PREPARING DATA FOR TAX POLICY AND REVENUE ANALYSIS

George Ramsey
Franchise Tax Board, P.O. Box 2229, Sacramento, CA 95810-2229

KEY WORDS: Administrative record research, income tax, tax model

INTRODUCTION

During the early years of tax programs in California, limited statistical data were collected for the entire taxpayer population of less than half a million. Due to the relative simplicity of tax laws, statistics were primarily used to summarize data for administrators and for public dissemination.

Now, the California Franchise Tax Board (FTB) receives over 14 million tax returns each year from individuals and corporations. The increase in the taxpayer population has been accompanied by ever increasing complexity of tax law. Pressure to modify existing tax provisions originates from many sources, including the executive branch, the legislature, and, through the initiative process, directly from California taxpayers.

Because of the ongoing interest in taxes, the need for statistical products has expanded beyond simple summarization and presentation of information. Today's tax policy and revenue analysts have developed a diversity of modeling and forecasting techniques which rely on statistical data in formats amenable to computerized processing.

In contrast to federal tax programs, which entail the transcription of enormous amounts of tax return data to computer files, only enough California tax information is data entered from each return to allow for computerized validation of the taxpayer's computation of tax liability, balance due, and refund amounts. This approach saves much in labor costs since federal data can later be used to enhance state data for purposes of compliance and enforcement.

For purposes of tax policy and revenue analysis, however, detailed data about the characteristics of California tax returns are required much sooner than can be afforded by the cooperative state/federal data exchange programs. Statistical sampling is thus employed to identify income tax returns to be subjected to extensive data collection.

California sampling programs currently include cross sectional and longitudinal samples of personal and corporate tax returns. As a result of state conformity to the federal tax reform act of 1986 and in order to accomodate a variety of legislative and administrative mandates, the amount of data collected from returns has increased fourfold.

In addition to the traditional data summaries and revenue impact analyses, sample data are being used to prepare analyses of specific tax law provisions of interest to state policy makers. These include such areas as the treatment of capital gains income, net operating loss carry forward, passive losses, and income from S corporations.

The Personal Income Tax (PIT) Sample consists of roughly 100,000 tax returns from which as many as 350 data items may be acquired. This sample serves as the primary tool for tax policy and revenue analysis for the FTB. While other sampling programs, such as that for corporate taxpayers, are also very important in terms of developing useful analytic tools, the PIT Sample is illustrative of statistical programs administered by the FTB. Design and development of this sample, data collection procedures, data enhancement, expansion to the population, reliability, and analytic products are discussed in detail below.

DESIGN AND DEVELOPMENT

The PIT Sample is a stratified random sample. For the 1990 tax year, 120,000 tax returns were sampled from a population of 14 million. Stratification is based on the type of return filed (short form versus long form), amount of adjusted gross income, remittance status (whether payment accompanied the return or a refund was requested), and taxability (whether or not net income resulted in a tax liability). Sampling ratios vary from 0.2 percent to 100 percent.

Stratification and other design elements are intended to result in a sample which is representative of the taxpayer population in terms of the distribution of income and tax liability within taxpayer filing status. In addition, the sample produces representative data for most sources of income, adjustments, and deductions. Low frequency events which are reported on tax returns (such as certain credits or business items) are representative only to the extent that their occurrence is limited to strata with high sampling fractions. Gross geographic distributions (such as north versus south) are adequately represented, while fine detail (such as county or city) are not.

Due to the annual revision of tax laws, return forms, and processing systems, sampling specifications undergo annual revision. The annual change process involves the determination of pertinent data items to be acquired, enhancements to edits to be performed for sampled returns, inclusion of external data sources, and review of sampling criteria and sample size.

Sample selection procedures are integrated with the FTB's principal return processing systems. Detection of tax returns which satisfy sampling criteria is accomplished through systematic interrogation of data during computerized validation of taxpayers' calculations of tax liability. Sampling ratios of less than 100 percent are achieved through systematic selection of document locator numbers (DLNs) which are sequentially assigned prior to data entry. This DLN numbering system allows for tracking, filing, and re-locating return documents.

DATA COLLECTION

Tax returns which are found to satisfy a sampling criterion are posted to a temporary master file tape. This file grows through the tax return filing season until data collection staff are trained and available to begin transcribing return data. At that time the temporary file is reformatted and copied to the permanent sample master file. Various data items which were captured in return processing are carried forward to this file. Other data collection materials which are produced at this time include a printed list of selected returns, document request forms, and formatted data collection transcripts.

