EVOLUTION OF THE PERSONAL INCOME TAX SAMPLE OF THE QUEBEC MINISTRY OF REVENUE

Claude Boivin, Robert Paré, Québec Ministry of Revenue

INTRODUCTION

The Québec Ministry of Revenue (QMR) exists in its present form since 1961 and, since 1974, has published a set of personal income tax statistics based on a sample of income tax returns for each fiscal year.

The sample selection of taxpayers for the 1983 taxation year featured a radical change in comparison to the preceding eleven years. The functional analysis of an "ideal" statistical system (Fellegi [1]) served as a framework for using data from administrative records to produce statistics. The purpose of this paper is to describe the adaptation and implementation of some of the functions of the "ideal" statistical system to the particular environment of the QMR.

The first section includes a brief description of the administrative file from which the sample of taxpayers was drawn (this sample is called the Personal Income Tax Sample). The second section summarizes the sampling design used until 1982 (taxation years 1972 to 1982) and explains the reasons behind the major modifications applied to the 1983 sample selection. The third section summarizes the functional analysis of the "ideal"statistical system and describes the first applications carried out at the QMR. Finally, the last section describes the expected short-term developments of the statistical system of the QMR.

1. THE PERSONAL INCOME TAX SAMPLE

Each year, a data file called the Personal Tax Return Master File is opened in order to collect information from assessed income tax returns. In addition to containing a large amount of information transcribed from the tax returns, this Master File also includes data from the Assessing system. For example, edited tax fields and details about the correspondence exchanged between taxpayer and the QMR may be found on this file. The Personal Tax Return Master File includes three types of records: original assessed returns, amended returns and cancellations.

The population is defined as all taxpayers filing for a specific year. Thus, all the income tax returns filed for 1983 constitute the population for this taxation year. These returns were, for the most part, processed and added to the Personal Tax Return Master File between March 1984 and the beginning of February 1985, when the final version of the Master File was produced. However, some 1983 returns are recorded in the 1984 Master File, either because their assessment necessitated several delays, or because they were received late. Similarly, the 1983 Master File contains income tax returns for previous years. Since our target population consists of all tax returns filed for a specific taxation year, the sampling frame (the Master File) necessarily differs slightly from this.

Using the Personal Tax Return Master File for statistical requirements would create a number of problems because of the file's content of over three million records with more than three hundred variables. In 1974, the Personal Income Tax Statistical System was implemented in order to obtain a sample of taxpayers from the Master File and to ensure that requests for various types of information could be processed and forwarded within a reasonable time limit. The three major goals of this system were:

1) - to produce statistics on individuals who filed an income tax return for a given taxation year. These statistics included the distribution of different types of income as well as the deductions and exemptions used. They are published annually, with a two year lag, and the first edition was available in 1974 (1972 tax year);
2) - to obtain records from the Personal Tax Return Master File for auditing purposes;
3) - to permit simulation studies in order to evaluate the impact of proposed fiscal changes. Before the 1983 taxation year, the sample was built continually between March and February, a new selection being made each time a batch of records (tax returns) was appended to the Master File. In 1983, the sample was drawn only when the final version of the Master File was ready.

2. HISTORY OF THE PERSONAL INCOME TAX SAMPLE

Each Personal Income Tax Sample drawn from 1972 to 1982 was a stratified random sample. This section describes the major changes made to the stratification up to the 1982 taxation year.

In 1974, a first sample of taxpayers was selected from the 1972 Personal Tax Return Master File. Stratification was based on the type of income tax return used (Long or Short), the size of the municipality where the taxpayer resided and the taxpayer's total income. Each type of return was subdivided into 15 strata according to three categories of municipality size and five income

levels. This first stratification had, therefore, thirty strata (2 types of returns by 3 categories of municipality size by 5 income levels). The long returns were completely enumerated while the short returns were selected with various sampling rates. The overall sampling rate for the short returns was 10%.

The first change in stratification occurred during the 1974 sample selection. The type of return was discarded as a criterion and was replaced by the type of taxpayer (business or non-business taxpayer). A business taxpayer was defined as a taxpayer having at least one of the following types of income: business income, professional income, commissions, fishing or farming income. All other taxpayers were included in the non-business group. The other stratification criteria remained the same. Business taxpayers were completely enumerated and non-business taxpayers were selected with various sampling rates.

In 1976, a fourth category of municipality was added; moreover, business and non-business taxpayers were classified according to new levels of income. Thus, ten strata of business taxpayers and 20 strata of non-business taxpayers were defined.

