a result of imputed item non-response, or (iii) were deemed to be potential EI recipients.

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, Greenbook income data for high-income recipients, and 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 SLID estimates for high income individuals (as well as the income for these individuals) are lower than is indicated by personal income tax records. Both under-reporting and non-reporting of several income and deduction items are dealt with in the creation of the SPSD. Figure 2 provides an overview of this high-income adjustment process.
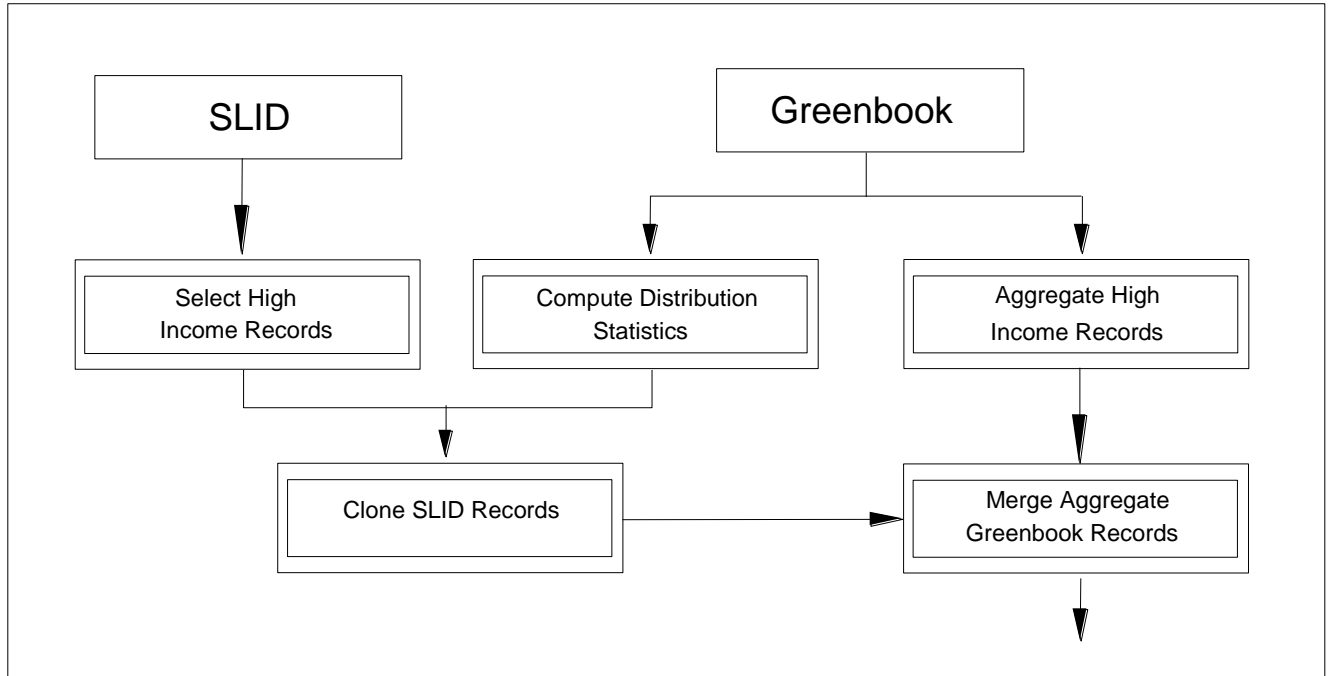

Figure 2: High Income Adjustment Process

MICRO-RECORD AGGREGATION

Non-reporting by high-income individuals in the SLID is ameliorated by using the Greenbook counts for high-income filers. The weights of each high-income record on the SLID are adjusted so that the sum of the weights corresponds to the Greenbook. These SLID records are used as the "hosts" for accepting the more precise information from the Greenbook. 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 SLID, 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 and some deduction components on these records with plausible but non-identifiable sets of income and deduction items from the Greenbook.

