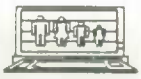
SPSD/M 

Database Creation Guide

The Database Creation Guide which describes in exhaustive detail each individual step in constructing the SPSD was not completed at the time of publication. In its place we have included an edited version of a paper delivered at the symposium on the statistical uses of administrative data, November 1987. The original title of that paper is "The Social Policy Simulation Database: An Example of Survey and Administrative Data Integration"

Table of Contents

1. Introduction 1

2. Objectives, Data Sources, and Techniques 3

3. The Host Data 7
   3.1 Suppression of Outliers 8
   3.2 Randomization 8
   3.3 Iterative Proportionate Adjustment (IPA) 9
   3.4 Splitting Database 10

4. Categorical Matching 11

5. High Income Adjustment 11
   5.1 Micro-Record Aggregation 12
   5.2 Categorical Match 13
   5.3 Evaluation 14

6. Unemployment Insurance History Imputation 16
   6.1 UI Donor Dataset 16
   6.2 Categorical Match 17
   6.3 Evaluation 18

7. Household Duplication 20

8. Stochastic Imputation of Income Tax Information 21
   8.1 The Donor Data 22
   8.2 Data Transformations 22
   8.3 Deriving Distributional Statistics 24
   8.4 Imputation 27
   8.5 Evaluation 27

9. Family Expenditure Survey Data Imputations 30
   9.1 Inflating the 1982 FAMEX 30
   9.2 Determination of Imputation and Matching Variables 31
   9.3 Categorical Match 34
   9.4 Evaluation 35

10. Conclusions 39

11. References 41

Table of Figures

Figure 1: SPSD Database Creation Process 7
Figure 2: High Income Adjustment Process 12
Figure 3: Capital Gains Pre and Post Duplication 15
Figure 4: Green Book Distributions of Capital Gains 23
Figure 5: Green Book Distribution of Charitable Donations 23
Figure 6: Capital Gains Distributions Pre/Post Imputation 28
Figure 7: Charitable Donations Pre/Post Imputation 29

Table of Tables

Table 1: Comparisons Between Matched UI Records 19
Table 2: Variables and Classifications for FAMEX Match 32
Table 3: Expenditure Vector Comparisons 37
Table 4: FAMEX Vector Match Income Comparisons 39

Abstract

This paper describes the construction of a prototype database explicitly designed to support analysis of personal income and sales tax and income transfer policies. Tax and transfer policies increasingly require integrated analysis that cuts across traditional jurisdictional and program lines. The Social Policy Simulation Database/Model (SPSD/M) was constructed to support micro-analytic modeling by combining individual administrative data from personal income tax returns and unemployment insurance claimant histories with survey data on family incomes and expenditure patterns. Considerable use of additional aggregate administrative data was made in both the database creation and modeling phases of the project. Input-output data were also applied in modeling 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.

The paper represents an intermediate draft. An initial "beta test" version of the SPSD/M was completed in the fall of 1987. The text of the current draft describes the construction of the SPSD for that version. We have since revised the database for final release to the general public, based on subsequent research, including that of colleagues in the Methodology Branch of Statistics Canada. The more important changes are described briefly in parenthetical italicized comments at the appropriate point in the text.

1. Introduction

In Canada, a few 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 ministry (including other federal ministries and provincial governments) have 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 both countries in the area of macro-economic policy where many agencies regularly provide independent analyses and forecasts.

The first commercial release of the Social Policy Simulation Database and Model (SPSD/M) later this summer will bring about a major change. With the SPSD/M from Statistics Canada, anyone will be able to 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 means microsimulation. In order 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 unemployment 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 current version of the SPSD/M is similar in this regard - the modeling software only does accounting calculations, though the capacity for behavioral modeling may be added in the future.

A matter of higher priority in the development of the SPSD/M has been to provide an integrated framework for tax/transfer analysis. At present, there are three federal ministries with major microsimulation capabilities: Finance for personal income tax, Employment and Immigration for unemployment insurance, and Health and Welfare for the Family Allowance and Old Age Security transfer programs. 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 unemployment 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, unemployment 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 paper 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.

2. Objectives, Data Sources, and Techniques

In developing the SPSD, every attempt has been made to maintain the variety and utility of the original source data while ensuring the non-confidentiality of these data so that the resultant database and model can be publicly released. Four central objectives thus 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 unemployment 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 child care 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 deduction.

These central objectives are highly interdependent and compromises among them have been made. 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 National Accounting, there has been a growing strand of concern about the lack of microdata foundations for macro-economic aggregates, for example in the writings of 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). So far, the 1984 SPSD is probably best considered as a prototype, but with biennial production as currently planned, the SPSD may well grow to be more closely linked to the National Accounts.

The SPSD has been constructed from four 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;

- unemployment insurance (UI) claim histories: a specially drawn one percent sample of histories from the Ministry of Employment and Immigration administrative system; and

- the *Family Expenditure Survey* (FAMEX): 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).

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 FAMEX), or in the form of summary tables (Taxation Statistics), or not at all (UI 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 the 1981 Census, Canada Assistance Plan (welfare) administrative reports, Statistics Canada's 1981 census, Vital Statistics, and Health and Welfare summary reports.

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

Iterative Proportional Adjustment (IPA) refers to a technique for reduction of bias by forcing agreement between data and known control totals. For example, survey weights may be adjusted to ensure that the population by age and sex represented by the survey corresponds to the "known" population by age and sex (i.e. based on census data).

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 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.

(Conversion is a technique that will be used in the next version of the SPSD to adjust for under-reporting of welfare and UI benefits. It is a method for adjusting based on the assumption of item non-response. The initial version of the SPSD was premised on all under-reporting being total non-response.)

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. We turn next to the main part of the paper where each step in the construction of the SPSD, as shown in Figure 1, is described.

