THE EVOLUTION OF THE SAMPLE DESIGN FOR PERSONAL INCOME TAX STATISTICS

Elaine Hoskins and Neil Barclay - Revenue Canada, Taxation

INTRODUCTION

Through the past decade there has been considerable attention given to the statistical uses that can be made of administrative records. Increasing emphasis has been placed on how these records can be manipulated, aggregated and linked in order to develop new statistical series or to eliminate or supplement surveys. However, the use of administrative records tracks back much farther than the last ten years, and actually began when government started issuing statistical reports on some of its operational programs. The pioneers of this movement are thus those who initiated the analysis of unemployment insurance programs, health and welfare schemes and, of course, taxation programs. This paper will describe, at the briefest of levels, the evolution of the Personal Income Tax Statistics program of Revenue Canada, Taxation, a program which has produced an annual set of statistics on the incomes and income taxes of Canadians since 1946.

The purpose of this exposé is to give other practising statisticians the benefit of our experiences in developing and using data from an administrative source. We begin with a history of the program, highlighting the events that have altered and improved it into its present form; then we will turn to the methodology of the current sample design, followed by a description of how our statistical system fits into the administrative operation and some of the problems encountered within that operation. We will conclude our paper by outlining possible plans for the future.

DEVELOPMENT OF THE PERSONAL INCOME TAX STATISTICAL SYSTEM

Taxation of personal income in Canada has existed since 1917. Although the Department's formal statistical program was not implemented until 1946, it did prepare and publish certain information such as revenue collections from the beginning years. In 1933, the first set of statistics which analyzed income distributions was published through the Dominion Bureau of Statistics, now known as Statistics Canada. Government recognized as early as 1933 the statistical value of tax records, as we see from the following quote: "Statistics of income that come to the knowledge of income tax officials have long been regarded as furnishing a guide to both the aggregate amount of national income and more particularly to its distribution by income classes and by occupations." (Incomes Assessed by Income Tax in Canada, 1933).

For the next ten years, the statistics published were fairly basic and were related to the government fiscal year in which the revenue was collected. Their publication was ad hoc. In 1943 the Department of National Revenue took the first steps towards the development of a statistical system to collect and publish detailed income tax statistics. The decision to make the change was occasioned by the growing interest in income tax statistics from an economic and social standpoint. The nature of the statistics that were published was changed from strictly administrative (reporting on the accomplishments of the Department) to both administrative and fiscal socio/economic. Defining the population as those filing for a specific tax year (equivalent to a calendar year), the first set of statistics (for the 1941 tax year) was released in 1946 under the title Taxation Statistics. This publication has been released annually since 1946, the latest of which (1982 edition) analyzes 1980 tax returns. The statistical system behind the publication is the basis for the discussion in this paper.

In the forty year history of the Personal Income Tax Statistical system, changes, modifications and enhancements have been effected and this evolution has been influenced largely by changes in user needs, changes in the tax structure and the population covered by that structure, and by changes in the operational environment. The personal tax return (T1) has always been the survey instrument, and sampling has always been used, under the management of departmental officials; those are the only factors that have not been altered over the years.

In 1946 a simple sampling scheme was used whereby selection was 20% of returns with taxable income of less than $5,000 plus all returns of over $5,000. (Taxable income is defined as income after deductions and exemptions.) Tables released were simple, showing only income and occupational distributions.

During the 1950's, various (still simple) sampling schemes were used to select about 6%-7% of the whole population. The tables that were prepared were broadened and indeed tables of "General Interest" were added, indicating that the publication was intended for wider use.

By the early 1960's the advent of electronic data processing was having a marked positive effect on the operation of the Department, and correspondingly on the personal income tax statistical system. Because of the new EDP technology, processing of tax returns was now centralized in Ottawa. A system was developed whereby returns were selected for the sample through a computerized selection routine, and a base file of assessment and identification data was created from the universe master assessing file. Through a data collection operation the remaining required detailed data was transcribed, keyed and then added to the base file. A series of programs to validate and edit the data was developed and implemented by the late 60's. Routines for the estimation of missing data (such as gross income) and classification (e.g. occupation) were also developed. The use of stratified sampling was increased as more variables such as tax status (i.e. taxable or non-taxable) were introduced into the design.

