FRAME CREATION FOR INSTITUTIONAL SURVEYS AT STATISTICS CANADA

L. Swain, M. Brodeur, S. Giroux, K. McClean, J. Mulvihill, J. Smith, Statistics Canada
L. Swain, Social Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6

KEY WORDS: Institutional surveys; frame creation and evaluation; Business Register

1. OVERVIEW

In examining the creation of frames for institutional surveys at Statistics Canada, this paper presents an overview of the institutional surveys program with examples of the frames used. The basic underlying themes in frame creation and evaluation across the various subject matters are emphasized and future initiatives are suggested for the improvement of frame creation and evaluation.

2. INSTITUTIONAL SURVEYS

In order to examine institutional survey frames at Statistics Canada, it is first necessary to clarify what is meant by institutional surveys. Combining definitions contained in the Oxford English Dictionary and the Concise Oxford Dictionary, an institution may be defined as an establishment, organization or association of public or general utility and founded especially for educational, charitable, religious or social purposes (e.g., a school, college, hospital, asylum, reformatory or the like). Basically and simply stated, institutions provide a public service or social benefit.

At Statistics Canada, the institutional surveys program is generally seen as comprising surveys in the subject matter areas of health, justice, education, culture and public institutions, for which organizationally the following divisions are responsible: the Canadian Centre for Health Information; the Canadian Centre for Justice Statistics; the Education, Culture and Tourism Division; and the Public Institutions Division.

Although most surveys carried out by these divisions are of institutions and the services they provide or use institutional frames to collect data about individuals, their programs also include some household surveys and some business surveys to collect their subject matter. On the other hand, institutions are included within some surveys done elsewhere in Statistics Canada. Thus, a definition based on an organizational division of subject matter does not coincide with one based on target population or frame unit.

For the purposes of this paper, institutional surveys include those surveys at Statistics Canada with institutions as the target population (in part or in total) or which use institutions as a frame to collect data about or from individuals.

Although these surveys are individually well documented, very little attention has been paid to examining methodological aspects of institutional surveys as a whole. Besides being generally informative, identification of commonalities and differences in methodologies among the various surveys may generate ideas for improvement in approaches for individual surveys. This paper focuses on the creation of frames for institutional surveys.

The next section presents examples of the variety of frames used for institutional surveys with reference made to specific surveys and several subject matters. Section 4 contains the underlying themes that arise from the examination of the institutional survey frames. In section 5, possible future initiatives follow from these underlying themes.

3. INSTITUTIONAL FRAMES

A variety of frames are used for institutional surveys at Statistics Canada. Since space precludes a complete listing and discussion of all these frames, a few examples are provided to give a flavour of their diversity.

In the subject area of health, the frame for the "Survey of Residential Care Facilities" comprises lists from provincial and territorial (hereafter denoted as "provincial" for simplicity) government ministries of health and/or social services. The facilities are generally those that are approved, funded or licensed by these ministries and are restricted to those with four beds or more. For the "Survey of Registered Nurses", the frame is the set of provincial nursing associations. The data collected for this survey are obtained as a result of the process of registration of nurses with these associations.

The post-censal "Survey of Health and Activities Limitations" acquired the institutional component of its frame from the list of collective dwellings provided by the 1991 Canadian Census of Population. Six types of institutional collective dwellings were included in the survey: chronic care residences, nursing homes, residences for senior citizens, general hospitals, psychiatric institutions and centres of the physically handicapped. The institutional frame was used as a first-stage sampling frame for a selection of institutions from which the frame for the target population (residents of the sampled institutions) was obtained.

In the justice area, the frame for the "Adult Criminal Court Survey" is the set of provincial ministries responsible for courts (e.g., Ministry of the Attorney General). Case characteristics are extracted from the automated court files of these ministries and sent to Statistics Canada.