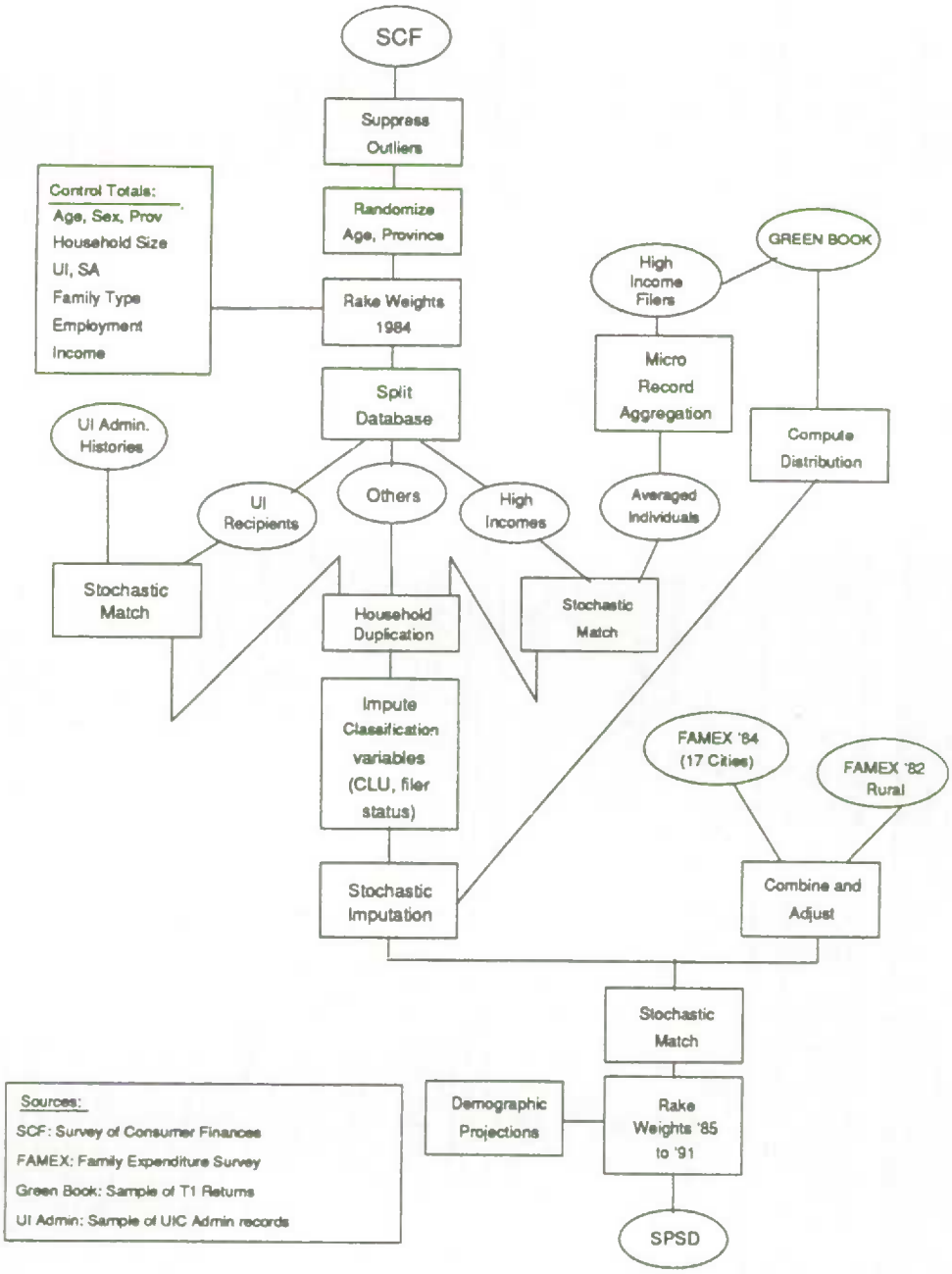


Figure 1: SPSD Database Creation Process

3. The Host Data

The target or "host" dataset is derived from the 1984 Statistics Canada *Survey of Consumer Finances* (SCF), 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 Household Income Facilities and Equipment (HIFE) form details the characteristics of the dwelling, and certain kinds of equipment contained in it. In 1984 the survey consisted of approximately 36,000 households containing 98,000 individuals.

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 being challenged by the SPSD. Our objective is a fully hierarchical database including individuals, census families, economic families and households.

The information from the UI, Greenbook and FAMEX 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 UI file (about 30,000). Records representing high income individuals (those with an income of over $80,000 in 1984) were duplicated to correspond to the number of high income records derived via micro-record aggregation from the Revenue Canada sample (about 5,000). 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. The resulting database contains over 170,000 records with a high proportion of the records representing households containing unemployed or high income individuals.

3.1 Suppression of Outliers

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 are deleted from the public release household file, and census families with more than four UI recipients or more than 6 earners are deleted from the public release census family file. The initial step in SPSD database construction was to suppress each household that met any of the SCF screening criteria (i.e. the criteria applied at the household, economic family, or census family levels).

In addition to suppression of entire households, 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.

3.2 Randomization

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.

Disclosure of the precise age-sex composition and location of a household may increase the risk of a breach of confidentiality. However, this risk will be considerably reduced by randomizing the ages of household members within five year age groups and by randomizing the sex of children (i.e. aged <= 15). This randomization, however, will not affect estimates of the costs or distributional impact of various child related tax/transfer programs.

Similarly, the geographical location of unusual household types may be 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).

(The independent randomization of the sexes of children and ages has proven to be problematic. For example, many cases of twins result, as do unrealistic age differences between parents and their children. Thus a somewhat different strategy will be followed in the next version. With 50% probability, all the sexes of the children will be "flipped". Also the same age shift factor, drawn uniformly from integers between -2 and 2, will be applied to all persons in the household, subject to some boundary conditions.)

3.3 Iterative Proportionate Adjustment (IPA)

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. This is not done at present because Statistics Canada public release databases are provided with separate weights at individual, census family, or economic family levels.

In order to provide this consistency, multi-level IPA was employed. The procedure is a generalization of the ordinary IPA (popularly termed 'raking', see Deming and Stephan (1940)). This specially developed multi-level IPA procedure was applied to the SCF to obtain individual level weights that are consistent with known age-sex control totals, and simultaneously unique to households. It may be thought of in terms of successive (proportional) adjustments to survey weights to bring them in line with pre-determined control totals. In multi-level IPA, the adjustments may be applied at household, family, and/or individual levels with an additional step which replaces individual (adjusted) weights within a household by the household average.

In addition the SCF also exhibits reporting biases which restrict its utility for modeling tax and transfer programs, for example:

- non-reporting of high-income individuals,
- under-reporting of social assistance (welfare) income, and
- under-reporting of investment income.

Using iterative proportional adjustment, the SCF record weights were recalculated to correspond to external control totals such as the number of high income (over $80,000 in total income in 1984) individuals, family size by province, private pension income and Social Assistance benefits by province.