Data collection transcripts are three-page documents on which up to 350 data items may be entered. In addition to identification and demographic information, data from the California tax return and associated schedules as well as federal tax returns and schedules are transcribed to the data collection transcripts. Data items which were data entered for basic return processing are preprinted on the transcript. The organization of data on the transcript approximates the sequence of forms and schedules included in the tax return.

Transcription of information from tax return documents to the data collection transcripts involves a three-stage cycle of events: transcription, key data entry, and automated edits. During the initial cycle, all available data from the tax return is transferred to the transcript, data entered and edited. If the information transferred to the transcript is found by the automated edit to be free of errors, the information is posted to the permanent sample master file. If an error in transcription or data entry is detected by the automated edit, a second cycle transcript is printed which includes edit flag indicators adjacent to data items which produced the error condition. This transcript is used to modify erroneous data and is recycled until the automated edit detects no errors and posts the record to the sample master file as complete.

Specific types of edits incorporated in the automated edit program include verification of detail data summing up to total within certain schedules, cross reference checks to ensure that totals from substantiating schedules are properly carried to summary schedules and the primary tax return, and limitation of particular entries to statutory maximums and minimums.

Many of the errors detected by computerized edits are errors committed by the preparer of the tax return. Of the 120,000 returns processed for the 1990 tax year, over 14,000 contained errors attributable to the preparer of the return. The errors committed in preparing returns are not corrected. The sample is designed to represent the income, deductions, and tax liabilities originally reported by taxpayers. Determination of the proper entries that should have been made is discussed below under expansion and enhancement.

In order to post records with detected preparer errors to the permanent research master file, it is necessary to bypass computer edits. Depending on the source of the error, one of various "edit bypass" codes is entered on the data collection transcript. Each edit bypass code identifies specific items which, because of incorrect preparer entries, resulted in the error condition that was bypassed. When an edit bypass code is invoked, the computerized edits are deactivated, and the record is posted as complete to the permanent master file.

Another situation which causes difficulty for computer edit programs is when insufficient detail schedules are attached to validate entries on the return. In particular, federal schedules which are pertinent to amounts recorded for California are often absent from documents available to the data collection staff. For example, the federal return, with its detail of income sources and deductions, may not have been attached to the state return. Such detail is required for accurate expansion and modeling. However, some California taxpayers have no federal filing requirement and, thus are not required to file and attach a federal return. These conditions require that various criteria for filing requirements (among others) be evaluated to allow exceptions to edit programs. However, when certain critical data is unavailable, a unique completion status code is applied by the computer edits. In such cases, returns with both federal and state filing requirements are posted as incomplete.

Unpredictable circumstances invariably result in sub-populations of returns which are not subjected to sampling procedures. Exclusions of such returns could seriously bias sample estimates. Examples include ambiguous return preparation instructions which cause taxpayers to make irrational entries, new tax law provisions which result in unpredictable taxpayer behavior, and certain processing functions which do not adequately account for idiosyncracies of individual tax returns. Such conditions often require that returns be processed manually, thus circumventing computerized sample selection protocols. To prevent potential bias, such returns (usually fewer than 30,000) are examined outside the automated environment to determine whether they should be sampled.

ENHANCEMENT AND EXPANSION

Of the 120,000 returns sampled for the 1990 tax year, about 107,000 were posted to the master file as completed, cross-sectional records. Just under 2,000 return documents could not be located for statistical transcription of information. Roughly 5,000 did not contain enough detail to satisfy modeling needs. About 5,000 were completed for the longitudinal panel. Less than 1,000 were deleted for other reasons. Of the completed records, about 1,000 were returns for previous tax years, which were set aside for separate analysis.

About 13,000 sample records were compiled for taxpayers who filed the California short tax form. Such taxpayers were not required to attach the federal return. Therefore, detailed data about income and deductions were not included. In order to obtain income and deduction details about these taxpayers, the state sample was matched against the tax return data files supplied by the Internal Revenue Service.

The Individual Master File (IMF) and the Individual Return Transaction File (IRTF) are acquired by the state primarily for compliance purposes. However, these files contain a wealth of income, adjustment, and deduction data which are quite useful for tax modeling at the state level. In cases where insufficient information about income and deductions is available from the state return and attached federal return, the federal files are used to augment statistical data collected by the FTB. More than 13,000 sample returns were augmented by federal data acquired via the state/federal exchange programs for the 1990 tax year. Data captured from the federal data files is not allowed to replace data transcribed during the data collection process. Only missing data fields are subject to augmentation. Other administrative record data are added to the sample file later.

