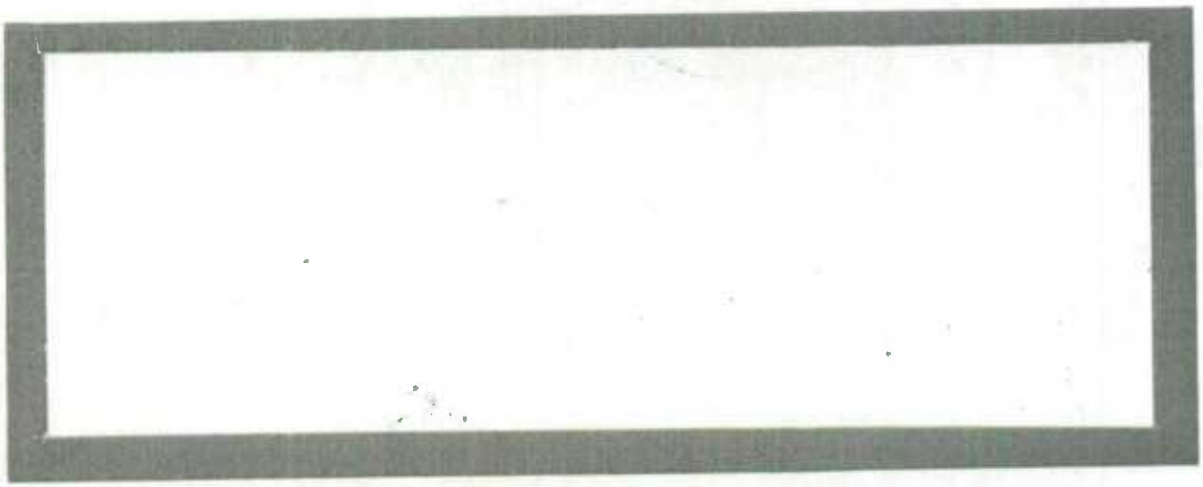
Statistics Canada | Statistique Canada

DRAFT



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes auprès des entreprises

Canada

ANNUAL SURVEYS OF ECONOMIC PRODUCTION: METHODOLOGY FRAMEWORK

M. Colledge

Working Paper No. BSMD 85-046E

ANNUAL SURVEYS OF ECONOMIC PRODUCTION:

METHODOLOGY FRAMEWORK

Michael Colledge, B.S.M.D.

Draft April 9, 1985

Contents

1. Introduction

The objectives of this document are to elaborate upon the methodological framework for annual surveys of economic production as laid down in the Infrastructure Project Strategy [1] , and to outline a practical means of implementation.

The document is primarily concerned with the provision of frame data and use of income tax data. Adoption of a common strategy for frame and tax data, however, will provide the basis for standardizing all aspects of survey methodology.

Section 2 of the document contains a background to annual surveys of economic production. It comprises an overview of the objectives, the basic procedures, the frame data requirements and the practical constraints under which these surveys are designed and operated, and a brief summary of the approach recommended in the Strategy.

Sections 3 and 4 describe in much more detail the methodological framework implicit in the Strategy and its practical implementation in terms of the Central Service Function.

In the final section of the document a number of issues and unresolved problems are identified and suggestions are made for their examination and resolution.

2. Background to Annual Surveys of Economic Production

2.1 Basic Objectives and Structure

The objectives of an annual survey of economic production are to collect and publish a full range of financial, production and commodity data, including the "principal statistics" [4], broken down by 3-4 digit SIC (1980), by province/major urban area and, possibly, by size. The appropriate statistical unit is the establishment, defined in the 1980 SIC [4] but modified in the Strategy [1, page 16, item A7]. Data collection is traditionally by annual mail questionnaire.

In principle there could be a single annual survey covering economic production in all the industries for which data are required. In practice the target population is divided by industry into (supposedly) mutually exclusive subpopulations each of which is covered by an independent survey. One of the goals of the Strategy is to ensure that these separate operations are standardized and coordinated so that, in effect, they may be collectively viewed on one survey [1, p. 31, item I].

Annual surveys of economic production provide control counts and totals for use by the corresponding subannual surveys. Conversely, subannual data may be used to help maintain annual survey frames.

2.2 Operations

For the purpose of this paper, the major activities of an annual economic production survey may be described in terms of operations and responsibilities as follows:

(a) production of the frame for sample selection, (CSF responsibility);

(b) sample selection, i.e. creation of an "occasion" file with initial sampling weights inversely proportional to selection probabilities, (joint CSF/survey operations' responsibility, for which the CSF will provide facilities);

(c) data collection, i.e. mailout, capture, follow up, etc., (survey operations' responsibility, using CSF facilities where available for mailout and follow up);

(d) data processing, edit and imputation, (survey operation's responsibility);

(e) estimation and tabulation, i.e. calculation of final sampling weights obtained by adjusting initial weights for non-response, outliers, edit failures and births not included in the frame originally created for sample selection (survey operations' responsibility using updated frame counts provided by the CSF).

The timing of these operations for reference year Y data is along the general lines:

sampling	-	December, Y
mailout	-	February, Y+1
data capture	-	March-October, Y+1
estimation	-	February, Y+2

In practice there may be more than one mailout, and two or more sets of estimates, e.g. preliminary, revised and final.

At any give point in time survey operations may be involved with three reference years. For example, in March Y+1 ongoing activities include:

preparation of estimates	-	year Y-1
mailout	-	year Y
building frame	-	year Y+1

2.3 Frame Data Requirements

The frame data requirements associated with annual surveys of economic production in reference year Y may be summarized as follows.

(a) **Sampling** - A list of establishments is required for sampling. It should contain all statistical units for which there was economic production during the given annual reference period (year Y), but no duplicates nor units which are irrelevant, e.g. inactive or out of scope. Associated with each unit should be all the variables likely to be used for stratification, in particular, industry, geography and size.

In practice, at the time of sampling late in year Y, the list may not be complete.

(b) **Contact** - Just prior to mailout in, say, February Y+1, contact information, e.g. name, address, "for attention of", "report for", is required for each sampled unit.

(c) **Estimation** - As the frame may not be complete at the time of sampling, any additional establishments discovered during the course of year Y+1 to have been active in reference year Y should, in principle, be added to the frame (and sampled if possible). In practice, this does not require production of a new frame for year Y, simply an updating of the frame stratum counts and corresponding sampling weights for estimation when this takes place.

Data collected by survey questionnaire and follow-up are put onto the "survey response file" for year Y. Some of these data may indicate changes in the value of frame items. However, with the exception of birth information they are not used to update the frame for year Y but rather to build the frame for year Y+1.

For edit and imputation purposes it is important to be able to compare response data for the reference year with that for previous years, at least for all large establishments. This implies a frame requirement to "track" establishments through changes of ownership etc., over the course of time.

2.4 Implementation Considerations

There are three important practical factors to be taken into account in designing and implementing annual surveys of economic production.

Frame maintenance costs. There are a very large number of small establishments, in total from 800,000 upwards (depending upon the particular criteria for used to define what is in scope). It would require a huge commitment of resources to:

(a) detect and add all births;
(b) detect and delete all deaths;
(c) maintain precise, accurate, up-to-date classification data;
(d) track all establishments through time.