The control totals employed in constructing weights for SPSD represented: (a) individuals by age and sex, (b) individuals by income class, (c) individual UI claimants, (d) households by family composition and labour force participation, (e) households by Social Assistance benefits, and (g) individual pensioners. Each of these control totals was disaggregated by province.

It has been shown (unpublished research by George LeMaitre, Social Survey Methods Division, Statistics Canada) that IPA adjustments of this sort lead to improved estimates of population characteristics. In particular, use of multi-level IPA with control totals provided only by population by age and sex produces estimates of family level characteristics with a 50% reduction in sampling variance compared with the principal person method currently employed. At the individual level, the IPA procedure results in a sampling variance that is essentially the same as is produced by current methods.

(Further research by George LeMaitre has suggested that the under-reporting of UI and welfare income is more likely due to item non-response. Thus IPA, which implicitly is based on an assumption of total non-response, is inappropriate. For the next version of the SPSD, IPA control totals will not include any reference to UI or welfare receipts. Instead, in a separate process, selected records will be identified as item non-respondents, and they will be "converted" to UI or welfare recipients. Appropriate benefit amounts are imputed in the course of conversion.)

3.4 Splitting Database

Splitting refers to a mechanical data preparation step that partitions the SCF (after suppression of outliers, randomization and IPA) into three mutually exclusive subsets: high income individuals, UI 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 UI recipients. There are, in fact, a handful of such cases but UI recipients in these households are treated as though they received no UI. High income individuals are those with incomes over $80,000 while UI 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).

4. 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 was 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 FAMEX data, UI data, and Green Book income data for high-income recipients. The technique allows the preservation of inter-item correlations from the donor record. Each of the matching procedures is described more fully below, where it is also noted that these categorical matches virtually preclude the possibility of an exact match.

5. 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.

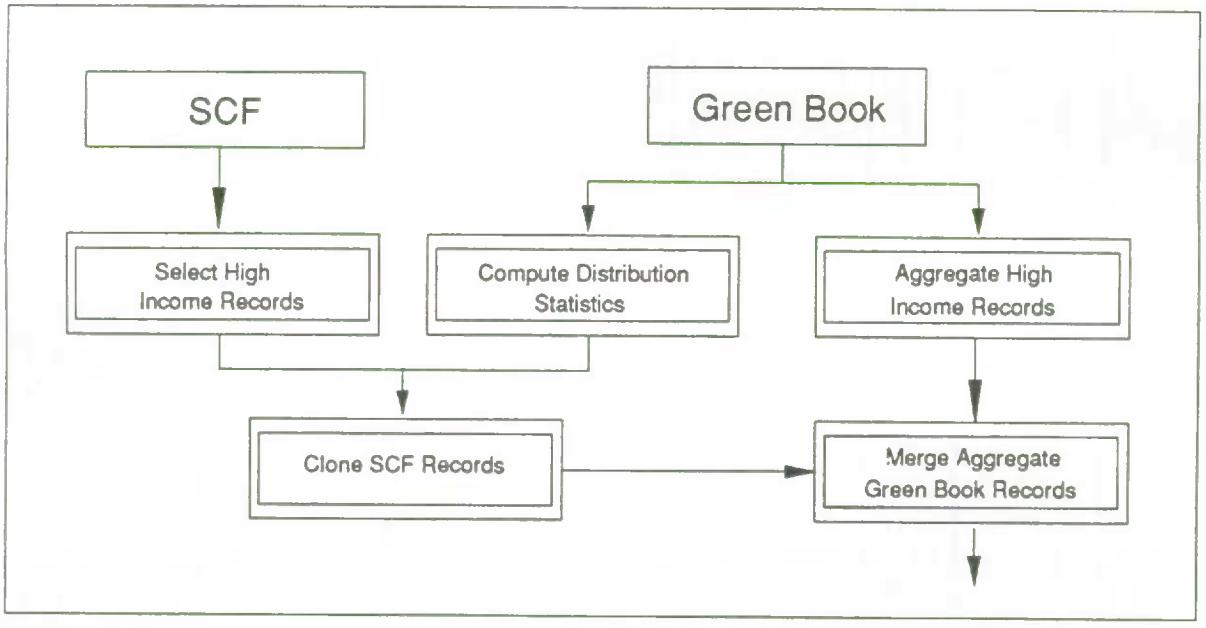


Figure 2: High Income Adjustment Process

5.1 Micro-Record Aggregation

Non-reporting by high-income individuals in the SCF is ameliorated by using the Green Book counts for individuals with income over $80,000 as an IPA margin. The IPA then increases the weights of each high-income record on the SCF so that the sum of the weights corresponds to the Green Book.

There are approximately 300 such records. The IPA 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, or five-tuples, 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 4,676 pseudo microdata records representing 24,556 Green Book Records, in turn representing 133,650 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.

5.2 Categorical Match

The original 300 SPSD records are duplicated to match the number of aggregated Green Book high income records (4,676). These 300 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.

5.3 Evaluation

Although this method of micro-record aggregation assures that correlations between the income and deduction items (shown in Table 1) are generally maintained, the univariate distributions of the synthetic records tend to have less variance than the original Green Book records. This is a result of the aggregation of several records into one. Very often, for sparse items such as Allowable Other Years Capital Losses, the five records to be aggregated contain several zeros which are included in the average. The average is maintained but the distribution tends to be less dispersed.

Figure 3 provides an example of the distortion in the distribution of Capital Gains introduced by this method. In effect, five-tuples of individuals were all attributed small values for Capital Gains instead of four with zero values and one with a higher value.

(In the next version of the SPSD, a change in method should address this problem. Instead of using a simple arithmetic average over the 5-tuples of high income individuals to create one pseudo microdata record, a weighted average will be used. One of the five actual records will be chosen at random to receive a weight of 80%, while the other four records will each be given a weight of 5% in computing the weighted average. This method is thus effectively a mixture of drawing a subsample (80%) and blurring by simple averaging (20%).)