Throughout these first twenty-five or so years (i.e., 1946 to 1970) Revenue Canada,

Taxation developed a smooth running system for collecting personal income tax data. The users and clients were a fairly well defined group and the population under analysis was very stable, growing normally as compared to the Canadian population as a whole. Then in 1968, the Carter Commission studied income taxation in Canada and recommended vast changes in the system in order to implement income redistribution. The result, the Tax Reform of 1972, greatly changed the structure by taxing unemployment insurance benefits, integrating Armed Forces personnel into the system, and greatly increasing personal exemption limits as well as changing many other components. The population of taxfilers began to grow faster than had been previously the case. Provincial governments started introducing tax credit schemes such as property tax credits which brought in more filers, mostly to claim the credits. Moreover in 1978, the federal government introduced the refundable Child Tax Credit, adding an estimated one million women with little or no income to the taxrolls. The personal income tax system was now used as a vehicle to implement social programs as well as to raise fiscal revenues.

Because of the high usage of the system, fiscal analysts and now social policy analysts needed a precise way of analyzing the fiscal and income redistribution effect of policy changes. By the late 1970's more than 90% of all Canadians were covered by the system. Users were growing and requiring different analyses (for example, occupational analyses, data for market research, etc.), and Revenue Canada, Taxation itself became a user as it needed statistics on the processing system to evaluate its performance.

How did we cope with this explosion in requirements for personal income tax statistics? The answer: we expanded our existing system. To satisfy our fiscal analyst clients, additional data items were transcribed and a statistical subsample was built as the basis of a simulation model. To satisfy new clients data tables were increased, some published, some not, and a customized tabulation service was initiated. To satisfy our own Department's statistical requirements, another subsample was created containing tax data supplemented by administrative data related to taxpayer errors, time required for assessing, etc.

In addition to these changes, the sample design was reviewed, a more responsive design formulated and a procedure to monitor and update the design was put into place. The new design, using replication and stratification, was put into place in 1974. It has since been modified in terms of strata delineation, but the variables used to define the strata have not been altered. The following section will discuss the current design and how it has been maintained since initiation.

SAMPLE DESIGN AND SELECTION

Because we are dealing with an administrative operation where the universe is at hand, we have the opportunity of addressing what may be called a "clean" sampling situation. Some of the interesting features are:

(a) the frame is clearly defined;
(b) universe information both in terms of counts and amounts is available on a wide variety of variables;
(c) a computerized set of data spanning five previous years is readily accessible;
(d) published yearly statistics date back to 1946;
(e) an efficient collection network has been established; and
(f) similar samples collected throughout the world may be used to compare sampling techniques and quality of results.

The sample is selected from a computer file of all personal tax returns subsequent to their assessment. Selection criteria and sampling rates are programmed within the computer system and applied continuously as returns for the current year are received in the various centres across the country. The universe is the set of taxfilers who file between January 1 and December 31, resulting in the following implications: the sample does not reflect (i) prior years returns filed, (ii) current year returns filed after December 31 and (iii) returns that are reassessed. It is felt that inclusion of these returns would be far too costly in relation to gains.

The sample scheme is a stratified replicated systematic sample without replacement. Systematic selection is utilized primarily because of simplicity of implementation in a computerized environment. Periodic variation in universe values (for example, the tendency for certain taxfiler types to file at particular times of the year) does not occur sufficiently to impair the representativeness of the sample. Replication simplifies the estimation of strata variances.