Consequently it is unrealistic to contemplate maintaining a complete frame with all the derived characteristics.

Response Burden and Costs. Small businesses are reluctant to respond to survey questionnaires. Thus STC are obliged to consider use of income tax data wherever possible, to supplement or replace survey data, both for economy and for response burden reasons.

Heterogeneity of Establishments. Not only is there a very large number of small establishments but there is also a small group of very large ones. The universe is not at all homogeneous. This is graphically

illustrated by conceptual and practical examples in fig. (1), (2) respectively: 8% of companies in scope for SEPH account for 79% of employment, and although companies do not coincide with establishments, this serves to illustrate the point.

2.5 Basic Elements of the Strategy

The basic elements in the design and implementation of annual surveys of economic production as given in the Infrastructure Project Strategy are as follows.

(a) No attempt will be made to maintain a complete, unduplicated, fully classified, up-to-date list of active establishments.

(b) Maintenance resources will be allocated according to the size and significance of establishments, the larger ones receiving a higher proportion of resources than smaller ones; the target of precise, accurate classification will be confined to establishments which are large or are sampled in the current year.

(c) Income tax data will be used for two distinct but related purposes: to help in frame maintenance, and to supplement survey response data.

(d) Data for smaller units which cannot be obtained from income tax returns will be collected on a reduced or "short form" survey questionnaire. Very small units will be defined out of scope for direct survey.

(e) The operations presently engaged in the processing of frame and income tax data will be organized and suplemented to form a Central Service Function (CSF). The CSF will have two components: (1) concerned with the provision of frames, using tax data to complement other frame data maintenance procedures; and (2) concerned with the provision of tax data to supplement and replace survey response data.

3. Methodological Framework Frame and Income Tax Data

3.1 Introductory Remarks

The Infrastructure Strategy specifies the basic framework for economic sureys and describes in some detail the functions of the CSF and corresponding Central Frame Data Base (CFDB). The Strategy contains explicit reference to annual surveys of economic production in items A6, A8, C17-18, H1 and I1-I6 of Chapter 4. This intention of this section is to elaborate upon and extend these ideas.

3.2 Division of the Target Population for Data Collection

For data collection purposes the target population will be divided essentially on the basis of size and significance into three categories: (A) large; (B) small; and (C) out of scope.

(A) Large (Long Form) This category will contain the following:

(A1): very large establishments, individually significant in terms of their impact upon published estimates;

(A2): establishments which although not in A1 are individually significant in terms of their impact upon published values in certain strata;

(A3): establishments not in A1 or A2 which are significant for operational reasons, for example, small establishments of a large enterprise, or small establishments not covered by income tax returns;

(A4): establishments which are individually insignificant but which, measured in terms of a prespecified size threshold, are sufficiently large to warrant coverage by a full scale "long form" survey questionnaire; the size threshold may be industry dependent but has yet to be precisely specified (see Section 5).

The establishments in categories A1-A3 will be surveyed annually with certainty. Those in category A4 will be sampled on a probability basis, with sampling rates increasing with size. It is conceivable that category A4 will be null. All sampled establishments will be mailed a "long form" questionnaire requesting the full range of financial, production and commodity statistics.

(B) Small (Tax Based) This category will contain all the individually insignificant establishments which are not large enough to warrant a long form (i.e. not in category A) but which, measured in terms of a prespecified lower size threshold, are large enough to be in scope for survey (i.e. not in category C). The threshold will be expressed in terms of income tax data items captured by Revenue Canada (RCT) for all tax returns thus will not be industry specific. All establishments thus defined will be identified by, and assumed coincident with (i) corporations filing T2 tax returns, or (ii) unincorporated businesses belonging to individual T1 tax filers (or to partnerships of T1 tax filers) reporting self-employed income from economic production. This set of establishments will be referred to as "tax-based". The precise specification of the lower threshold and of what constitutes "economic production" has not yet been determined (see Section 5).

The set of tax-based establishments will precisely complement the set of large establishments covered by long form. A sample will be drawn annually, the corresponding tax returns located and financial data obtained from them. A second sample will be mailed a "short form" questionnaire requesting limited financial, production and commodity statistics. Sampling in both cases will be on a probability basis with sampling rates increasing with size. For simplicity of estimation the short form sample should be a subset of the tax sample but this

approach has yet to be finalized and evaluated (Section 5). It is conceivable that the same questionnaire will be used for short and long form samples.

(C) Out of Scope Any economic production which is reported on tax returns falling below the specified lower threshold will be defined as out of scope for survey. The corresponding tax returns will not be sampled for mailout purposes. However, data on tax returns will be used as the basis for estimating the effect of omitting such activity from explicit survey coverage.

3.3 Data Collection by Direct Survey: Long Form

For data collection by long forms in reference year Y, say, for each survey a frame is required of all establishments which satisfy the criteria for category A, i.e. all establishments which are individually or operationally significant, or are above the specified size threshold in year Y. Associated with each establishment in the frame must be all the classification information for sampling and reporting arrangement information for mailout and data collection. The results of the sampling process will be the long form sample file with weights determined by the selection probabilities.

Data collected from long forms and follow up will be used to convert the long form sample file into the (re)weighted long form response file, where reweighting adjusts for non-response and for outliers. In general, deaths and changes in classification values, detected during survey data collection and follow up will not cause changes in weights. Such information, together with changes in contact data, will be incorporated in the frame for the following reference year (Y+1) and will not require a new frame to be generated for year Y.

Profiling, nature of business enquiries, PD and tax data will be used together with survey feedback from the reference year Y-1 to ensure the long form frame is as complete as possible at the time of sample

selection. Nevertheless, it is inevitable that some establishments satisfying the criteria for coverage by long forms will not be on the frame. Typically the exclusions will be new births and establishments which have recently grown above the size threshold. Some effort may be made to send long forms to these establishments which are detected in time to do so, and to adjust the frame and sample weights accordingly, but this procedure will probably not cover all births. However, by ensuring that the corresponding tax returns are sampled and data captured, complete coverage of the target population will be obtained, although, of course, not all the required data items will be available.

For edit, imputation and validation purposes it is important to access response data obtained in previous years from the same establishments. This implies the need to trace all establishments on the long form frame over time, through changes of ownership, etc. so that year-to-year comparisons can be made and continuity assured. In this context the precise definition of "continuity" as it applies to establishments has yet to be finalized, see Section 5.

3.4 Data Collection by Tax Acquisition

Tax data obtained from samples of T1 and T2 returns will be used to complement the coverage of the long form survey frame.

Tax returns will be sampled as they pass through the RCT Tax Accessing system. The sampled returns will be assigned an SIC code and tax data extracted. The data will be stored on tax response files. The precise nature of the SIC assignment and data capture process, in particular whether capture will be directly from the returns or from photocopy or microfilm facsimiles, has yet to be determined (Section 5).

Subject to final validation and some possible variations such as the incorporation of RCT sampled tax data (Section 5), the procedures for sampling and estimation will be as indicated in the following paragraphs.

In addition to the natural division according to source (T1 and T2) the tax sample will comprise two basic components, the "cross sectional" sample and the "prespecified" sample.