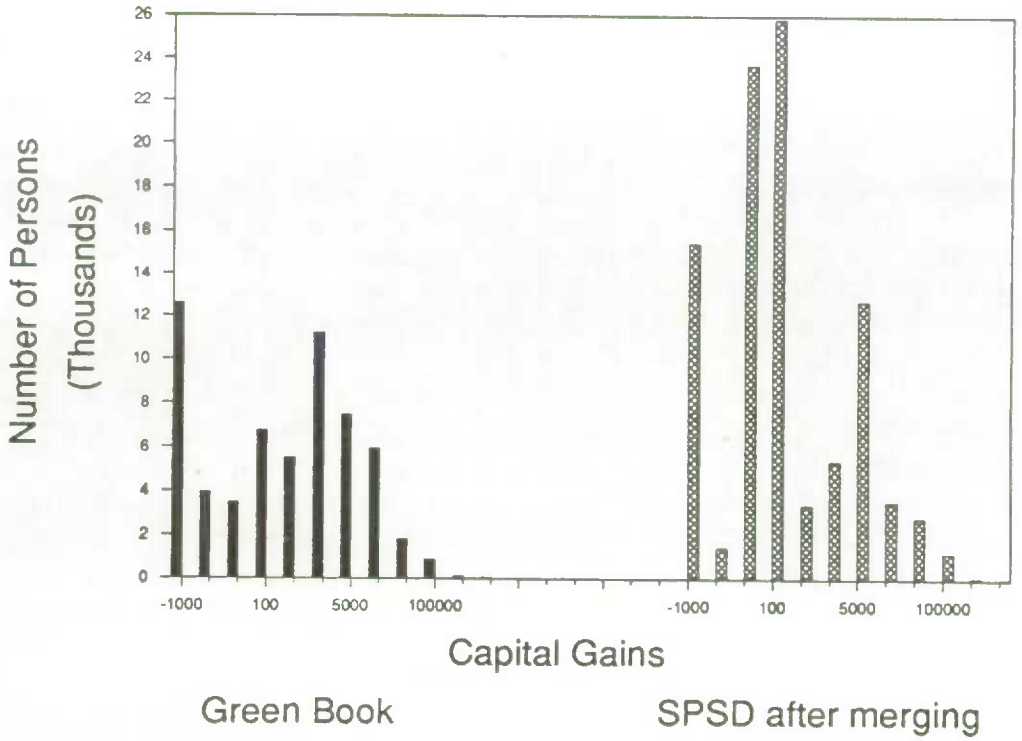


Figure 3: Capital Gains Pre and Post Duplication

6. Unemployment Insurance History Imputation

Unemployment Insurance (UI) 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 UI 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".

UI income is an important component of both disposable and taxable income. On its own, UI income and simulated variants 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 UI simulation module and identification of a limited set of "program relevant" UI variables (Table 2) that could serve as input to the UI simulation module.

6.1 UI Donor Dataset

The UI administrative histories imputed to SPSD were based on a 1% sample of administrative records from the population with some UI claim activity within the 1984 calendar year. The sample consists of about 30,000 individuals and represents about 40,000 claims. The content of this dataset was specially designed. On one hand, it had to be rich enough to capture the weekly labour force history relevant to application of UI program regulations. On the other hand, it had to be compact and general enough to be non-confidential. This was accomplished by thinking in terms of an event history, so that the durations of various activities became the focus rather than weekly activity records. The staffs of Employment and Immigration Canada and of the Forget Royal Commission were helpful in designing this dataset. The following list shows the set of variables employed as input to the UI model.

<u>UI History Variables</u>

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)
Weeks of Work (prior to current claim)

Average Weekly Earnings (prior to claim)
Penalty for Voluntary Quit (weeks)
Week Claim Established
Benefits Paid in Calendar Year (1 or 2 claims)
Weeks of Benefits Paid in Calendar Year

Because of the interrelatedness of these UI claim history variables, each of the 30,000 claimants' records (which may consist of one or two claims) was categorically matched to SCF records which had some reported UI income in the calendar year. In addition to the UI claim history variables identified above, administrative data on claimant age, province and sex are used as matching keys. These same variables were available on the host dataset for individuals with UI income.

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 host dataset by (i) identifying UI recipients aged 65+ (retirement benefits), (ii) identifying UI recipients with occupation coded as "Hunting, Fishing, Trapping" (fishing benefits), and (iii) identifying female UI recipients with a child aged 0-1 (maternity benefits). No distinction could be made between sickness and regular benefit types on the host dataset.

6.2 Categorical Match

Matching was carried out by first partitioning the donor administrative (UI) 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 UI and host datasets had equal numbers of records. If in any given cell the number of host records exceeded the UI records, then the UI records were uniformly duplicated (UI data were a simple random sample). Correspondingly, if the number of UI 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 host 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 UI claimant population. Initially, the host dataset contained 10,381 such records, while after duplication there were 31,585 records. This expansion of the dataset was intended to ensure full use of the UI histories available from the 1% sample.

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

6.3 Evaluation

Table 1 provides an indication of the success of the match. The correlation between benefits reported on the SCF and the corresponding (matched) benefits from the donor UI dataset indicates the 'accuracy' of the match, since benefit ranks rather than benefits per se were used in the match. Difference quartiles represent the 25%, 50% and 75% cutpoints for the distribution of differences between SCF and UI benefits. These differences might be interpreted as 'errors' resulting from the substitution of UI benefits from administrative data for those reported (and imputed via 'conversion') on the SCF.

Table 1 - Comparisons Between Matched UI Records

Distributions by Province & Sex and for Canada
(i) n - Number of Pre-duplication Records
(ii) r - Correlation Between Host & Donor UI Benefits ($)
(iii) Difference Quartiles - [Host($) - Donor($)]

Province/Sex	n		r	Difference Quartiles		
	Host	UI		25%	50%	75%
NFLD-Male	795	929	0.953	-192	140	417
Female	445	549	0.925	-270	-14	232
PEI- Male	241	246	0.631	-1,159	-290	789
Female	210	213	0.871	-363	11	531
NS- Male	496	787	0.931	-271	45	528
Female	294	528	0.919	-197	-38	147
NB- Male	604	798	0.941	-531	-45	589
Female	390	573	0.905	-102	158	669
QUE- Male	1,116	5,471	0.970	-162	86	341
Female	784	3,961	0.958	-112	103	324
ONT- Male	787	4,990	0.960	-149	36	207
Female	687	3,837	0.953	-110	74	306
MAN- Male	343	611	0.932	-360	-69	294
Female	272	508	0.866	-115	-49	496
SASK- Male	369	548	0.918	-239	231	489
Female	283	394	0.954	-83	75	311
ALTA-Male	691	1,648	0.946	-88	68	448
Female	482	1,072	0.951	-174	16	264
BC- Male	625	2,281	0.953	-112	186	470
Female	467	1,638	0.954	-185	68	461
CANADA	10,381	31,582	0.953	-155	69	352

In most cases, the differences between benefits reported on the host dataset and benefits sampled from UI administrative data are relatively small. Discrepancies as large as 255 dollars may be expected, since they could represent a UI benefit payment for a single week (i.e. the

minimum discrepancy in benefit weeks). Moreover, the differences are small in comparison to median benefit levels, which were $2,972 for males and $2,050 for females, at the national level.