In the educational subject matter area, the survey of "Elementary-Secondary School Enrolment" has as a frame the provincial ministries of education (for public and most private schools), the federal Department of National Defence (for schools on Canadian military bases overseas), the federal Department of Indian and Northern Affairs (for Indian and Inuit students) and federations of independent schools. Schools for the blind and deaf are also included. The "Survey of Private Business Schools" also uses lists from the provincial ministries of education. However this survey also obtains lists from the federal Department of Employment and Immigration through a certification process for private business schools necessary for tax purposes.

For culture, the "Heritage Institutions Survey" uses the Directory of Canadian Museums and Related Institutions from the Canadian Museums Association as well as information from provincial museum associations and from contacts in federal or provincial ministries responsible for areas such as parks, archives and communications.

In the area of public institutions, the "Public Sector Financial System" uses federal and provincial government Public Accounts so that the frame is the ministries of finance. For data on special funds and certain agencies, boards, commissions, financial statements from various government organizations are obtained, which are then part of the frame. For the "Waste Management Survey", the frame is the set of local governments (generally municipalities).

Other frames include units such as libraries, transition homes, cancer registries, registrars of vital statistics, cultural associations, universities and colleges.

Institutional frames may be used to survey institutions and the services they provide (the target population is institutions); to survey institutions to obtain information about the registrants, users or clients of the institutions without actually contacting the individuals (the target population is people but the frame units are institutions); or to survey the registrants, users or clients of the institutions (the target population is people). In the last case, the institutional frame is used as a means to identify the individual persons in order to contact them directly.

4. UNDERLYING THEMES

Institutional surveys are very much like business surveys (i.e., the data collected may be about the institution where the content may be finance, employment, salaries or the services that the institution provides but could also be about the registrants, users or clients of the institution where the content may be vital statistics, cancer events, crime incidents, types of residents of transition homes). In fact, the "business" of an institution is service. The service provided (e.g., education of elementary school children or medical treatment in a hospital) can be viewed as analogous to that provided by service industries in the business sector. These institutional services are generally referred to as non-marketed services (Statistics Canada 1993).

Institutional frames and the information about the institutions and their registrants, users or clients are often based on administrative data that exist because of funding or licensing arrangements, registration or a government statute.

Many of the frames are multi-sourced, i.e., multiple frames. The Elementary-Secondary School Enrolment survey, the Heritage Institutions Survey, the Public Sector Financial System described above all demonstrate the multiplicity of sources and the "detective" work required to build a complete frame.

Given provincial jurisdictions in health, justice, education, culture and public institutions, many of the frames are at least in part provincially-based. As a result, there may be variations among provinces in the content and definitions associated with the frames.

There are often differences in definitions of inclusion on the frame (e.g, mandatory vs. non-mandatory registration). Differences in the legal basis for the existence of the institutions will also affect inclusion. The source of the frame may differ (e.g., automated vs. hard-copy records) as may the timing of updates to the frame (e.g., births and deaths). The variables which are maintained about the institutions or their registrants, users, clients (e.g.,family-based vs. person-based health number; event-based vs. person-based diagnostic information) may vary across the provinces as may their definitions (e.g., secondary school). The provincial jurisdictions may be reluctant to provide some information to Statistics Canada (e.g., names, addresses and telephone numbers of students to use as a frame). In compiling national statistics, these differences affect coverage, quality, accessibility and comparability.

Except in cases where there is a simple and obvious frame (e.g., provincial ministries responsible for courts), coverage evaluation is usually not done (in this example, the evaluation is trivial). However, even in such a simple case, the provincial ministries usually collate data from institutions (e.g., the courts themselves) and therefore have their own frame of institutions. In such a case, Statistics Canada does not generally perform an evaluation of the coverage of the provincial frame. In other situations, coverage evaluation is often very difficult to do as some surveys have their own master lists of institutions obtained by collating lists together, lists which are each often incomplete frames for the target population. Whenever a new source (often incomplete as well) enters the picture, comparisons are made with the existing frame and new units are added. There are few independent checks possible.