Cross-sectional Sample

An all-purpose, "cross-sectional" sample will be drawn. Stratification will be on the basis of current year values of variables captured annually by RCT for all tax returns and hence available at the time of sampling. These variables include gross business income, province and, for T2 returns only, assets. The sample may also be stratified using SIC codes carried forward from previous years or assigned by an automatic procedure based on business name and description (if any) captured by RCT. (The precise strategy has yet to be determined, see Section 5).

It is worth noting explicitly that, in order to select and draw the cross-sectional samples it will not be necessary for STC to maintain complete lists of tax-based establishments. During the course of tax assessing RCT build complete lists of T1 and T2 taxfilers for the reference year from which sampling can take place.

The "cross sectional" sample will be the major component, meeting some or all of the needs of every individual survey.

Prespecified Samples

For certain surveys the cross-sectional sample may not be sufficient to provide good estimates at the required level of industrial detail. Hence the sample may be supplemented by survey specific samples in particular industries. These samples are "prespecified" in the sense that, unlike the cross-sectional sample, the units are chosen before the tax returns are filed and processed.

In order to select prespecified samples a frame of tax-based

establishments with full classification and RCT identification information will be required. However the frame will not need to be anywhere near complete as overall coverage will be provided by the cross-sectional sample. A suitable prespecified frame will be furnished simply by using tax response files for previous years. Ideally the frame will be restricted to units sampled in the most recent available year so that classification data are as up-to-date as possible and the number of taxfilers who have since ceased or changed business will be minimized (see Section 5).

Subsampling

For certain surveys, in particular the Census of Construction, additional data items will be captured from the subsample of the combined tax samples. The frame for this subsampling will simply be the sample response file itself.

Weighting and Unduplication

The probabilities of selection of the cross-sectional and prespecified samples will be combined to produce an initial weight for each sampled return.

When RCT tax data processing for the reference year has been (essentially) completed the corresponding universe files of T2 filers and T1 filers reporting self-employed income will be obtained from RCT. These files will be used to provide control totals for reweighting the tax response data, to unduplicate the tax and long form survey frames, and to assist in building the long form survey frame for future years. The processing steps will be as follows.

First, the T1 (SIN) and T2 identifiers on the tax universe files will be matched against the list of T1 and T2 identifiers associated with establishments on the long form survey frames for the reference year. Common identifiers will indicate tax returns covered by long form

in a survey, i.e. duplicates. These duplicates will be flagged and removed from the tax universe counts and control totals. A complete unduplicated universe will thus in effet be created, comprising a long form component for each survey and residual all-industry tax component.

The tax response file for the reference year will then be matched against the list of duplicates and for those members of the list which were sampled the corresponding sampling weight will be adjusted to eliminate the effect of duplication.

Finally, the tax response file will be reweighted using counts from the tax universe files, to adjust for the effects of tax "non response" (i.e. failure to obtain tax data), and to improve the efficiency of the estimates. Unlike the tax universe file the weighted tax response file will be fully industrially coded, thus, in combination with the weighted long form response file will it provide coverage of the universe without duplication or omission for each individual annual survey. This will complete processing of tax data for the current reference year.

An additional use of tax data for reference year Y will be in helping to build long form survey frames for future reference years, Y+1, Y+2 which are as complete as possible. To this end, the identifiers of all tax returns on the universe and/or sample files which indicate economic production above the long form threshold will be matched against the list of T1 and T2 identifiers currently associated with the long form frames for year Y+1 and/or Y+2 (depending upon the timing of the matching). Units not on the frames will be investigated and, if appropriate, new establishments corresponding to such units will be added to the long form frames.

3.5 Data Collection by Direct Survey: Short Form

Direct data collection by a reduced, "short form", or "other characteristics", survey specific questionnaires will be required for

some surveys to supplement the financial data which can be acquired from tax returns, also (possibly) to calibrate tax and survey data against one another.

For simplicity of estimation the short form survey sample should be a subset of the tax sample. Thus the frame used for sampling should coincide with that provided for selection of the prespecified tax sample. In the case of T2s, as many reference year Y returns are actually processed during year Y, the frame for the short form sample could be supplemented by units already tax sampled (see Section 5).

4. Methodological Framework: Implications for the Central Service Function

4.1 Introductory Remarks

The methodological framework for annual surveys of economic production defined in Section 3 defines a number of requirements for provision of frame and income tax data and for survey sampling, mailout and data capture, which can be most efficiently and effectively handled centrally. This section is concerned solely with the first two items, i.e. provision of frame and tax data. The extent to which survey sampling, mailout, data capture, follow up and editing operations will be organized centrally and the relationship between the CSF and Head Office Operations Division have not yet been decided (see Section 5).

The division of the CSF into two components is based on the conceptual distinction between the use of income tax data to provide a frame, and their application in supplementng or replacing survey response data. However this distinction is not really very precise. For example it is the weighted sample response which actually provides the universe counts and totals by industry, to complement the long form coverage for each individual survey; conversely data items available on tax universe files can be used to supplement survey response data.

The two components of the CSF are as follows.

CSF(1). The CSF(1) will acquire tax universe files and use these data together with frame information from other sources, including the tax response files, to provide frames for selection of long form, short form and tax prespecified samples. It will provide all the facilities required for drawing the survey samples, and it may actually perform the sampling, to survey specifications.

CSF(2). The CSF(2) will select, acquire, code, data capture and weight income tax data and will disseminate corresponding weighted tax response files to surveys.

There will be passage of data between these two functions. In particular, production by CSF(2) of final weights will require tax universe counts and control totals (unduplicated against the long form frames). Conversely tax response data will be used by CSF(1) in the construction of prespecified tax sampling frames.

It is important to note that the CSF will have functions additional to those cited here to meet the needs of annual surveys of economic production. It will be responsible for the provision for frame data for all other annual and subannual surveys and will be the focal point for exchanges with RCT.

4.2 Central Service Function (1)

For each reference year the CSF(1) will perform three basic tasks.

(a) It will provide a frame of establishments to each annual survey for sampling and contact by long form questionnaire. The long form frame will be as complete as possible at the time of sampling and will contain all relevant, classification and reporting arrangement data. There will be facilities for subsequent adjustment of frame units to include any new establishments which were not detected until after sample selection but were in time to be added to the sample.

(b) It will provide to each annual survey a frame of tax-based establishments for prespecified tax sampling. The prespecified frame will not be complete but it will contain all relevant classification data and RCT identifiers. This same frame, augmented possibly by other establishments appearing in the tax sample for the current reference year, will also be used for short form sampling.

(c) It will provide to the CSF(2) the identifiers of tax returns duplicating the coverage of establishments in the long form frames, and the corresponding tax universe counts, net of such duplication. These data will enable the CSF(2) to unduplicate and reweight the tax response file.

In order to identify duplication of long form and tax coverage the CSF(1) will associate with each establishment or group of establishments all related T2 and T1 (SIN) identifiers. The relationship links between establishments and tax returns may be simple (1-1) or complex (m-n), as illustrated in figure 3, provided that they account for all potential overlap. The links will also facilitate the use of tax information for frame maintenance, in particular for updating classification data and for tracking of establishment through changes of ownership.