SLID Income and Deduction Items Replaced for High Income Individuals

- **Employment Related Income**

     idiemp - Earnings from Employment
     idise - Self-employed Income

- **Investment Related Income**

     ididiv - Amount of Canadian Dividends
     ididive - Amount of Eligible Canadian Dividends
     ididivie - Amount of Other than Eligible Canadian Dividends
     idiinvnd - Other Investment Income
     idicapg - Taxable Capital Gain/Loss For Year
     idipens - Pension Income
     iditrrsp - RRSP withdrawals

- **Other Income**

     idialimo - Alimony
     idiworkc - Workers Compensation
     iditoth - Other Taxable Income

- **Deductions from total income**

     idrpp - RPP deduction (207)
     idrrsp - RRSP deduction (208)
     iddues – Annual Union, Professional or Like Dues (212)
     idiloss - Business investment loss (217)
     idmovexp - Moving expenses (219)
     iddalimo - Support payments made (220)
     idcarry - Carrying charges and interest expenses (221)
     idexplor - Exploration and development expenses (224)
     idalexp - Other employment expenses (229)
     idothded - Other deductions from total income (232)

- **Deductions from net income**

     idpartlo - Limited partnership losses of other years (251)
     idnclos - Non-capital losses of other years (252)
     idcloss - Net capital losses of other years (253)
     idcapgex - Capital gains deduction (254)
     idaddded - Additional deductions from net income (256)

- **Non-refundable and refundable tax credit information**

     idtuitn - Tuition fees for self (323)
     idmedgro - Gross medical expenses (330)
     idcharit - Charitable donations (340)
     idgifts - Cultural and ecological gifts (342)
     idpolcon - Total federal political contributions (409)
     idlabtxg - Labour-sponsored funds tax credit (414)

idmincar - Minimum tax carry-over (427)
idfortx - Non-business income tax paid to a foreign country (431)
idforinc - Net foreign non-business income (433)
idgstreb - GST rebate (457)
idprvftc – Provincial foreign tax credit
idtransit – Public transit pass amount

For each province, records from the Greenbook 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 a weighted 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. In addition, the addition of a weight to the average adds uncertainty.

The resultant aggregate contains thousands pseudo microdata records representing several tens of thousands of Greenbook records, in turn representing more than three 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 high-income records are duplicated to match the number of aggregated Greenbook high-income records by province. These 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 age group (6 groups), sex and region using the same procedure described in a subsequent section (**Stochastic Imputation of Income Tax Information**). For each province, this new imputed value of total income was used as a key to sort the SPSD records before merging the similarly sorted, aggregate Greenbook pseudo microdata records.

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

Employment Insurance History Imputation

Employment Insurance (EI) is a complex 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 an 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.

Given the fact that the number of people who receive EI will vary during the SPSM model's time frame, the SPSM needs to have a mechanism whereby it can increase and decrease the number of recipients over and above the changes that small program changes would entail. In order to accomplish this, a regression determines which of the people who did not receive EI are most likely to receive it in the future. In order to allow for an 80% increase in the number of EI recipients, the most likely candidates will be duplicated and then will also have EI histories imputed to them. See the sections of this guide on conversion and splitting the database for more information.

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 over 200,000 individuals and represents nearly 250,000 claims. People on the SPSD can have up to two claims per calendar year. 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
       Main Benefit Type
       Type Change Flag
       Weeks of Benefits (current claim)
       Hours of Work (prior to current claim)
       Minimum divisor weeks
       Average Weekly Earnings (prior to claim)
       Week Claim Established
       Weeks of EI benefits in each of the last five years previous to claim (1 or 2 claims)
       Effective weekly benefit rate
       Local unemployment rate
       EI exhaustee flag
       Weeks of training benefits
       Training benefit weekly rate
       Weeks of other benefits (other benefits include items like job creation or work sharing benefits)
       Other benefits weekly rate
       New entrant / re-entrant flag

Received parental benefits

To this list, we also add a variable for weeks of work prior to the claim. This variable is not available on the EI administrative file since it was only needed to derive EI benefits. This variable is estimated by imputing a 'weekly hours of work' variable to the claim and then using it along with the total hours worked prior to the claim to derive weeks worked prior to the claim.

Each SPSD record which had some reported EI income in the calendar year or who was flagged as a potential future recipient 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 SLID dataset by

(i)   identifying EI recipients with occupation coded as "Hunting, Fishing, Trapping" (fishing benefits),

(ii)   identifying female EI recipients with a child aged 0-1 (maternity benefits),

(iii)   identifying EI recipients attending school (training benefits).