It is expected that some host data may represent biased responses and that others may contain benefit components not included in the UI data or model (e.g. training allowances). To the extent this is the case, then error in the host dataset would make an important contribution to the benefits differences.

Correlations are high, except in PEI where little gain from duplication was possible. In the absence of substantial duplication, age and claim type matching constraints will reduce marginal correlations in benefits.

High correlation is not a necessary consequence of the matching technique. Matching of corresponding ranks guarantees a monotone association, but not necessarily a strong linear association. This use of ranks can be interpreted as matching corresponding quantiles of independent samples. Thus, a strong linear association indicates that the two samples (host and donor) are from similar density functions.

Further direct evaluation of the results of the match is difficult, since essentially all common factors between datasets have been employed in the match. The UI data provide an extension and replacement of host data in which UI variables are unbiased and consistent with the UI program structure.

7. 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 UI 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 UI 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 UI 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 probably most are most likely to resemble the institutional population. The weights on these records are adjusted to reflect estimates of the institutional population by age, sex and province (based on administrative statistics on institutional bed days).

8. 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 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.

8.1 The Donor Data

The source data for the imputation were derived from a Revenue Canada sample of 1984 Individual Tax Returns. This contained 2.4 percent of all returns (380,419 returns), the same sample used to compile the *Taxation Statistics* (the Green Book) publication. 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 1984 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).

8.2 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. Figure 4 provides an example of the difference in capital gains between two income groups. A comparison of charitable donations between the same groups is provided in Figure 5.

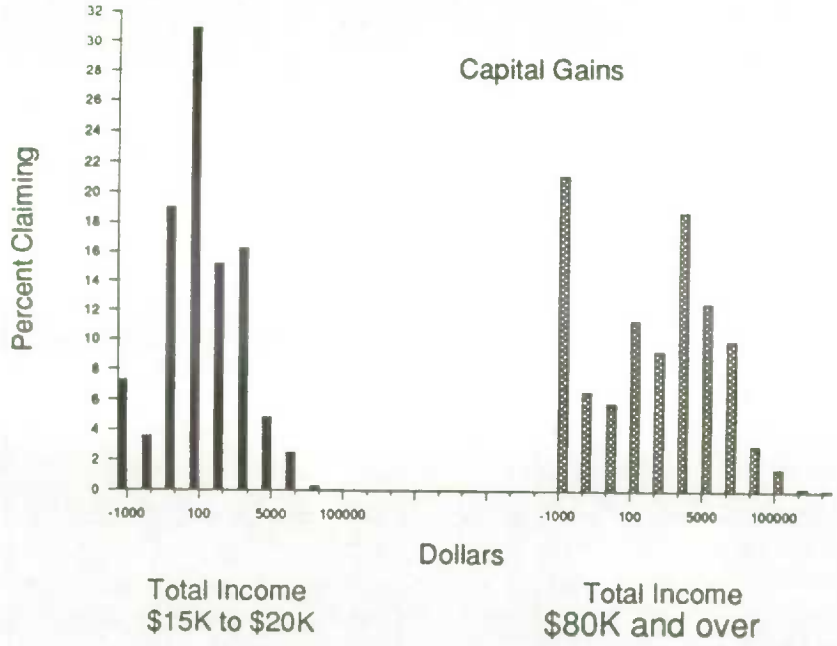


Figure 4. Green Book Distribution of Capital Gains for Two Income Groups

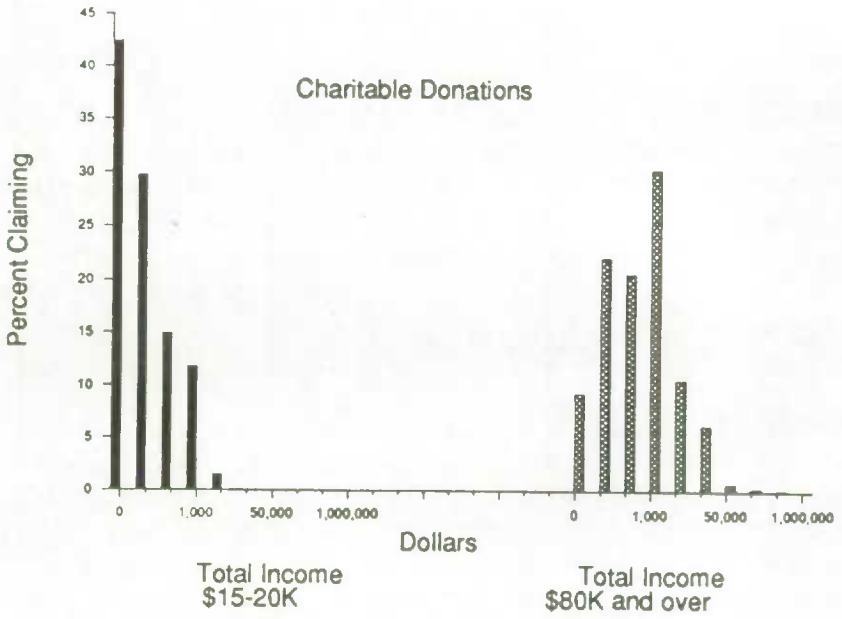


Figure 5. Green Book Distribution of Charitable Donations

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 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 ($3,500 in 1984). 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.

8.3 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 in section 8.2 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 sub- ... sub-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.

8.4 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 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.

8.5 Evaluation

Figures 6 and 7 provide some examples of results of the imputation process. These are both aggregated to the level of the entire sample.