At any given point in time the CSF(1) may be involved in the processing of frame data for three different reference years. For example in March, Y+1 activities will include:

(a) unduplication of the tax universe file for year Y-2;

(b) provision of long form mailout information for year Y-1;

(c) provision of a frame for prespecified T2 tax sampling for year Y.

In addition the CSF(1) may be required to prepare interim tax universe counts and duplicate lists for preliminary estimates.

As stated in the Strategy, the primary tool used by the CSF(1) in fulfilling its function will be the CFDB. The CFDB will have two components, the Integrated Portion and the Non-Integrated Portion.

CFDB Integrated Portion

The establishments which constitute the long form frames together with their associated tax linkages will be created and maintained in the Integrated Portion (IP) of the CFDB. The IP may also contain additional establishments not in scope for long form sampling, e.g. establishments below the long form size threshold, establishments relevant to subannual surveys, etc. It will also include much additional related information concerning ownership, payroll deduction accounts, legal, operating and reporting structures, etc.

Maintenance of the IP establishment data for long form surveys will make use of survey feedback payroll deduction (PD) data, tax data and profiling/nature of business enquiries so that, at the time they are produced, the long form frames will be as up to date as possible. The maintenance processes which result in addition or deletion of establishments and updating of classification or contact data for reference year Y will be as outlined in the following paragraphs.

(a) Survey feedback for year Y-1 will be used to update classification and reporting data and to flag establishments which have ceased economic activity. A few new establishments may also be identified.

(b) Information received from RCT concerning each newly opened PD account during year Y will be analyzed. If the potential number of employees or remittance data exceed certain threshold values (to be specified, see Section 5) and the account will be matched against all units in the IP. If no association can be found then a profiling/nature of business enquiry will be initiated to decide whether the PD account truly indicates the existance of an establishment which should be in the IP and is not. If a new establishment is justified the corresponding classification information, reporting arrangement and tax identifiers will be obtained and the data added to the IP.

(c) At least once during year Y the active file of PD accounts not currently associated with any estblishment in the IP will be scanned. Accounts for which remittance data exceed a certain threshold value (to be specified, see Section 5) will be identified and investigated as in (b).

(d) The T4/T4A summary file (for year Y-1) obtained from RCT in year Y will be used to update wage/salary and remittance data on the PD universe file. Accounts not currently associated with the IP for which wage/salary values exceed a certain threshold (to be specified, see Section 5) will be identified and investigated as in (b).

(e) The tax universe files for Y-2, or Y-1 if available, will be matched against the list of tax identifiers currently associated with the IP and non-matching tax return with gross business income above the threshold for inclusion in the IP (to be specified, see Section 5) will be identified and investigated as in (b).

(f) Late in year Y the (incomplete) tax response files for year Y-1 may be used in similar fashion to the tax universe files in (e).

CFDB: Non-Integrated Portion

Frame data relevant to annual surveys of production not maintained in the IP will be stored in the non-integrated portion (NIP) of the CFDB. These data will compromise the tax universe files (with long form survey duplicates flagged) and the tax response data used in the generation of frames for prespecified tax sampling.

Whereas the IP will be maintained on an ongoing basis, i.e. updated continually as new data become available, most tax data in the NIP will be updated annually, i.e. "replaced" rather than "maintained". During the course of year Y+1 the final tax universe file for Y-1 will be acquired and that for Y-2 will no longer be needed. In the same

year the tax response file for Y-1 will be utilized in generating the prespecified T1 frame file for year Y and prespecified T2 frame file for year Y+1.

4.3 Central Service Function (2)

The CSF(2) will be responsible for the selection, acquisition, capture, processing, weighting and dissemination of income tax response data.

Selection. The CSF(2) will administer the basic strategy for survey selection of prespecified T1 and T2 tax samples. It will consolidate the resulting list of tax returns to be sampled and transmit the file of identifiers to RCT together with the parameters for selection of the cross-sectional sample.

Acquisition and Data Capture. The CSF(2) will capture required response data from tax returns either by direct data entry from the returns as they pass through RCT assessing system or by copying the returns and capturing the data at STC. At the same time as data are captured CSF(2) will assign an SIC code. Data and codes will be recorded on (initial) weighted tax response files.

Processing. The CSF(2) will pass tax response data through edit and imputation, and geocoding routines.

Weighting. The CSF(2) will reweight the tax response data to remove, in effect, duplication with establishments covered by long form survey and to ensure the weights sum to universe totals. For this purpose the CSF(2) will require from the CSF(1):

(a) the identifiers of all tax returns associated with establishments subjected to sampling by long form survey, i.e. duplicates;

(b) the (unduplicated) tax universe file counts in each stratum to be used in weight adjustment.

Dissemination. The CSF(2) will transmit to each annual survey weighted tax response data for the appropriate industries together with measures of sampling variability.

All tax response data will be stored on a data base which will certainly be accessible to the CFDB, if not actually an integral part of it. The strategy regarding feedback of survey response data and/or corrections to tax response data has yet to be decided (see Section 5).

4.4 CSF: Summary of Processing and Data Flows

A schedule of major CSF processing events and related survey operations relevant to the production of annual survey data for reference year Y is shown in Figure 4. All items refer to reference year Y data unless otherwise specified. For brevity, certain intermediate and potential additional processing steps, for example, leading to the production of preliminary estimates, have been omitted.

The dates in Figure 4 are for illustrative purposes only. They were derived by consideration of the present Census of Construction operations with some improvements in tax and frame data processing. For example, it is assumed that a parameter driven T1 sampling program will be installed in the RCT Assessing system and that the file of identifiers for the T1 prespecified reference year Y sample will not be required until Feb. Y+1. In practice none of the dates involving tax data can be changed substantially without negociations with RCT. Survey specific dates, e.g. for mailout of long form and short form questionnaires, might vary somewhat from one survey to another although, in principle, they should be the same. Some survey operations may wish to produce preliminary estimates using tax files which are not fully complete.

Figure 5 indicates the flow of data and the data files relevant to the production of estimates for reference year Y. For simplicity, the flows files and dates for T2 tax data have been omitted. They follow the same pattern as for T1 tax data except that RCT processing begins nearly 12 months earlier and certain dates must be brought forward correspondingly.

5. Issues, Problems and Studies

In Sections 3 and 4 reference was made to a number of issues to be settled, techniques to be validated and problems to be solved. The following paragraphs elaborate upon these items and suggest the subprojects and activities within which they should be addressed.

Target Population (ref. Section 3.2)

The target population for annual surveys of economic production, has yet to be precisely defined. This includes the categories of self-employed income which are deemed to be in scope and the corresponding minimum size threshold.

Factors to be taken into account are:

(a) economic production by size of tax return for each class of tax return (corporate, business, farming, fishing, combined, etc.), as a proportion of total production, for each industry group covered by a major survey;

(b) availability of data on tax returns which fall within the range of sizes where the boundary is likely to be;

(c) cost of capturing and processing the required data from tax returns in this size range.

This problem will be addressed in Subproject No. 2 (Subpopulations).