In 1977, the criterion "rental income over $10,000 " was added to the definition of business taxpayers and the number of strata thus increased from 30 to 31.

To ensure a better representation of taxpayers with a low business income but with a high total income, two additional strata were defined in 1978 and the number thus grew to 33.

A complete revision of the sampling design was carried out for the 1980 taxation year. The number of strata exploded with a fourfold increase (33 to 122), and the overall sampling rate (14%) was slightly greater than in 1979 (13%). The new criteria for the stratification were:

(A) the regional office (Montréal or Québec City);
(B) the type of taxpayer (business or non-business);
(C) the source of income:
- non-business taxpayers of each regional office were stratified according to four principal sources of income subdivided in eleven levels of income. 88 strata were thus formed;
- business taxpayers were classified into 17 categories for a total of 34 strata (17 strata for each regional office).

Figure 1 summarizes the changes that occurred between 1972 and 1982 regarding the population and sample sizes, the overall sampling rate and the number of strata. It also indicates modifications to the sampling design in any given year. It can be seen that, since 1976, the sampling rate stabilized between 12% and 15% with the sample sizes ranging between 400,000 and 500,000.

The large sample size reflected the great number and diversity of variables as well as the numerous domains of study. In addition, other non-statistical parameters, including supplementary data collection requiring a stratification by regional office and auditing programs of business taxpayers which imposed a very high sampling rate (42% in 1982), also contributed to the large size.

Furthermore, it was difficult for the team assigned to the Personal Income Tax Statistical System to improve the stratification and optimize the sample size since a great amount of their time and resources was spent planning, managing and checking the different steps leading to the sample in its final form: adding descriptive variables, coding, transcription of information, cross-checking summaries at each selection etc.

The large sample size also increased the team's confidence in their ability to provide all the numerous statistical requirements and justified the lack of data quality control as well as measurements of precision of the estimates. However, data collection became more complex and less efficient, with the end result that the data processing activities (specifications of the selection criteria, the various coding, cross-checkings and definitions of statistical tables to be published) were draining most of the available resources, leaving but little for data analysis.

The problems encountered with a large sample size (not easy to use, difficulty of management, few resources left for data analysis) stimulated us to look for a general model which could define all the major components of a statistical system. The general model we selected is the Functional Analysis of an "Ideal" Statistical System of Fellegi; this analysis allowed us to identify strengths, weaknesses, lags etc. in the Personal Income Tax Statistical System.

3. QMR'S STATISTICAL SYSTEM

We describe briefly the Functional Analysis of an "Ideal"Statistical System of Fellegi [1] and subsequently illustrate the first developments of QMR'S statistical system.

The functional analysis (figure 2) identifies the functions and subfunctions of a statistical system which can provide coherent, relevant, timely, well-understood and readily accessible statistical information. The two main functions are (as defined in Fellegi's paper):

Function 1: obtaining, processing and disseminating data.

The first function is designed to provide statistical information and for this purpose, six separate subfunctions have been identified:
1.1 Analyze Requirements.
1.2 Identify Feasibility, Priority and Methodology.
1.3 Assemble Data.
1.4 Analyze, Interpret, Transform Data.
1.5 Disseminate Data.
1.6 Maintain Data Bases.

Function 2: maintaining and adjusting the framework within which the first function operates.

Function 2 is designed to maintain the framework for the short term operating and analytical activities of function 1. The three components of this function are:
2.1 Medium-term planning.
2.2 Development and promulgation of standard concepts and classifications.
2.3 Development and promulgation of standard tools and practices.

At present the QMR's statistical system consists essentially of the subfunction 1.3. The diagram of this subfunction (figure 3) illustrates the different activities which are part of it. The purpose of this data processing function is to produce clean and reliable data within a minimal time limit. The five principal activities are: sample selection, data processing, production of preliminary estimates, statistical estimation through a customized program tabulation procedures, adjustments to the sampling plan. The following describes each of these five activities.

3.1 Sample Selection

Prior to selecting the sample, feedback from the evaluation of the previous sampling design may necessitate modification of stratification. Sample size necessary to reach the desired degree of precision must be determined. The selection program must be modified to take into account possible changes in stratification and in sample size; finally, a verification of the selection, in order to detect unexpected increases or decreases in population sizes and discrepancies between actual and expected sample sizes, must be performed. We thus guarantee that any change in the structure of the taxpayer population will be monitored and we also verify that the parameters of the sampling selection have been accurately respected by the selection program.