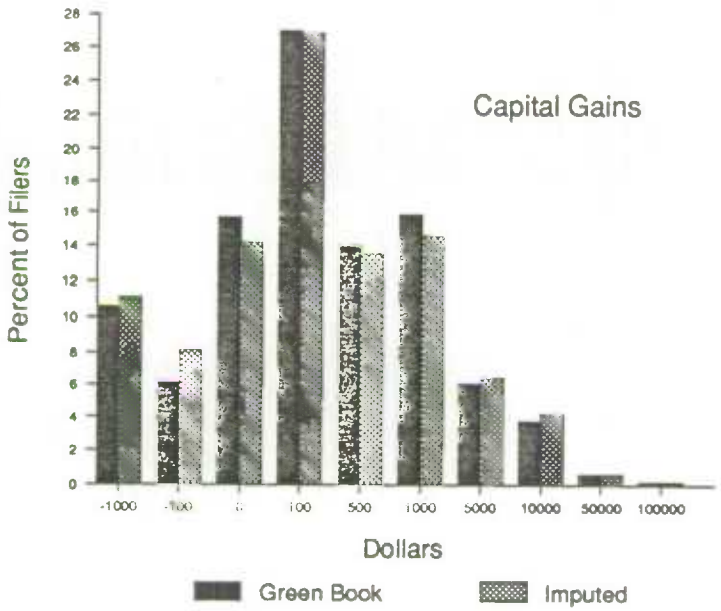


Figure 6: Capital Gains Distributions Pre/Post Imputation

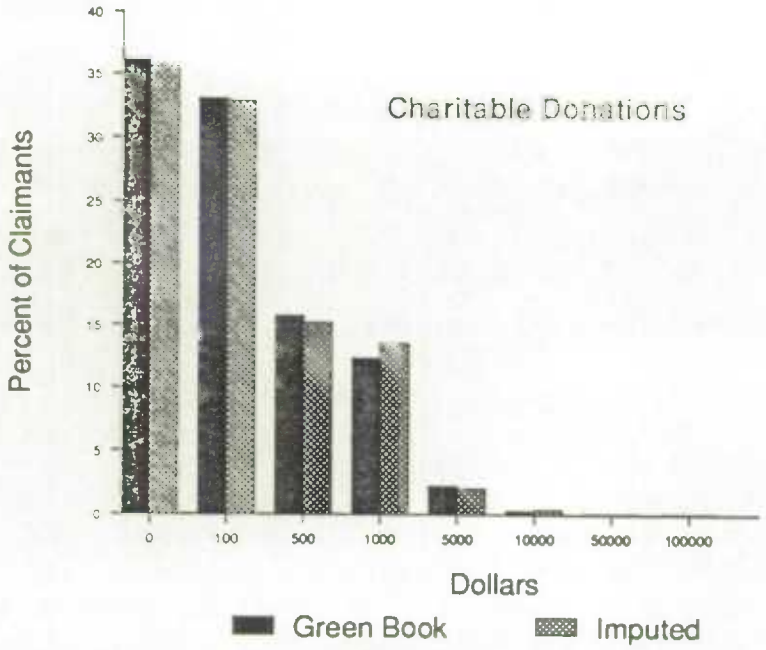


Figure 7: Charitable Donation Distributions

This method of imputation tends to use the full richness of the source data in regenerating plausible distributions on the host dataset. Overall distributions make sense but often individual cases do not. For example, since capital gains are imputed according to total income, age, sex and province, it is not impossible for a social assistance recipient to be imputed capital gains. In this case, the social assistance recipient is treated exactly the same as a retiring farmer who has sold his farm and has received several hundreds of thousands of dollars in capital gains.

Another problem with this method is that joint distributions are lost to the degree that they are not accounted for by the classification variables. In simple cases, most deduction items are well correlated with income and income is normally an important classification criterion. Where the inter-correlation between items (e.g., RPP and RRSP Contributions) is more important than their correlation with income, the correlations are lost unless the method is modified.

One outcome of the loss of correlation between deduction items is that the individuals, especially the high income group, on the host dataset appear not to be optimizing their tax situation. Since the high income group consists of all individuals with income over $80,000, an individual with a total income of $90,000 has the same probability of being imputed a million dollar deduction as another person with $2 million in income.

(In the next version of the SPSD, this problem will be addressed by applying the micro record aggregation method to impute more deduction items for the high income group. This would have the effect of preserving correlations between income and deductions as well as correlations among deductions.)

9. Family Expenditure Survey Data Imputations

The family 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 on the family expenditure dataset (about 10,000), it was decided to perform three separate synthetic matches. This allowed for a specific tailoring of the classification categories to the nature and determinants of the vector of expenditure items to be matched. For example, a household's expenditures on child care depends substantially on the number of children and the labour force status of the parents, and as such these should be the primary classification variables in any match. On the other hand, shelter costs are more strongly correlated with the number of rooms and residential tenure; a classification by number of children would do little to improve this match.

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

- Construction of a National 1984 FAMEX database
- Selection/Grouping of expenditure items for imputation
- Selection/Construction of Matching Variables
- Categorical Matching (Weighted Duplication)

9.1 Inflating the 1982 FAMEX

The family expenditure survey was last conducted for all Canada in 1982. The 1984 survey, which matches the time frame of the SCF host dataset, was restricted to a 17 city sample. The first step in matching was to prepare a 1984 "all Canada" version by inflating the values of the 1982 non-17 City records to 1984 and then adding these pseudo 1984 observations to the set of actual 1984 observations. All money items on a given 1982 family expenditure record were "grown" by using the same inflator rather than using specific CPI and income inflators. This simple process was dictated by the requirements of the commodity tax model where a complete accounting identity of a household's income, expenditure and saving patterns must be maintained.

This approach assumes that expenditure patterns of (non 17-city) households remain constant and thus avoids implicit assumptions of behavioral response to price fluctuations. This assumption was supported by an analysis of shifts in the proportions of total expenditures spent on individual items between the 1982 and 1984 17-City samples. The differences between proportions remained within one percent for all categories of expenditure. The largest differences were a one percent increase in mortgage interest as a percent of total expenditures (4.7 to 5.7), and a 0.6 percent decrease in automobile and truck purchases (5.4 to 4.8).

The inflators were computed specifically for each non 17-city household record on the 1982 FAMEX. They were based on average growth of income by family type for each of six income sources as reported on the Survey of Consumer Finances. A household's inflator was calculated as a weighted average of the six individual growth rates for their household type where the weight was the proportion of income received from each source.

9.2 Determination of Imputation and Matching Variables

Table 2 summarizes the variables for imputation as well as the matching variables used in each of the three categorical matches. The figures in parentheses represent the number of classification levels.