The stability of institutional survey frames ranges from the simple and very stable (e.g., the provincial ministries of education) to the sublime and less stable (e.g., the multiple frame for heritage institutions).

Generally, institutional frames (e.g., list of universities and colleges) are more stable than frames at lower levels when the institutional frame is used as an earlier stage in a multi-stage survey (e.g., lists of graduates obtained from universities and colleges).

The correspondence between Statistics Canada's integrated Business Register and some institutional frames is not generally good (e.g., Survey of Private Business Schools, Survey of Residential Care Facilities). This occurs because the Standard Industrial Classification (SIC) used to classify units for the Business Register is primarily business oriented and not institutions oriented. However, for government institutions, there is an annual reconciliation between the Business Register and the Public Institutions Division on federal, provincial and local governments.

Reconciliation of frames across surveys with similar target populations doesn't seem to be done (e.g., Residential Care Facilities Surveys and the Census institutional collective dwelling frame). Even with differences in the definitions of target populations and in frequency of collection, some reconciliation should be possible.

5. POSSIBLE FUTURE INITIATIVES

Following from these underlying themes for institutional frame creation, a few possible future initiatives arise.

The first initiative would be to examine further the relationship between the Business Register and various institutional frames to determine correspondence (or lack thereof); to identify what would be required to improve correspondence; to discover how the Business Register (with or without changes to incorporate institutional survey needs) could be used as a frame or for evaluation purposes for institutional surveys. Perhaps, better direct coverage or unduplication of effort could be achieved. Should the framework of the Business Register not meet the needs of institutions, an "Institutional Register" may be worth considering.

Secondly, given the expected 1997 revisions to the Standard Industrial Classification (SIC) system (recently underway as a project), it is an opportune time to explore an improved classification of institutions. Considerations already underway to improve classification of specific marketed service industries in the business area (Statistics Canada 1993) could be expanded to include the institutional services sector. By incorporating institutional survey needs better into the SIC, frame creation and use of the Business Register may be improved for these surveys.

Thirdly, where reconciliation of frames across surveys with similar target populations is not done, it should be considered.

For the future, it may be worthwhile to examine other

topic areas across institutional surveys, e.g., use of administrative data for collection, quality assurance, record linkage activities (exact and probabilistic), disclosure avoidance. Again, the intent would be to discover any underlying themes across the broad spectrum of institutional surveys from which concrete initiatives may follow.

In conclusion, the goal of this paper was to examine institutional survey frames across various subject matters at Statistics Canada and see what happened. The approach was more concerned with the "forest" rather than the "trees". Such an exercise along with a pursuit of the initiatives outlined above may or may not lead to a consolidated approach to institutional surveys at Statistics Canada. However, if pursued, it should lead to improvements in at least some of the institutional surveys. After all, good "forest" management does lead generally to better "trees".

ACKNOWLEDGEMENTS

The authors wish to thank the many persons who provided us with information about institutional surveys and their frames. Detailed documentation is available on the numerous institutional surveys conducted at Statistics Canada. For further information, please contact the divisions mentioned in the paper or the authors. The authors also wish to thank Bryan Lafrance, Jean-Louis Tambay and M.P. Singh of the Social Survey Methods Division for their useful comments on draft versions of the paper.

REFERENCE

Statistics Canada (1993). Data Gaps - Services: Introductory Notes for a Presentation to the National Statistics Council. Statistics Canada.