Accurate expansion of the sample file to the population is critical to tax policy and revenue analysis. Even though specific sampling ratios are designed into the sample, various factors, as stated above, result in less than a perfect yield. In addition to returns with incomplete data, and manually processed returns, some returns are not located for transcription within time constraints for producing the completed sample.

An additional factor, which is assumed inherent in sample selection, is that the temporal distribution of tax return filings is sampled representatively. However, due to technical constraints involving the assignment of document locator numbers, this assumption is not always true. Therefore, ex post facto stratification of both the sample and the population is performed to develop appropriate sample expansion factors. As a result sample unit weights varied by as much as 12 percent from planned values for the 1990 tax year sample.

For purposes of tax law simulation modeling, completed records to which edit bypass codes were applied and those containing omissions of certain data are identified. For this group of records, "residual" amounts are computed such that the sum of amounts posted to the completed records and the residuals results in a balanced tax return. A residual is computed, for example, when the sum of income source items does not equal the total income entered by the return preparer. In this case the sum of income items and the income residual equal the total income amount entered by the preparer. Similar computations occur when certain schedules are internally consistent but are not consistent with entries made on other forms.

RELIABILITY MEASURES

Two forms of sample data use are relevant to determination of reliability of data produced by the sample. One use involves estimation of overall program revenues for state budget preparation. The other involves estimation of the impact of tax policy revision and who may be affected.

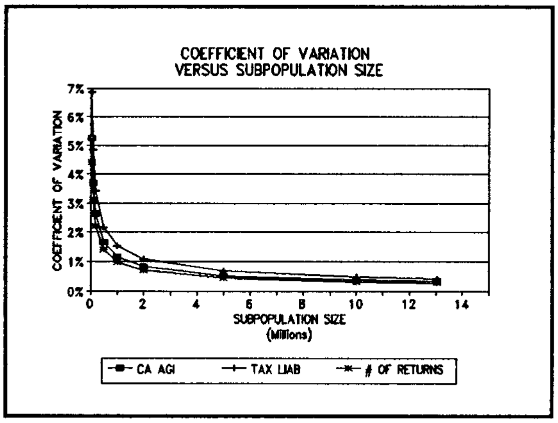
Overall program revenues depend on factors such as population, employment, personal income, and other economic factors such as job creation, business profits, and capital development. The PIT Sample provides information relevant to analysis of these factors in California. Reliability of such information is generally computed for the aggregate population level.

Tax policy analysis, on the other hand, often deals with smaller sub-populations. Disaggregation of information to the impact group arena is of greater importance, especially as it applies to low income taxpayers, high income taxpayers, families, business, seniors, benefit recipients, etc. Therefore, reliability of estimates for disaggregated sub-populations is often of greater importance.

To accommodate the needs of both revenue and tax policy analysts, coefficients of variation for various estimated population and subpopulation sizes are computed for aggregate adjusted gross income, tax liability, and number of returns filed as shown in Figure 1. Calculation of these coefficients assumes a proportionate distribution of sample returns across sample strata. As shown, as sub-population size increases, the coefficient of variation of estimates decreases.

Because sampling fractions approach 1.0 for high income (or loss) returns, estimates of income and tax liability for wealthy taxpayers are much more accurate than for middle and low income taxpayers. Figure 2 illustrates the relationship between average adjusted gross income and the coefficient of variation within income category.

Figure 1

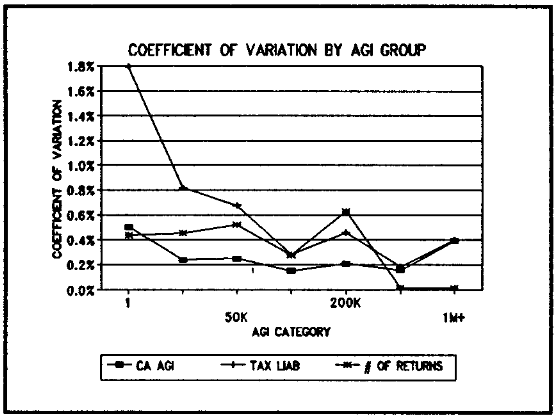


ANALYTIC PRODUCTS

The products of the PIT sample are comprised of printed summary statistics, computer files used for preparing ad hoc analyses, a computer file for use as input to the FTB's tax law simulation model, and special study subsamples.

Printed summary data consist of a distribution of each income, adjustment, deduction, and tax computation item across adjusted gross income and taxable income categories. Other data, including California adjustments to federal adjusted gross income, personal and dependent exemption credits, selected special tax credits and administrative information, such as preparer type and preparation media, are also included on the printed summaries.