Table 2: Variables and Classifications for FAMEX Match

	Shelter(126)	Child Care(36)	Expenditure Vector(390)
Imputed Variables	Rent Mortgage Interest Property Taxes Insurance Premiums Utilities Repairs Other Shelter Costs Value of Home Balance on Mortgage	Child Care Expenses	"Savings" Other Money Receipts Household Income Account Balancing Difference Expenditure Vector (50) (See Appendix A)
Matching Variables	Residential Tenure Number of Rooms Urbanization Geographic Region Household Income	Family Type Employment Status # Children (0-4) # Children (5-15) Household Income	Income (discrete 6) Family Type(5) Residential Tenure(3) Age of Head(4) Sex of Head(2) Geographic Region(5) Family Size(2) Number of Children(3) Urbanization(2) Income (continuous)

The variables for the shelter match were selected and grouped so that estimates of major shelter costs and imputed rent could be made. The chief intended use was for modeling social assistance (welfare) payments, and secondarily for use in modeling tax credits provided by some provinces. The high level of aggregation reflects the coarse way in which social assistance can be modeled due to the lack of other data relating to eligibility and benefit levels. For example, welfare benefits may depend on asset eligibility tests or fire insurance or both, while FAMEX reports nothing on total assets and only total home insurance.

Child care costs are composed of day care costs inside or outside the home as well as kindergarten tuition fees. This definition is intended to follow current federal legislation regarding the child care expense deduction. No attempt has been made to exclude costs that may be

disallowed for tax purposes due to the absence of receipts. Other items such as infant's clothing or other variables which may be desired when modeling an expanded definition of costs are not imputed.

The third and most ambitious match is designed to support modeling of commodity taxes. This is particularly relevant in the current policy context because major reform of federal and provincial sales taxes is under discussion. The selection and grouping of FAMEX income and expenditure variables for the expenditure vector was based on the structure and composition of the personal expenditure dimension of the Canadian medium level aggregation input-output tables and the requirements of the commodity tax model. Expenditures having some indirect taxes and duties were placed in the corresponding input-output personal expenditure category. Variables not having an indirect tax, or an indeterminate indirect tax were placed in a residual category (e.g. real estate commissions).

Additional variables were also included in the vector (e.g. income, taxes, savings) in order to complete the basic household accounting identity - income equals expenditure plus saving. In turn, this allows various simulation options - for example the allocation of a change in disposable income between saving and consumption. A number of conceptual differences between FAMEX and the system of national accounts on which the input-output tables are based still remain. Nevertheless, as shown in Adler and Wolfson (1987), the SPSD and National Accounts household sector estimates for 1984 are in reasonably close accord.

The determination of matching variables was restricted by the availability of similar variables on both the host and donor datasets. From this limited set, individual analyses were conducted to determine the optimal selection and configuration of the matching variables for the three matches. The techniques used to identify variables included correlation, factor analysis and difference of means tests. Four main interdependent criteria guided the selection and creation of matching categories or bins:

Expenditure Levels: The variables used for classifying households should be highly related to both the level of expenditure as well as the distribution among specific commodity elements.

Expenditure Categories: The bins should be created in such a way as to restrict the attribution of costs to appropriate populations. For example, childless couples should not have child care expenses and unattached women should not have large men's clothing expenses.

Reporting Categories: The bins should reflect to as great a degree as possible the categories that will be used in final reporting. For example the SPSD and model are likely to be used for comparative analysis of different provinces and regions, different levels of income, and different family types, so these variables should be used in the matching process.

Sample Size Within Bins: When creating the bins, it was judged that both the host and donor databases should have at least five observations in any bin. This practice was adopted to prevent the maximum number of duplications of FAMEX records from being too large. The last matching variable in all cases, income, was used to rank all the records in a given bin in both the host and donor datasets. Since a fair number of bins contained very large numbers of records, the final sort on income was often a key element of the fit for all three matches.

The subsequent likely analytical uses of the data (e.g. tables by province, income group, or family type) were taken into account in creating the final matching and binning categories. The hierarchical organization of the variables was constructed manually in a flexible asymmetric manner that allowed for different breaks for different types of bins, or even different variables. Thus for shelter costs at the second level of the hierarchy (number of rooms), homeowners with or without a mortgage were classified into groups of less than 6, 6-7, and 8 or more rooms while renters were in groups of less than 4, 4, and five or more rooms.

9.3 Categorical Match

The categorical match of records was performed at the household level and required only the duplication of FAMEX records. In order to make the fullest possible use of the FAMEX data without having to duplicate SCF records, matching bins were created in such a way as to ensure that the FAMEX 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 FAMEX counterpart. The match took the form of a weighted duplication of FAMEX records, and was designed to force the FAMEX sample counts within bins to match the corresponding host bin.

The general task in this weighted duplication procedure is to increase the number of FAMEX 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 FAMEX records. However, it would be inappropriate simply to make five clones of each FAMEX record because this would in effect treat the FAMEX as a simple rather than as a stratified random sample; no account would be taken of the FAMEX sample weights. Instead, those FAMEX records with higher weights are cloned proportionately more than those with smaller weights.

More precisely, a weighted probability of occurrence of FAMEX 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 FAMEX 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 the FAMEX dataset,

at least one match is assigned to every FAMEX record, and then the probability is multiplied by the difference between the sample sizes of the host and donor bins. In other words, every FAMEX household is given at least one match and the number of duplications still required to hit host bin size are distributed across the FAMEX 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}).

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

Where: i = the i^{th} FAMEX household
j = the j^{th} matching bin
W = the weight of the FAMEX donor record
N^h = the sample size of the SPSD host bin
N^d = the sample size of the FAMEX donor bin

Each FAMEX 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 FAMEX 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 FAMEX data, at least until SPSD weights are associated with it. The difference between the SCF and FAMEX weights can however create distortions in the matched distributions.

(The different treatment of child care expenses in the tax system has resulted in a need to perform an additional imputation. This will involve the allocation of household expenditures on child care expenses to each child in the household. The allocation will be based on a regression model of that distribution.)

9.4 Evaluation

Several tests to assess the quality of results and assist in subsequent analysis were performed. The distributions of the aggregate expenditures are extremely similar before and after matching. The only real sources of distributional and aggregate difference are attributable to the different (SPSD) weights now associated with the FAMEX data and the minor impact on FAMEX weights of imposing a minimum duplication of one. Benchmark control totals for most expenditure data are not readily available. As such the central test for these aggregate totals was

how the post-match totals compared to the FAMEX totals. The differences between the individual item totals imputed during the shelter and child care matches were all within five percent. Table 3 presents the results of the expenditure vector match.