No distinction could be made between the other benefit types on the SLID dataset.

CATEGORICAL MATCHING

Matching was carried out by first partitioning the donor administrative (EI) and host (SLID) datasets on the basis of age group, province, sex, and claim type. Duplication of records within these cells was carried out to ensure that corresponding cells of the EI and SLID 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). This process is done twice, once for SLID records who receive EI (or who were converted to becoming EI recipients), and once for SLID records who were flagged as potential future EI recipients.

The outcome of the cell match and duplication steps was an increase in the number of records representing the EI claimant population. Initially, the SLID dataset contained about 7,500 such records, while after duplication there were approximately 30,000 records with EI and 22,000 extra records who could potentially receive EI in the future.

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).

Quebec Parental Insurance Plan (QPIP) Imputation

The Quebec Parental Insurance Plan (QPIP) which began in 2006 provides for payment of

financial benefits to all eligible workers who take maternity, paternity or parental leave in Quebec. It replaces the maternity and parental benefits offered under the Employment Insurance program.

As the administrative data for that program was not available, the variables needed to model the QPIP program were derived from SLID and from the imputed EI claim data. QPIP recipients are still categorically matched to EI administrative records in order to be able to model EI for years prior to 2006. EICS (EI Coverage Survey) data were also used to impute weekly earnings for QPIP recipients via estimated regression coefficients. The method of imputing the QPIP variables is as follows:

1- The total amount of QPIP benefit received in the year (idiqpip) is taken from the converted amount of EI from SLID, and adjusted according to the aggregates from the publication of Quebec statistics.

2- The start date of the QPIP claim (idqpstrt) is initially assumed to be the same as the start date for the EI claim.

3- Impute weekly earnings for QPIP using estimated regression coefficients based on EICS data, and followed by additional steps of adjustments to match the distributions from Quebec statistics.

4- Arrange data by family unit, grouping family units as: only mothers have QPIP, only fathers have QPIP, and both mothers/fathers have QPIP.

5- Within each group, randomly determine whether the family elects for special type or regular type of QPIP.

6- Impute QPIP weeks in calendar year for either special or regular type. Adjust QPIP claim start date if the initial start date on the file is unreasonable.

7- For family units where both parents have QPIP, restrict total parental weeks to be within the maximum allowed for each type.

8- Calculate QPIP amount in calendar year according to the imputed QPIP weeks in calendar year, weekly earnings and QPIP rates by type, so that the imputed QPIP amount can be compared to the target QPIP amount from the publication of Quebec statistics.

9- Impute QPIP weeks for the next year for those who might have their QPIP continued to next year. Then total QPIP weeks can be calculated.

The resulting database variables used to model QPIP are:

idiqpip: Original/adjusted amount of QPIP
idqpstrt: Week of QPIP start
idqpfrst: First type of QPIP claim
idqptype: Type of QPIP claim
idqpweek: Weeks of QPIP

idqpern: Weekly earnings to qualify for QPIP

Household Duplication

There are three conditions under which duplicates of SLID 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 SLID, 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 SLID records provides the basis for 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 SLID 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 required for the calculation of income tax liability. The following list of items was imputed from the Greenbook onto the SPSD in version 20.0. These are items which are not well represented on SLID (e.g., capital gains), entirely absent (such as carrying charges) or not easily modeled (e.g., disability deduction). Only items with enough sample, usually at least 200 observations, are imputed.