3.2 Data Processing

The sample drawn from the Personal Tax Return Master File is an incomplete data base and is difficult to use for statistical purposes. The function of data processing is: to update, validate and clean this sample; to create descriptive variables (age, sex, geographic code, etc.) by recoding information already in the Master File; to add and impute a small sample of the special cases of the Taxation Act in order to obtain a sample representative of all taxpayers; finally, to add the variables needed to use the estimation procedure SESUDAAN [3] developed by the Research Triangle Institute. This SAS [2] procedure computes estimates and standard errors from sample data.

3.3 Preliminary Estimates

Preliminary statistics are obtained from a raw version of the taxpayer sample. This file is a smaller version of the original sample and is limited to the principal variables of interest (about one hundred). The variables needed by the Research Triangle Institute estimation procedure are also part of this file.

3.4 Customized Tabulation Program

We have tailored a tabulation program by using SESUDAAN [3] and SAS procedure TABULATE. SESUDAAN was developed by B.V. Shah at the Research Triangle Institute. It computes certain rates, means or totals, and their standard errors from the data collected in a complex multistage sample survey. The ratio estimates and their standard errors are computed for various domains (subgroups) of the population.

The statistical approach used for computing the standard error is a first-order Taylor approximation. This method for obtaining approximations of standard errors in large samples and in domains of study is well known.

Shah notes that his program provides one of the best known numerical approximations of standard errors for a large number of ratio estimates available in the literature (1981). Even though it is designed to handle many types of estimations, at the present time we only use this program for the usual estimates of totals (frequency and amount) by strata or by domains of study.

This procedure is not easy to use and, as B.V. Shah points out, "The use of this program is recommended only under the supervision of a statistician who fully understands all implications of the sample design used for data collection."

This warning lead us to write a program in the SAS MACRO-language which easily permits estimates of rates, means or totals and their standard errors (whether by stratum or by domain of study). The user needs only to specify which variables he wants to estimate, the desired estimation, the variable(s) which defined specific subgroups (stratum or domain of study). The program calls the SAS procedure "TABULATE" to format and to document results from SESUDAAN.

3.5 Sampling Plan Evaluation

This activity evaluates the sample and looks for improvements in the stratification and precision of the estimates. The sample selection activity (3.1) uses the findings of (3.5) the following year.

The subfunction "assemble data" is only in its first stage of development. We plan to modify and complete it each year according to inputs from data analysis and feedback from its own activities as well as from users.

4. SHORT-TERM DEVELOPMENTS

The short-term developments may be better understood by looking back at the functional chart of the "ideal" statistical system (figure 2). These developments will be: to complete the subfunction 1.3 "assemble and process data", to allocate more resources to subfunction 1.4 "analyze, interpret, transform existing data" and finally to "develop, maintain and promulgate standard concepts and classifications" (subfunction 2.2).

The subfunction "assemble data" will be completed by an editing procedure (validate and correct sample data). While it may seem strange to speak of editing data when all income tax returns filed are assessed, corrections made on returns by the assessing system might distort actual incomes, deductions or exemptions. Indeed, the assessing system may break down a large field value into other irrelevant fields or put a value in a field to guarantee that totals are exact. These distortions will introduce biases in the data; an editing procedure is therefore necessary to ensure proper statistical uses of the administrative records selected.

In 1985, much of the available resources were spent on the implementation of subfunction "assemble data"; we now plan to work on subfunction "analyze, interpret, transform existing data", that is to do more "real statistics". Fellegi [1] gives a list of the activities involved: "transform data through estimation modelling and prevision, evaluate the extent of coordination and integration through data analysis and confrontation of data with economic and social models". This analytical function will give statisticians a better knowledge and a better understanding of the taxpayer population.

Finally, we must develop standard concepts and classifications to provide well documented and well understood statistics. There will be a revision and an updating of the occupational and geographical codes and a data dictionary will provide exact descriptions of existing information. It can be very confusing when several variables are related to the same topic: for example, the Personal Tax Return Master File has nine variables on child care deduction: which one should be used to obtain estimates? A data dictionary will solve these problems and thus will enable us to produce the right estimates. Additionally, the number of variables in the data base can be controlled by eliminating useless or redundant variables.

CONCLUSION

The Personal Tax Return Master File of the Québec Ministry of Revenue (QMR) is a large administrative file. The sampling of taxpayers from this file has always yielded a large sample and the QMR's statistical system was. at the time of revision, a time consuming and complex process aimed solely at data collection.