Monitoring the Extent of Out-of-Scope Production (ref. Section 3.2)

Some economic production will be reported on tax returns falling below the boundary for direct survey coverage (as discussed in the previous paragraph). There is a need to monitor the extent of such production so that periodic adjustment can be made to the boundary if need be.

A rough breakdown by industrial sector is desirable to enable surveys to indicate, in footnote form, the appropriate magnitude of the corresponding undercoverage. It may be possible to obtain measurements of sufficient accuracy using totals provided by RCT and allocation by sector, taking a sample periodically to validate or adjust the allocation.

This problem will be addressed in Subproject No. 3M (activity a2).

Definition of Establishment Continuity (ref. Section 3.3)

Year-to-year comparison of establishment data for validation and imputation purposes requires a precise definition of establishment "continuity" over time. It must be stated under exactly what set of changes in size, location, activity, output, ownership, etc. an establishment is deemed to have "died" as opposed to "continued".

Factors to be considered are:

(a) the practical availability of data required to establish continuity or death (for each definition)

(b) the use to be made of year-to-year comparisons. Clearly the definition of continuity should reflect, in some sense, the validity of year-to-year comparison of data items.

This problem will be addressed in Subproject No. 2 (Concepts).

Survey Long Form - Short Form / Tax Boundary (ref. Section 3.2)

The boundary between coverage by long form questionnaire and by tax/short form has to be specified for each survey.

Factors to be taken into account are:

(a) the survey quality objectives for data items which are not available, or are conceptually different, on tax returns;

(b) the absolute and relative costs of processing data for the same units from long forms, short forms and tax returns;

(c) response burden.

It is not sufficient to make a statement of the type "long forms must cover 90% of all gross revenue/output by industry/commodity" without (quantitative) justification in terms of the quality of published estimates.

This problem will be addressed in Subproject No. 2 (Subpopulations).

Threshold Criteria for the IP/NIP Boundary (Section 4.2)

The criteria for inclusion in the long form component for each survey must be translated into criteria for creation and maintenance of units in the IP. In particular the thresholds for each data source defining the IP/NIP boundary have to be specified.

This task will be carried out in Subproject No. 3M (activity b1).

Overall Strategy for Tax Data Sampling and Estimation (ref. Section 3.4)

The overall strategy for tax data sampling and estimation is described in Section 3.4. It has been the basis for Census of Construction use of tax data for nearly 10 years. However the sampling variances of the estimates have never been determined and should be. Also, various possible improvements to the approach, and alternative approaches, as outlined in the following paragraphs should be systematically investigated.

(a) Consideration should be given to making more use of data captured by RCT on a universe basis for T2 returns (CORPAC) and on a sample basis for T1 returns (COMSCREEN) and T2 returns (tax model sample). It is conceivable that RCT-STC pooling and sharing of sample information could reduce cost and/or increase content for both agencies.

(b) Alternatives to the proposed cross-sectional / prespecified sampling strategy should be examined, for example the restriction of cross-sectional sampling to units not subject to prespecification.

This work will be undertaken in Subproject No. 3M (activity c1).

Strategy for SIC Code Maintenance (ref. Section 3.2)

The possibility of maintaining reliable industrial codes for the whole T2 universe should be considered. The advantages of such a procedure are that it would enable stratified estimates by industry in place of domain estimates and it would facilitate use of data captured by RCT for the universe. The disadvantage is the cost of adequate maintenance, or the bias which will result if maintenance is not adequate. The evaluation should include a comparison of the quality of estimates based on two different approaches which utilize the same total resources. In the first approach resources are divided between T2 universe, SIC code maintenance and coding and capture of the T2 sample; in the second approach they are all devoted to the sample.

It seems unlikely, in view of the very large number and industrial volatility of T1 tax returns indicating economic production, that maintenance of reliable industrial codes for the universe of T1 returns is desirable. However this perspective should be confirmed. The possibility of maintaining a crude industry sector coding for the universe by capture and automated coding of nature of business descriptions should be investigated.

This problem will be addressed in Subproject No. 3M (activity d1).

Frame for Prespecified Tax Sampling (Sections 3.4, 4.2)

Assuming that the cross-sectional / prespecified sampling stragegy is adopted a frame must be provided to each survey to enable selection of the prespecified tax sample. As units on this frame must be industrially clarified the frame itself must be drawn from tax sample response files for previous years or, possibly, from coded portions of tax universe files. The precise method of construction of such a frame has yet to be decided.

Factors to be taken into account are:

(a) there must be a sufficient number of units on the prespecified frame to enable selection of a sample of adequate size for any industry of importance to the survey which will not be fully covered by the cross-sectional sample;

(b) the classification data should be as current as possible to reduce the incidence of prespecified units which will be out of scope, e.g. which will have become inactive or changed industry;

(c) data storage considerations may suggest an upper limit to the number of years of sample response data which should be stored.

In combination these factors may suggest that the prespecified frame is restricted to units sampled in the most recent reference year, and that, for T2 returns, a second frame may be required for selection of a supplemental sample when sample data for the following reference year become available.

These ideas will be investigated in Subproject No. 3M (activity C).

Procedures for Acquisition of Tax Data (Section 3.4)

Tax returns will be sampled as they pass through the RCT Tax Assessing systems. The sampled returns will be assigned an SIC code and data extracted and stored on the tax response files. Presently, T1 returns are microfilmed and T2 returns are photocopied and the facsimilies sent to STC for data extraction and capture.

The possibility of extraction and data capture directly from T2 returns as suggested in the past by Valiquette and Adams[5] must be investigated. Provided quality can be maintained, adoption of such a procedure would have a significant benefit by vastly reducing the need for and cost of photocopying.

T1 returns are processed by RCT at Regional Centres so the opportunity for on line data extraction and capture by STC personnel does not present itself. The alternatives to be investigated in this situation are that RCT be requested to capture some additional data items such as nature of business information for all T1 returns or that RCT capture the data required from the sample on behalf of STC.

These possibilities will be addressed in Subproject No. 1 (Phase IV).

Survey Short Form Sampling (ref. Sections, 3.2, 3.5)

Each annual survey of economic production spans a particular range of industries, not the whole industral spectrum. The set of all tax-based establishments will complement long form establishments in providing complete coverage of the universe for any given survey. However, as members of the set of tax-based establishments will not all be industrially classified, it will not be possible to partition this set by industry and obtain explicitly the component which complements each individual survey. Instead, the weighted tax sample of all establishments coded to a survey will be used both to provide (domain) estimates of the number of tax-based establishments in the survey universe and to furnish a frame for short form sampling. This is the justification for the assertion in Sections 3.2 and 3.5 that the short form sample for any survey should be a subsample of the tax sample coded in scope for that survey. It raises two issues which must be addressed.

As the procedure represents a marked departure from current practice for the Census of Manufactures, it must be investigated and validated in this specific context. The work should be addressed in Subproject No. 4.

The best method for subsampling the tax sample should be determined. The basic alternatives are to:

(a) subsample the prespecified tax sample selected by the survey;

(b) subsample the prespecified tax sample together with any other establishments already tax sampled and available at the time of subsampling.

This work should be addressed in Subproject No. 3M (activity a2).