1. ididiv: Actual amount of Canadian taxable dividends (120)
2. ididive Actual amount of Canadian eligible dividends
3. ididivie: Actual amount of Canadian other than eligible dividends (180)
4. idiinvnd: Investment income other than dividends
5. idrpp: RPP deduction (207)
6. idrrsp: RRSP deduction (208)
7. idccett: Child care expenses (Form T778, line 214)
8. idiloss: Business investment loss (217)
9. idmovexp: Moving expenses (219)
10. iddalimo: Support payments made (220)
11. idcarry: Carrying charges and interest expenses (221)
12. idexplor: Exploration and development expenses (224)
13. idalexp: Other employment expenses (229)
14. idclergy: Clergy residential deduction (231)
15. idothded: Other deductions from total income (232)
16. idemplo: Employee home relocation loan deduction (248)
17. idstkded: Stock option and shares deduction (249)
18. idpartlo: Limited partnership losses of other years (251)
19. idnclos: Non-capital losses of other years (252)
20. idcloss: Net capital losses of other years (253)
21. idcapgex: Capital gains deduction (254)
22. idnorth: Northern residents deductions (255)
23. idaddded: Additional deductions from net income (256)
24. idothpe: Amounts for infirm dependants age 18 or older (306)
25. idcgtc: Caregivers amount (315)
26. iddisslf: Disability amount (316)
27. iddisoth: Disability amount transferred from a dependant other than your spouse (318)
28. idintstu: Interest paid on student loans (319)
29. idtuitn: Tuition fees (320)
30. ideducpm: Months in school part-time (321)
31. ideducfm: Months in school full-time (322)
32. idmedgro: Gross medical expenses (330)
33. idcharit: Charitable donations (340)
34. idgifts: Cultural and ecological gifts (342)
35. idtransit: Public transit pass amount (364)
36. idpolcon: Total federal political contributions (409)
37. iditc: Investment tax credit (412)
38. idlabtxg: Labour-sponsored funds tax credit (414)
39. idmincar: Minimum tax carry-over (427)
40. idfortx: Non-business income tax paid to a foreign country (431)
41. idforinc: Net foreign non-business income (433)
42. idgstreb: GST/HST rebate (457)
43. idprvftc: Provincial foreign tax credit
44. idprvpol: Total provincial political contributions
45. idrentpd: Total rent paid in Manitoba and Ontario (6110)
46. idproptx: Net property tax paid in Manitoba and Ontario (6112)

47. idhomstu: College residence in Ontario, Manitoba resident homeowners tax assistance (6114)
48. idmschpd: School tax paid in Manitoba (6122)
49. idfitexp: Fitness expense amount for children (365)
50. idsprecpr: Children's Sport & Recrn Expense(Prov) (5849)

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.

For some deductions, people with high income whose income was replaced with an average from the Greenbook will also have their deductions replaced with an average from the Greenbook. See the section on high income for a more detailed explanation.

THE DONOR DATA

The source data for the imputation were derived from a Canada Revenue Agency sample of individual tax returns in the base year. The sample is stratified by source of income, place of residence, tax status, and total income range. The sample includes extra strata for earners with total income greater than $250,000, outliers, and non-residents.

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 Greenbook income tax data with the SLID-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. Income group
6. Wage group
7. Number of children
8. Size of union dues

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. For example, a comparison of the distribution of the size of charitable donations for two different income groups in 2004 is provided in Figure 3.

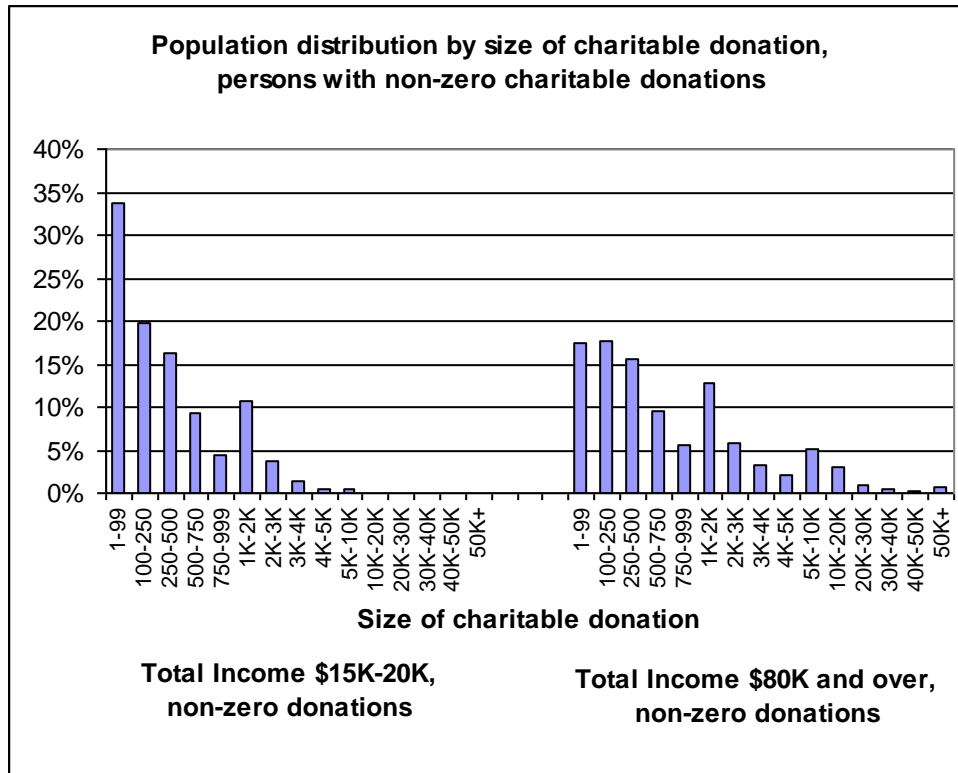**Population distribution by size of charitable donation, persons with non-zero charitable donations**