The functional analysis of an "ideal" statistical system (Fellegi [1]) highlighted functions already existing in the QMR's statistical system as well as missing or weak components. Fellegi's functional analysis was, therefore, used as a guide in the implementation of a true statistical system at the QMR. This implementation dealt first with the subfunction "assemble data". Within the next year, we expect to complete this subfunction with an editing procedure and carry out activities of the subfunctions "analyze, interpret and transform existing data" and "develop and maintain standard concepts and classifications".

REFERENCES

[1] FELLEGI, I, (1978), Functional Analysis of an "Ideal" Statistical System, Statistical Services in Ten Years Time, (Duncan, J. edit.), Toronto, Pergamon Press.

[2] SAS (1982), SAS User's Guide: Basics, SAS Institute Inc., Cary, North Carolina.

[3] SHAH, B. V., (1981), SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data, Research Triangle Institute, North Carolina.

Figure 1 : History of the Personal Income Tax Sample 1972-1982

year	sample size	population size	sampling rate	number of strata	new stratification
1972	669 090	2 536 761	26,4 %	30	yes
1973	741 641	2 689 854	27,6 %	30	no
1974	489 696	2 778 031	17,6 %	30	yes
1975	552 011	2 909 571	19,0 %	30	no
1976	381 918	3 007 047	12,7 %	30	yes
1977	404 331	3 062 780	13,2 %	31	yes
1978	392 948	3 180 650	12,3 %	33	yes
1979	443 326	3 328 743	13,3 %	33	no
1980	490 304	3 418 317	14,3 %	122	yes
1981	540 995	3 713 591	14,6 %	122	no
1982	447 347	3 605 295	12,4 %	122	no

Figure 2 : Functional Chart of the "Ideal" Statistical System

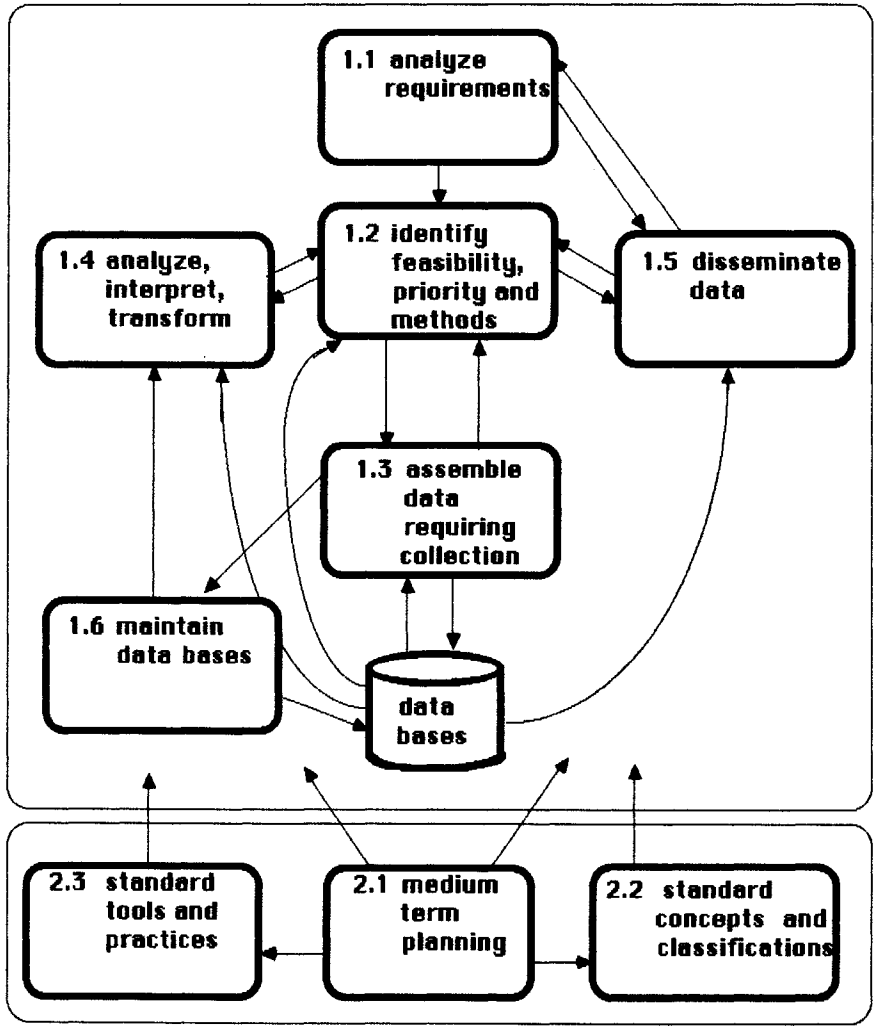

Figure 3: Chart of Subfunction "Assemble Data"