Estimation for Survey Short Form and Tax Subsample Data Items (ref. Sections 3.4, 3.5)

Certain data items will be acquired from short form questionnaires and from subsampling tax returns for additional industry specific information. Given that in both cases data will be available for a subsample of the overall tax sample the question is now best to form estimates. Three possible alternatives present themselves:

(a) Use "classical" [6, ch 12] two-phase estimation procedures which will provide unbiased estimates and measures of sampling variance but which will not utilize empirical relationships between data items to improve efficiency;

(b) Use model-based procedures to improve efficiency but with the risk of introducing bias into the estimates of both totals and corresponding sampling variances;

(c) Use mass imputation procedures along the lines of the present Census of Construction to improve efficiency by use of empirical by use of empirically established relationships without the need for explicit modelling, but with the risk of introducing bias, and with no known method for computing sampling variance.

This problem will be addressed in Subproject No. 3M (activity a2).

Maintenance of Tax Sample Response Data Base (ref. Section 4.4)

All tax data will be stored on a data base which will be accessible via the CFDB if not actually part of it. (The physical storage and access procedures will be determined in Subproject No. 3S.) The strategy regarding feedback of corrections or supplements to the tax data from survey sources has yet to be decided.

Factors to be taken into consideration are:

(a) provision of facilities for updating using survey feedback would add to the complexity of the systems design;

(b) updates from a survey which did not affect sample weights nor cause reclassification from that survey sector to another could be handled relatively easily as they would not involve other surveys;

(c) updates which cause changes in sample weights and/or reclassification of units from one survey sector to another would be more difficult to deal with as they would affect more than one survey and could lead to an iterative sequence of updates;

(d) the data base could be the focal point for production of small business statistics, and for survey data - tax data comparison and calibration; in this role the facility for feedback and storage of survey data would be very useful.

These issues will be addressed in Subproject No. 3M (activity C).

Roles of CSF, Survey and Headquarters Operations (ref. Section 4.1)

The extent to which survey sampling, mailout, data capture, follow-up and editing operations will be organized centrally and the relationship between the CSF, the surveys and Headquarters Operations have not yet been decided. This issue will be addressed in Subproject No. 4o.

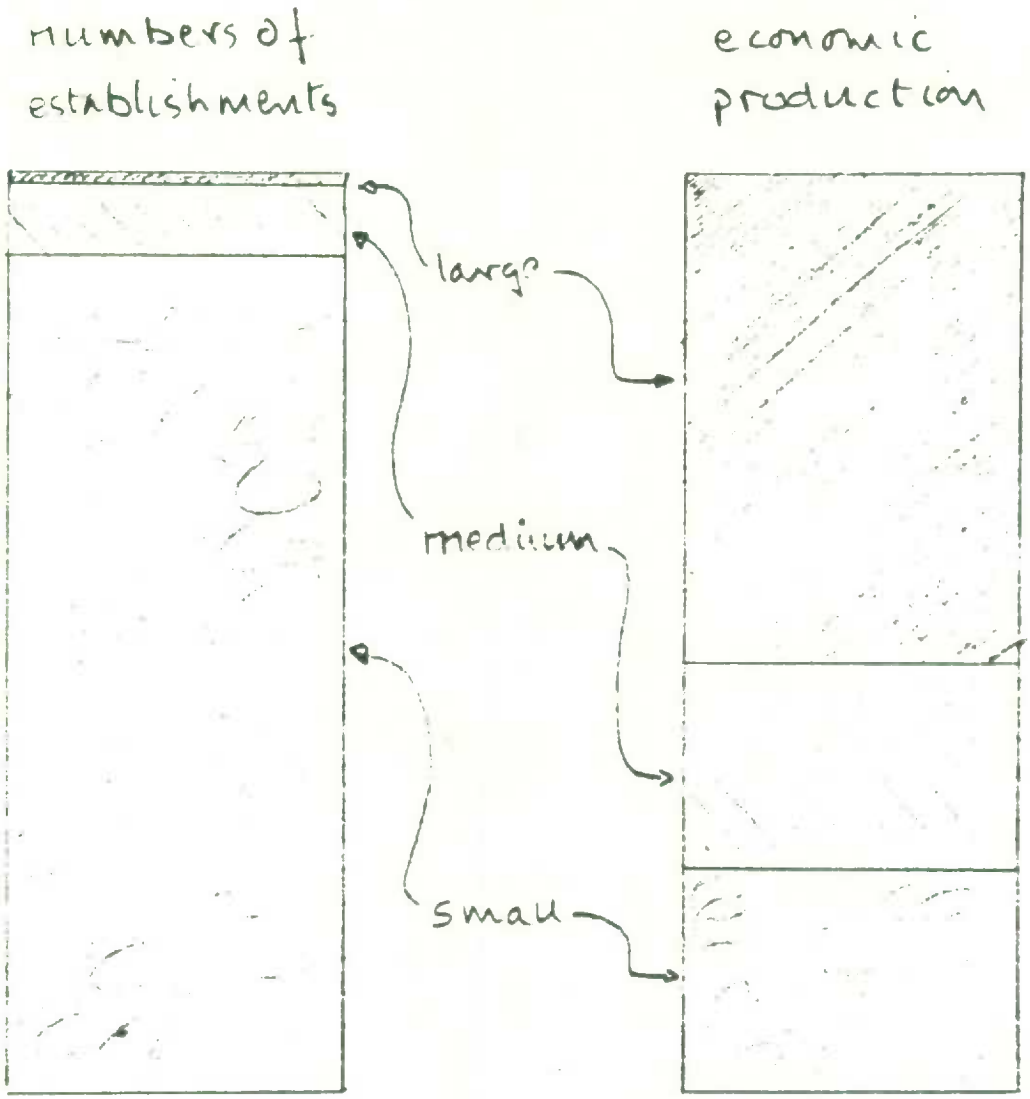


FIGURE (1) (Conceptual Example)

Comparison of numbers of establishments by size with corresponding economic production

SEPH-C001-P05
STATISTICS CANADA - SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS
COMPANY DATA BASE
SEPTEMBER, 1984
NUMBER OF COMPANIES AND NUMBER OF EMPLOYEES
(WEIGHTED AND UNWEIGHTED)
BY COMPANY SIZE
CANADA