Figure 3: Greenbook Distribution of Charitable Donations in 2004

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

For some deduction items, it was possible to identify eligibility on the host and sometimes also the donor dataset. Here are some examples.

Capital gains deduction	Received capital gain income
Child care expense deduction	Presence of children
Alimony deduction	Did not receive alimony income
Education amount for self	Went to school
Employee home relocation program	Received wages and salaries
Ontario college residence deduction / Manitoba shelter assistance	Full-time student in post-secondary post-education in Ontario or a home owner in Manitoba
Foreign income (as a percent to total income)	Total income must be non-zero
Foreign taxes paid as a percent of foreign income	Received foreign income

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 Child Care 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 RPP, RRSP for people without RPP contributions, and RRSP for people with RPP contributions separately.

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 Greenbook file. This requires a method of representing 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 over 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 were 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 and if 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.

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.

IMPUTATION

Using this complex set of distributional statistics generated from the Greenbook 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

Household expenditure data are intended to support simulations requiring information on shelter costs, simulations concerned with child care costs, and simulations of commodity taxes. Due to the limited number of records in SHS (about 15,000), it was decided to perform a consumption structure imputation using as many household categories as the data can support. A minimum number of observations are set 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 multivariate analysis. Households in one class are grouped according to the similarity of their consumption patterns, not their consumption levels. Variables used to define the consumption patterns are the 47 categories of expenditures, and some extra variables (e.g. savings and other expenditures).

The possible classification variables which are used to group similar consumption patterns are: household type, tenure, sex of household head, income class, age group, head working

full time, spouse working full time, region, presence of preschool aged children, and presence of school aged children are used as classification variables. The multivariate analysis evaluates the explanatory power of each classification variable and the best variables are used. The process iterates with new classification variables chosen until no more subgroups could be formed which contain a minimum number of observations. The first classification variable is forced to be household type in order to ensure 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.