Table 3: Expenditure Vector Comparisons, Selected Items

Income/Expenditure Category	FAMEX $ Millions	SPSD/ FAMEX	SPSM/ FAMEX
Food & Non-Alcoholic Beverages	30,805	3.10	
Alcoholic Beverages	4,959	0.38	
Tobacco & Related Products	3,453	3.69	
Men's and Boy's Clothing	4,462	-0.21	
Gross Imputed Rents	19,021	-5.36	
Gross Paid Rent	12,773	4.17	
Electricity	4,226	1.26	
Other Fuels	2,115	8.89	
Durable Household Appliances	3,292	-0.55	
Semi Durables	3,627	-0.98	
Non Durable	4,301	1.22	
Domestic Services	1,121	-9.25	
Other Household Services	2,000	0.40	
Medical Care	1,381	1.65	
Hospital Care	86	-3.46	
Drugs & Sundries	1,657	0.87	
New & Used Automobiles	10,014	-1.53	
Auto Parts & Repairs	4,458	3.63	
Purchased Transportation	3,086	1.43	
Communications	3,583	1.98	
Recreation,Sports, & Camp Equip.	7,514	-4.28	
Books, Magazines, & Stationary	2,261	1.50	
Recreational Services	4,412	0.66	
Jewelry, Watches, & Repair	1,033	-3.41	
Personal Care	2,333	-0.49	
Union & Professional Dues	985	2.75	1.80
Personal Taxes	45,148	-5.19	14.77
Unemployment Insurance Premiums	2,924	0.81	17.25
Retirement Pension Payments	6,108	0.44	18.88
Unallocated FAMEX Items	2,525	8.99	
Net Change In Assets/Liabilities	16,021	-5.49	
RRSP Contributions - Total	3,492	-7.56	36.76
Other Money Receipts	5,612	4.14	
Account Balancing Difference	1,245	-17.20	
Spending Unit Total Income	272,714	-0.87	6.99

Table 3 shows the relationship between the aggregate totals for FAMEX, SPSD, and SPSM modeled variables. The second column shows percentage differences between the pre- and post-matching value of the FAMEX items. As can be seen all of the totals for variables are within a few percent, the differences being largely attributable to the SPSD weights associated with the FAMEX expenditures. Account Balancing Differences are 17.2 percent smaller due to the fact that they are not an actual expenditure but the discrepancy between a family's receipts

and disbursements. The third column shows the percentage difference between the FAMEX data and the SPSM modeled and/or imputed variables. The larger differences are due to corrections for underreporting that have been made through the imputation of Green Book distributions.

The degree to which a FAMEX record was duplicated averaged 6 times for all three matches. The maximums of FAMEX household duplications were 28, 42, and 51 for the shelter, child care and expenditure vector matches. In 75 percent of expenditure vector matches the duplication was less than 12.

The correlation of host and donor incomes was high with values of .91 and .96 for shelter and child care imputations. The correlation is inversely related to the number of bins because of the final sort on income. For this reason, the expenditure vector match resulted in a weaker correlation (.86) of FAMEX and host dataset income.

Table 4 below gives a further indication of the quality of the match between the FAMEX and the host dataset (at this point most of the way along in its transformation from the SCF to the SPSD). The SPSD was first sorted by income and divided into quintiles. In addition, the top quintile was subdivided at the $80,000 level, corresponding to the point where the special high income imputation based on income tax records started. These six income groups correspond to the rows of the table. Then, within each SPSD income group, records were sorted in ascending order of the ratio of SPSD income to the income on the FAMEX record with which it has been matched. The levels of these ratios have then been displayed at various percentile cut-points. For example, the .887 figure in the second column and second row indicates that in the second quintile SPSD income group, 5% of the records (always in weighted terms) had a ratio of SPSD to FAMEX income less than .887.

The differences in incomes tended to be the greatest at the tails of the distribution where the most change had been caused in the host distributions by IPA and high income adjustment. This fit is especially important to understand because of its effect on various commodity tax model options as well as the a priori relationship between income and expenditures. Overall, in 90 percent of expenditure vector elements imputed, individual household income differences were within fifteen percent (see bottom row of Table 4).

Table 4: FAMEX Expenditure Vector Match Income Comparisons

SPSD Income Quintile or Group	Percentile Cut-Points in the SPSD to FAMEX Income Ratio Distribution						
	1	5	25	50	75	95	99
1	0.010	0.5985	0.918	0.991	1.055	1.319	1.664
2	0.845	0.887	0.954	0.989	1.021	1.078	1.129
3	0.898	0.938	0.980	1.003	1.030	1.074	1.094
4	0.916	0.947	0.978	1.000	1.022	1.072	1.101
5 excl >$80,000	0.855	0.900	0.961	0.998	1.037	1.130	1.207
>$80,000	1.003	1.014	1.074	1.181	1.418	2.130	3.418
All	0.555	0.866	0.965	0.999	1.035	1.154	1.572

In all quintiles the median ratio difference between pre- and post-matching incomes was within one percent except for the over $80,000 group. This is because the maximum income on FAMEX is on the order of $250,000 while the maximum on the SPSD is about 11 million due to the high income adjustment. Certain commodity tax model options attribute indirect taxes based on dollars of imputed expenditure, and as such the relationship with income should be close.

10. Conclusions

The Social Policy Simulation Database and Model (SPSD/M) as just described is a work in progress. We are now in the midst of a third iteration in building the database and refining the model software, this time for the first commercial release by Statistics Canada.

In order to test the viability of the SPSD/M idea, it has been necessary to forge ahead often by making simplifying assumptions. The basic view has been that it is better to have a working, testable product with limitations sooner rather than a better version always under development.

The process of developing the SPSD/M has already had some valuable spin-offs. These have included suggestions for the providers of the source data sets which are now being implemented, such as revisions in the weighting system for the monthly Labour Force Survey. Furthermore the model has produced results that have already been useful in several instances of policy planning in Canada. These include the 1985 federal Royal Commission examining the unemployment insurance system, a special Ontario task force reviewing social assistance, published analysis of the 1988 federal income tax reform, and projections of the impact of Canada's aging population on the fiscal structure of the federal government.

Many methodological refinements of the database creation process have been implemented in order to adjust for gaps and inaccuracies in the data. Further improvements are of course possible, and will continue to be made as the SPSD/M matures to become an ongoing product of Statistics Canada.

11. 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.

Deming, WE and Stephan, F.F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known", Ann. Math. Statist. 11, pp 427-444.

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,