COMPANY SIZE (as defined by unique 7 digit BRID)	WEIGHTED				UNWEIGHTED			
	NUMBER OF COMPANIES	%	NUMBER OF EMPLOYEES	%	NUMBER OF COMPANIES	%	NUMBER OF EMPLOYEES	%
ALL COMPANIES	562,871	100.0	8,805,595	100.0	43,023	100.0	5,929,987	100.0
LESS THAN 20	517,357	91.9	1,822,307	20.5	26,423	61.4	110,061	1.9
20-49	27,736	4.9	826,828	9.3	6,093	14.2	194,024	3.3
50-99	9,337	1.7	634,214	7.1	3,793	8.8	265,186	4.5
100-199	3,931	0.7	551,543	6.2	2,445	5.7	344,407	5.8
200 AND OVER	4,410	0.8	5,050,703	56.8	4,269	9.9	5,016,309	84.6
		8.1%		79.5%				
NO EMPLOYEES	128,210	22.8	0	-	7,083	16.5	0	-
1-9	416,933	74.1	2,649,135	29.8	25,433	59.1	304,085	5.1
10-19	15,016	2.6	1,956,289	22.0	8,566	20.0	1,345,731	22.7
20 AND OVER	1,912	0.3	4,200,171	48.2	1,912	4.4	4,200,171	72.2
NO EMPLOYEES	128,210	22.8	0	-	7,083	16.5	0	-
1-4	247,270	43.9	543,003	6.1	10,027	25.2	23,851	0.4
5-9	91,037	16.2	592,063	6.7	4,467	10.4	29,459	0.5
10-19	50,040	9.0	687,241	7.7	4,046	9.4	56,751	1.0
20-49	27,736	4.9	826,828	9.3	6,093	14.2	194,024	3.3
50-99	9,337	1.7	634,214	7.1	3,793	8.8	265,186	4.5
100-199	3,931	0.7	551,543	6.2	2,445	5.7	344,407	5.8
200-499	2,498	0.4	770,532	8.7	2,357	5.5	735,138	12.4
500-999	953	0.2	666,652	7.5	953	2.2	666,652	11.2
1000-1499	325	0.1	396,734	4.5	325	0.8	396,734	6.7
1500-2499	291	0.1	563,367	6.3	291	0.7	563,367	9.5
2500-4999	203	-	690,472	7.8	203	0.5	690,472	11.6
5000 AND OVER	140	-	1,962,946	22.1	140	0.3	1,962,946	33.1

Remark:
79.5% of SEPH employment is accounted for by 8.1% of companies

Fig 2. Distribution of companies by size and corresponding employment

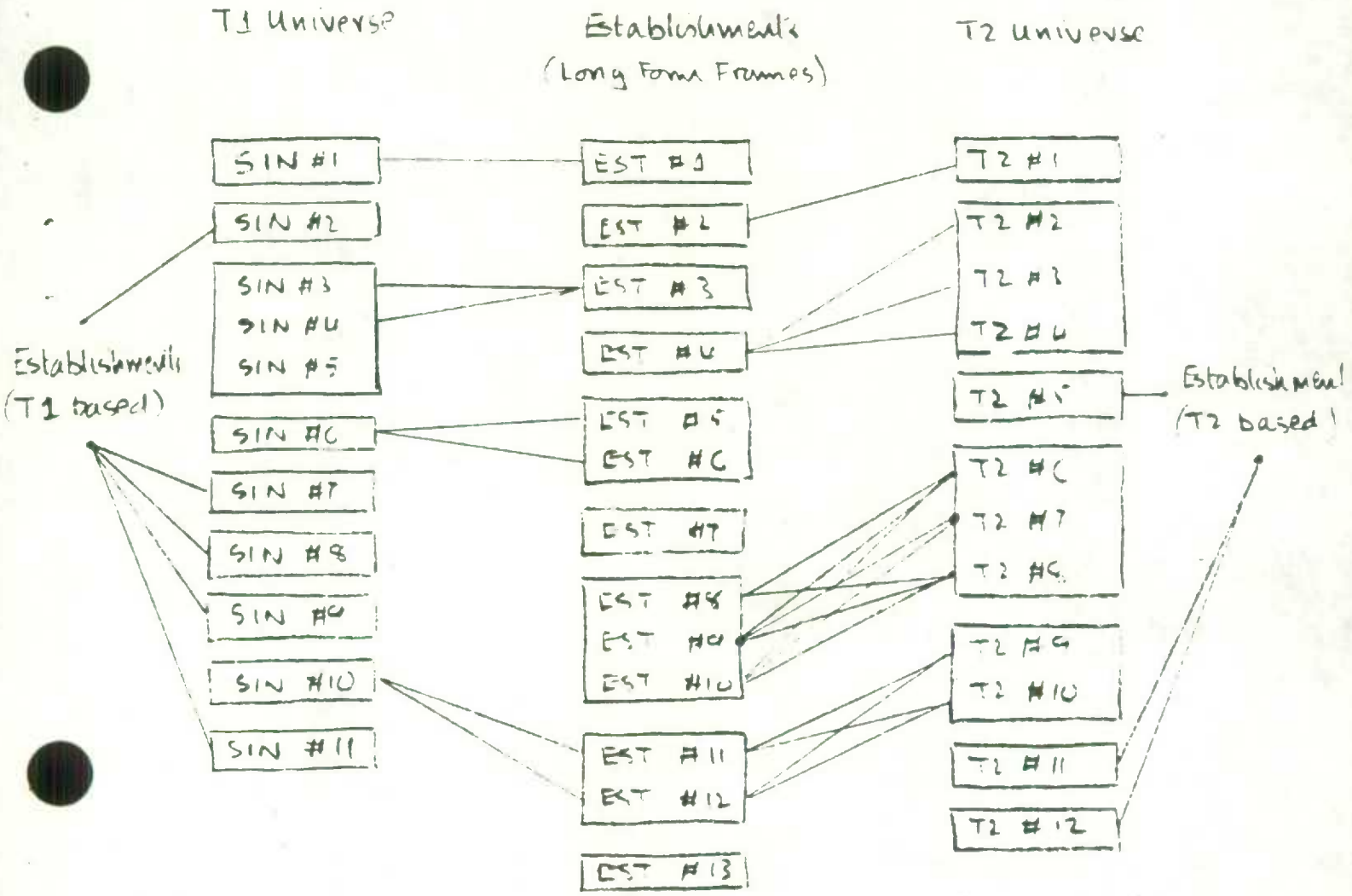


FIGURE (3) Linkages between tax universe files and long form establishments

{Note : tax units not linked to long form establishments constitute the tax-based establishments}

Fig 4. Summary of CSF Processing Relevant to

The Production of Annual Data for Reference Year Y.

(sheet 1/3)

Date	Function	Input Data	Action	Output/Result
Dec Y-1	CSF(1)	T2 tax response file and tax universe file for Y-1 (incomplete), Y-2	create frame for T2 prespecified tax samples	T2 prespecified frame for Y
Jan Y	Survey Ops	T2 prespecified frame	select T2 prespecified samples	T2 prespecified sample for Y
Feb Y	CSF(2)	T2 prespecified samples	consolidate T2 prespecified samples	T2 prespecified sample file for Y to ECT
March Y - Dec Y+1	CSF(2)	T2 tax returns	capture T2 response data for prespecified and cross sectional samples	T2 tax response file for Y
Jan Y - Dec Y	CSF(1)	PD20, PAY/DAC and TL/TL4A data	identify updates for IP	update establishment and linkages in IP
March Y - Dec Y	CSF(1)	survey feedback for Y-1	identify updates for IP	update establishments and linkages in IP
March Y	CSF(1)	T1 tax universe for Y-2	identify update for IP	update establishments and linkages in IP
April Y	CSF(1)	T2 tax universe for Y-2	identify updates for IP	update establishments and linkages in IP
Jan Y - Dec Y	CSF(1)	profiling and nature of business data	identify update for IP	update establishments and linkages in IP
Nov Y	CSF(1)	T1 sample response and universe files for Y-1 (incomplete)	identify updates for IP (optional)	update establishments and linkages in IP
Nov Y	CSF(1)	T2 sample response and universe files for Y-1 (incomplete)	identify updates for IP (optional)	update establishments and linkages in IP