The stratification criteria were initially developed with the objective of partitioning the filing universe into groups of individuals with similar tax income attributes and status, and in view of the primary data uses, the production of statistics by area, income level, etc. The universe of 15,000,000 tax filers is stratified according to the following criteria:

CRITERIA
1) Source of Income Employment Self-employment Investment
2) Urban Geographical Area 15 Urban groupings of cities with similar population sizes
3) Rural Geographical Area 13 Geographical areas: 10 provinces, 2 Territories and 1 Non-Resident. (Includes all areas not covered under urban)
4) Tax Status and Income Range Taxable: 4 income ranges Non-Taxable: 3 income ranges

The stratification defines 588 strata which are supplemented by two additional strata to cover high income individuals and invalid-coded returns.

Every year a sample size of 450,000 is set and is not allowed to grow larger because of resource constraints. Using regression techniques and five years of data, the income strata boundaries are adjusted annually so that these limits move with the frequency distributions, and the income strata represent the same taxfiler population from year to year. Because several subsamples are based on the personal income tax sample, stability both in terms of sample size and frequency distribution is a crucial element of the design. Next, using several years of universe data (for up to two years prior to current year), strata universe sizes are forecasted. The stratum population statistics and the means and population variances used to allocate the sample are then estimated for the current year. With the overall sample size fixed, the sample is then allocated to the strata on the basis of three variables, taxable income, total tax deducted and total exemptions, using Neyman allocation at the national level, and then adjusted through examination of marginal co-efficients of variation for the sub-population level. From these sample sizes the sampling rates are determined. As Neyman allocation does not provide adequate representation in relation to all data needs, there is consequently a requirement for certain imposed constraints. For example, in tax modelling simulation work several sub-populations are of special interest and are therefore required to be sampled at a higher rate of, say, 1:1 or 1:2, while on the other hand some strata have very large population sizes and small variances and may be required to be sampled at a rate of, say, 1:150 or 1:200.

The next step is computer programming of the selection criteria. Strata are assigned two random starts (one for each of the two replicates), then the program creates a stratum index for each return, compares the sequence of the return with programmed counters and either selects or does not select the return for inclusion in the sample. A weight, equal to the sampling rate is assigned to each selected return for use at estimation time. Returns are grouped by Assessing into some 70 batches to facilitate processing. We monitor batch by batch the strata population and sample counts, comparing them with historical ones to determine if the sampling is progressing as expected.

Previously, we have made reference to a "clean" sampling situation; however, this is not to say that we are not without problems. Any significant changes in legislation, economy and demography which may have an effect on the distribution of taxpayers must be taken into consideration. In addition, certain problems inevitably arise because we are operating within an administrative system.

As mentioned earlier, we currently use income indices to adjust the strata income limits. We started this technique in the 1977 tax year primarily because yearly rises in income caused significant shifts to higher income strata, which is a special interest group sampled at a higher rate. Also the introduction or change of tax credit legislation can alter (increase or decrease) the filer population in various strata. For example, the Child Tax Credit legislation in 1978 introduced 1,200,000 new filers to the little or no income strata. In 1980, Ontario senior citizens no longer had to file for provincial credits through the federal tax system, decreasing the filer population by about 200,000. Canada Savings Bonds maturing in 1978 and 1979 tax years caused a shift of parts of the population to investment income sources. Review of the subject matter usually enables us to foresee major population shifts such as these, from which we adjust strata population estimates before sample allocation.

The sample design may be affected by or must be adjusted because of operating within an administrative system. In the past we have experienced problems such as the implementation of incorrect geographical area codes requiring mid-stream correction. Our system tails the Assessing system and when there are disruptions, changes or reruns in the system the effect is felt all the way through. Reruns of the assessment of an entire batch of returns, for example, necessitate the adjustment of all our counters for sample and population sizes. We must continually watch the assessing system to identify such problems immediately.

We have discussed the basics of the sample design and how we handle some of the problems. The next section discusses briefly the data collection.