Let us give an example to illustrate the process. The first classification split is forced to be household type. Then for each type of household, a regression is run in order to determine which of the remaining classification variables have the most explanatory power. So, for example, unattached individuals may have household tenure as the next classification level whereas married couples with kids might have the presence of preschool aged children as their next classification level. Unattached individuals who own their own home might then be split by sex whereas unattached individuals who rent their home might by split by age group. The splitting continues until a further split would cause the resulting bin to have fewer than the minimum number of observations, either in the SHS or in the SPSD. The final categories are then assigned the variable hdevmv.
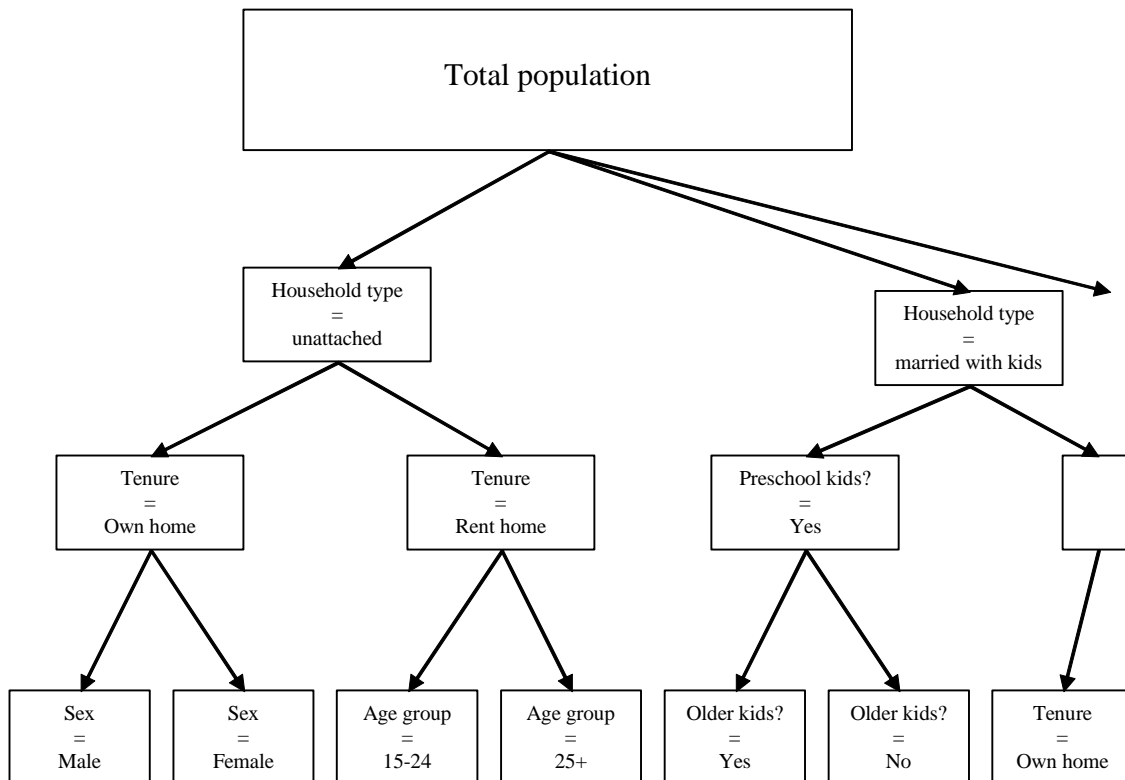
Figure 4: Illustrative example indicating how matching key is created

CATEGORICAL MATCH

Once the matching key is produced, the SPSD and the SHS can be split into the resulting bins. The next task is to assign an expenditure vector from the SHS to the SPSD households within these bins. A weighted duplication is performed on the SHS in order to ensure that every SPSD observation gets an SHS vector of expenditure. Within the bins, the observations on both datasets are sorted by total income. On average, there might be about six times as many SPSD 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, the 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. If the number is less than one, then it means that the SHS expenditure vector is not used in the matching process. 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 ith SHS household

j = the jth 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 SLID and SHS

weights can however create distortions in the matched distributions.

Other topics:

CHILD CARE EXPENSES

Child care expenses are imputed both from the Survey of Household Spending and from the T1 sample. The imputation from the SHS is independent of the total expenditure vector imputation. This was done so that more appropriate variables could be used to match similar records. The methodology used is similar as that used to impute the expenditure vector. The child care expenses from the T1 sample were first imputed in the same way as the other deductions imputed from the T1. For both types of child care expenses, the imputed expenditures were assigned to the children in the family.

IMPUTATION OF WEEKS WORKED TO EI CLAIMS

The EI status vector does not include weeks worked in the 52 weeks prior to an EI claim. The SLID PUMF on the other hand has weeks worked last year by individuals. In order to impute this variable onto each EI claim, a categorical matching process was used to match individuals from the SLID PUMF to individuals with EI histories on the SPSD. Individuals and claimants were matched by a combination of characteristics such as age, province, sex, industry, paid hours of work last year and child benefit received.

References

Adler, H.J. and M.C.Wolfson (1988), "A Prototype Micro-Macro Link for the Canadian Household Sector", The Review of Income and Wealth., series 34, no 4.

Ruggles, R. and N. Ruggles (1986), "The Integration of Micro and Macro Data for the Household Sector", The Review of Income and Wealth, series 32, no. 2.