Fig 4 (Sheet 2/3)

Date	Function	Input Data	Action	Output / Result
Dec Y	CSF(1)	IP establishments active in Y	create frame for long form samples	long form frame for Y
Jan Y+1	survey ops / CSF(1)	long form frame for Y	select long form samples	long form samples for Y
Feb Y+1	survey ops CSF(1)	long form samples for Y	obtain contact info from IP	mailout of long form questionnaires for Y
Jan Y+1	CSF(1)	T1 tax response file for Y and possibly earlier years	create frame for T1 prespecified tax samples	T1 prespecified frame for Y
Feb Y+1	CSF(2)	T1 prespecified samples for Y	consolidate T1 prespecified samples	T1 prespecified sample file for Y to RCT
Jan Y+1	CSF(1)	T1 and T2 prespecified samples for Y and possibly T2 response file for Y (incomplete)	create frame for short form samples in NIP	short form frame for Y
Feb Y+1	survey ops	short form frame for Y	select short form samples	short form samples for Y
March Y+1	survey ops CSF(1)	short form samples for Y	obtain contact info from NIP	mailout of short form questionnaires for Y
March Y+1 - Oct Y+1	survey ops.	long form questionnaires	capture, edit, impute, follow up, weight	weighted long form response file for Y
March Y+1 - Oct Y+1	survey ops.	short form questionnaires	capture, edit, impute follow up	short form response file for Y
March Y+1 - Dec Y+1	CSF(2)	T1 tax returns	capture T1 response data for prespecified and cross sectional samples	T1 tax response file for Y

Fig 4 (sheet 3/3)

[redacted]e	Functional	Input Data	Action	Output / Result
Feb Y+2	CSF(1)	T1 tax universe file from RCT	identify T1 long form duplicates	files of T1 long form duplicates and (unduplicated) T1 universe control counts and totals for Y
Mar Y+2	CSF(2)	files of T1 long form duplicates and (unduplicated) T1 universe control counts and totals for Y; T1 tax response file for Y	reweight T1 tax response file	weighted T1 tax response file for Y (unduplicated against long form universe)
Mar Y+2	CSF(1)	T2 tax universe file from RCT	identify T2 long form duplicates	files of T2 long form duplicates and (unduplicated) T2 universe control counts and totals for Y
Apr Y+2	CSF(2)	files of T2 long form duplicates and (unduplicated) T2 universe control counts and totals for Y, T2 tax response file for Y	reweight T2 tax response file	weighted T2 tax response file for Y (unduplicated against long form universe)
Apr Y+2	CSF(2)	weighted T1 and T2 tax response files for Y	merge T1, T2 files create survey specific files	weighted combined T1, T2 tax response file for Y for each survey unduplicated against long form universe.
Apr Y+2	Survey OPS	weighted unduplicated tax response file short form response file long form weighted response file	merge files, generate estimates	estimates for universe

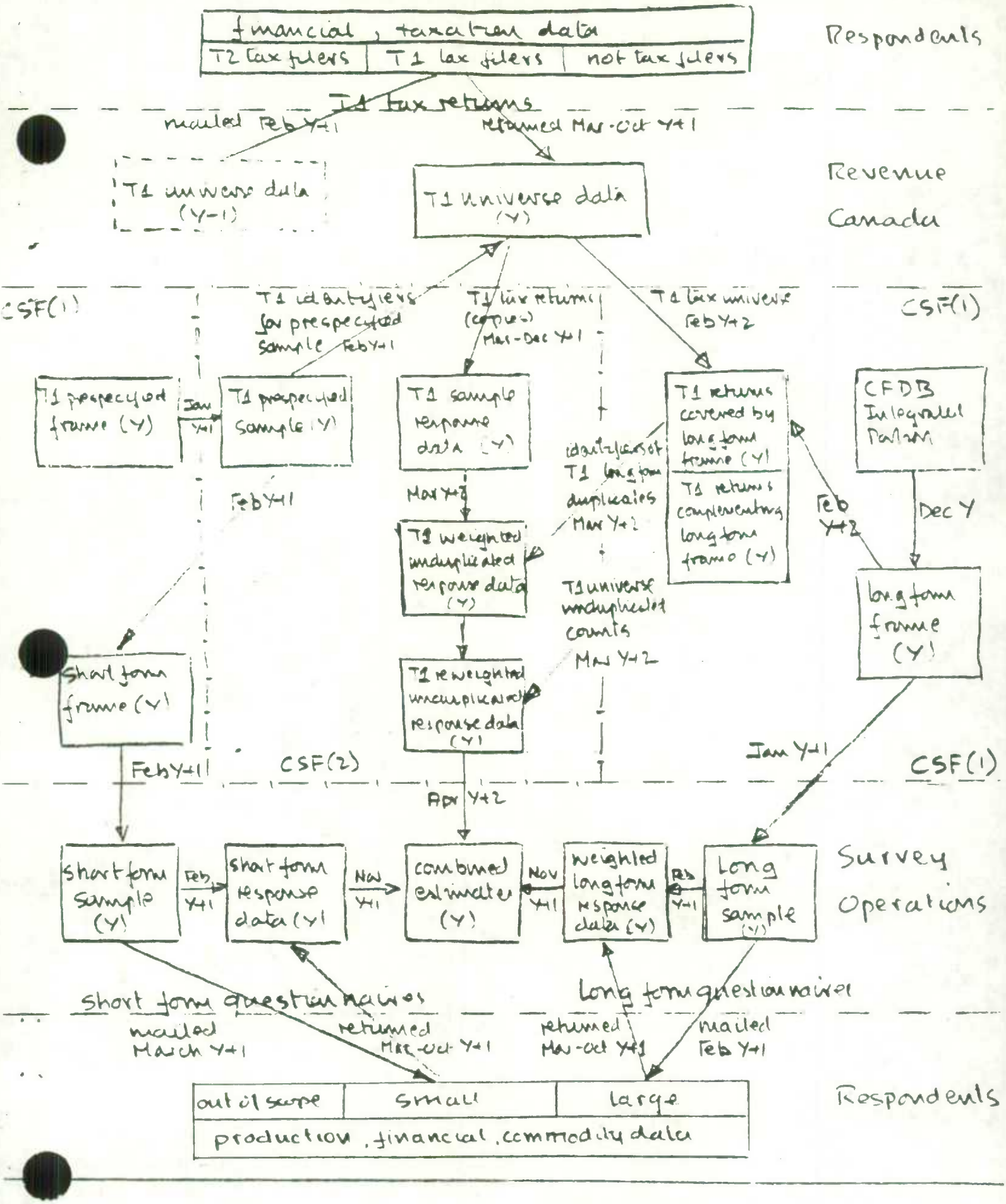


FIGURE (5) ANNUAL SURVEY DATA FLOWS (T2 not shown)

REFERENCES

[1] Infrastructure Project

[4] Standard Industrial Classification 1980, Statistics Canada

[5] Jcmt RCT/STC Study Reports. Valiquette and Adams. 1982

[6] Sampling Techniques, W G Cochran, 1977