DATA COLLECTION

For each return selected, a transcript containing identification and some assessed information is printed, and the selected returns and their transcripts are pulled and routed to the data collection unit. At that unit, clerks transcribe the required data, assign appropriate occupation codes and verify data classifications. The transcripts are then keyed and the data passes through validation routines comprised of balancing, comparisons, and cross-reference checks. Any transcript in question is sent back to the unit for resolution. This process is repeated for all selected returns. The process is called "on-stream" as returns are "intercepted" immediately after assessment and the statistical record is created shortly thereafter. The system is closed in that all selected returns must be added to the master statistical file. No return may be dropped. The file is "built" throughout the year, lagging the actual processing system by only a matter of days and thus allowing us to monitor the creation of the statistical data through analysis of the file. With respect to missing data, you will recall that the system reports on assessed tax returns. Thus missing data only occurs for variables that were not assessed such as "gross income". To correct this problem, we assign values at year end. After this stage, the file is complete.

SPECIAL FEATURES RELATED TO OPERATING WITHIN AN ADMINISTRATIVE ENVIRONMENT

Throughout the paper we have discussed how the system is designed and how it fits into the personal tax return processing operation. Naturally this discussion has touched on various aspects that are dictated or influenced by the administrative operation. Some of these influences are negative but a good portion are positive. This portion of the paper summarizes these influences and how they affect the statistical system.

Positive points include the following. The universe is at hand, and thus provides ease of analysis for sample methodology and selection. Statistics selected are from assessed returns and thus have had certain validation prior to entering the statistical system. Because our universe is the assessed filing population, we do not have non-response in the traditional sense. The cost of the entire system is much cheaper than conducting a similar income survey by questionnaire. Turning to the negative side, you must realize that we have little influence over the design of the survey instrument with respect to information such as residence or occupation, as the statistical purpose is considered secondary. The universe is constantly changing, particularly in recent years, as new items become subject to taxation or new tax credits are introduced. It must be mentioned that non-filers exist outside our defined universe and should be considered by users when making interpretations of the data, as these individuals come from specific segments of the population. Any change to the process affects the data collection operation since our system has been designed to be on-stream. For example, decentralization of the Assessing operation occurred in 1975 and now seven separate centres process the returns. Data collection had to be decentralized along with the other operations, and the informal relationship that existed between the statisticians and those running the data collection had to be formalized in order to maintain consistency between centres.

The issues discussed above illustrate many of the advantages, problems and challenges our statisticians are faced with.

FUTURE SAMPLE DESIGN CHANGES

Now we look to the future and what that holds for the sample methodology and the collection system.

A challenge will be to redesign the Personal Tax Sample by utilizing other sampling and estimation methods and developing new techniques in order to obtain an optimal design in terms of efficiency and a manageable sampling procedure. An area of current concern is the overall sample size and its control. The stability of the current sample size depends mainly on the accuracy of strata population size forecasts. In the future, we may consider other approaches such as multi-stage sampling, different stratification criteria, etc., as well as the use of more universe data in estimation procedures in part through ratio or regression estimators. Intercorrelations between tax variables have already been extensively studied in other projects and this information could be put to good use with respect to the sample design. To go one step further, a completely dynamic sample selection procedure could some day be developed.

Developments in the data collection procedures are envisaged within the next two years. Using direct data entry through preformatted prompting computer screens, required data can be keyed, validated and corrected instantaneously. With this new on-line procedure, clerical resources will be reduced, and data will become more timely.

Clearly many other approaches and procedures could be addressed and whatever focus or direction is taken in the future, the Personal Tax Sample will continue to be of interest to sampling experts in general.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the assistance of a number of colleagues for their comments. Particular recognition must go to Peter Pashley and Phil Mette who provided helpful discussion and input.

REFERENCES

All research and information used as input into this paper was obtained from internal documents of Revenue Canada, Taxation and from the publication series Taxation Statistics, available through Canadian Government Publication Centre, Supply and Services Canada, Catalogue No. RV 44.