Figure 2



A separate set of distribution tables is produced for each filing status and for taxpayers that claimed a renter's credit. Distributions for taxpayers with a tax liability are produced separately from those with no tax liability.

Usage of the printed summary data can be divided into three categories. The first use category involves validation of assumptions made to estimate total program revenues estimated for state budget preparation. Based on sample data, previous estimates for the tax year of the sample are revised and treated as actual. Previous estimates for the subsequent tax year are revised to incorporate the new data in the forecasting model. And estimates for the second subsequent tax year are prepared.

The second type of use is in preparing statistical material for publication. The principal public document which relies on printed summary data from the PIT sample is FTB's annual report. This report describes significant events of the tax year, discusses the impacts of new tax programs and tax legislation, and contains an extensive statistical appendix of income and filing status distributions.

Responding to inquiries is the third use category. Printed summary data can often be used to respond to inquiries about specific groups of taxpayers and about potential effects of proposed legislation.

While printed summary data is maintained in its ordinary form on paper, it is also maintained in print image form on computer media. The increased use of personal computers in the work place has resulted in greater demand for down loading of main frame computer files for use in analytic techniques developed for personal computer spread sheet and data base software. Applications of this data transfer technology have been implemented in all three of the major categories of use for printed summary data.

Another analytic product of the PIT sample is the sample data computer file. This file is maintained as an on-line SAS data set. It contains sample records with expansion weights and values for each of the 350 potential data items. The data set is partitioned into subsets of completed records, incomplete records, deleted records, records for which insufficient income and deduction detail were available, and longitudinal panel records.

The completed records subfile is particularly useful in responding to ad hoc inquiries, which are quite frequent. Because the file is stored on a direct access disk device, interactive statistical analysis is possible using SAS. Interactive processing allows analysts to address issues concerning optional target subpopulations and progressive changes to or limitations of selected tax provisions. The processing speed and efficiency of this type of analysis afford the capability of obtaining answers to specific inquiries and solutions to statistical problems in a matter of minutes.

Interactive processing is also used to test and debug specialized taxpayer behavioral models and tax law simulation models. Naturally, on line statistical analysis is limited by system defaults for core memory and working storage space. Consequently, a substantial amount of processing is performed in batch mode. Such processing is characterized by the need to address the entire sample file, extensive amounts of calculation, and the necessity of sequential analytic steps.

The analytic product of greatest importance for tax policy and revenue analysis is the PIT law simulation model. Each year, several hundred legislative proposals are evaluated. To address the distribution and extent of revenue impacts which could result from changes to the tax code, the model recalculates tax liability for each sample record. Proposed changes are identified to the model by use of input parameters which may modify or eliminate amounts of tax related data items or limit certain aspects of the population being targeted for tax law impact. The model compares base law to proposed law output values to identify how changes in overall program revenue are distributed across the population.

Before sample data is used by the model, growth factors are applied to each data item to allow for population growth, increase in income, and other economic projections. This extrapolation process allows the model to compute revenue impacts not only for the tax year represented by the sample, but for up to five years into the future.

While the PIT sample is invaluable in producing accurate analyses of tax policy issues and revenue impacts, the amount of data collected is not always adequate for the intensive study of specific provisions of the tax code. In order to accomplish studies of greater detail, the PIT sample is used to identify specific subpopulations to be subjected to in depth analysis. Examples of such spin off studies include the data collection and analysis of taxpayers' capital asset transactions, depreciation methods, business net operating loss deductions, alternative minimum tax, passive income and loss limitations, S corporation pass through income and loss, wage and salary withholding, and, through the longitudinal panel, taxpayer life cycles.

CONCLUSION

Sampling programs for state tax returns have grown considerably due to the availability of computer processing. Today, from the development and data collection phases through the analysis and modeling phases, it is difficult to conceive of working without computers. Development of computerized sampling, data collection, and analytic systems has expanded the arena of tax policy and revenue analysis. As we move forward to more efficient data collection and analytic systems, ever more complex and imaginative proposals will be subjected to sample based tax policy and revenue analysis. Even now, the development of an integrated tax burden model for the state is being proposed to unite state personal, corporate, sales, gasoline, and special tax program data with local property and business tax data. Such an undertaking is of grand scale even with respect to the current state of the art. Through the application of sampling and ever improving data technology, advanced projects will be enabled, to the benefit of all.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the contributions of Patricia Minor, Manager, Sample Development Unit and Larry Gilbert, Manager, Data Collection and Statistical Support Unit.

This report represent the opinions of the author and do not constitute official policy of the California Franchise Tax Board.