FRAME CREATION FOR THE SURVEY OF MINORITY-OWNED BUSINESS ENTERPRISES AND THE SURVEY OF WOMEN-OWNED BUSINESSES

Mark S. Sands
U. S. Bureau of the Census, Washington, DC 20233

Overview

The Survey of Minority-Owned Business Enterprises (SMOBE) and the Survey of Women-Owned Businesses (WOB) are conducted by the United States Bureau of the Census as part of the quinquennial Economic Censuses. The WOB survey collects data for Women-Owned Businesses. The SMOBE collects data for Black-Owned Businesses, Hispanic-Owned Businesses, and businesses owned by Asian Americans, American Indians, and Other Minorities. These surveys provide basic information on the total number of firms owned by minorities and women, the total receipts generated by these firms, the number of firms with paid employees, the number of employees, and the total payroll for these firms. Data published include aggregate totals for the United States (US) by industry classification, and aggregate totals for each individual State and for smaller areas within States (e.g., county, place, or Metropolitan Statistical Area). Data are also presented by legal form of organization, employment size, and receipt size. The data are presented in four separate reports, and a summary report that includes data from the three SMOBE surveys.

The Challenge

The big challenge (and the thing that makes SMOBE and WOB interesting surveys to work on) is "the needle-in-the-haystack" problem -- trying to make detailed estimates for a relatively small portion of a large population of firms.

A sample large enough to make estimates at the desired detail levels would be very expensive. However, if a methodology could be developed to pre-identify as many minority-owned firms as possible before sample selection, the costs could be held down and (hopefully) the desired detailed estimates for the minority-owned firms could be made.

This paper presents a brief history of the processes and procedures followed to create the frames for the SMOBE and WOB, from the first special study conducted in 1969 through the latest completed surveys conducted during the 1987 Economic Censuses.

The 1969 Special Study -- the first step

Frame Development

The Bureau received a file from the Internal Revenue Service (IRS) of basic information for businesses required to file one of the following tax forms: 1040C (sole proprietors), 1065 (partnerships), or 1120S (small corporations whose owners chose to be taxed as individual shareholders rather than as a corporation). This information included the name, address, and employer identification number of the firm, the social security numbers (SSNs) of the owners, partners, or shareholders (up to 10 owners for partnerships and small corporations), dollar receipts, and the principal business activity code.

The Bureau created a file of the SSNs of all business owners from the file received from the IRS. This SSN file was sent to the Social Security Administration (SSA). The SSA matched the SSN file against their files to get race information that was supplied by individuals when they applied for their SSN. Then the SSA sent the SSN file with the race information for each business owner back to the Census Bureau. This information was merged with the data received from the IRS.

The race code information received from the SSA had three categories: White, Black, and Other race. Using the race code information, the Bureau split the file of businesses into two separate files:

1. The BLK frame contained all 1065 and 1120S firms in which 50% or more of the owners/shareholders had a race code of "Black," and all 1040C cases

where the filer (or first) SSN had a race code of "Black."

2. The REM file contained all cases that were not placed on the BLK frame.

Estimates of the number of Black-Owned businesses were made by tabulating all firms on the BLK frame. No cases placed on the BLK frame were passed through any further processing. (This meant that Black-Owned firms could not also be classified as Hispanic-Owned or Other Minority-Owned.)

The method of identifying firms owned by persons of Hispanic ancestry was not as straightforward as for the Black-Owned firms. Since the race codes received from the SSA did not have indications of Hispanic ancestry, a different technique was developed.

The Bureau had developed a list of Hispanic surnames from the 1960 Decennial Census. Persons with these surnames were determined to have a very high probability of being of Hispanic descent. This list of surnames was matched against the REM file. Firms in which at least one owner had a surname that matched a surname on the Hispanic surname list were removed from the REM file and were placed on the Hispanic frame (HSP). Cases not placed on the HSP frame, but having at least one owner with a race code of "Other" were placed on the Other Minority frame (OTH). Cases not placed on either the BLK, HSP, or OTH frames were placed on the National frame (NAT).

A mail canvass of all cases on the HSP and OTH frames was conducted to determine the Hispanic ancestry (Mexican-American, Puerto Rican, Cuban, or Latin American) or Other minority ancestry (Japanese, Chinese, Filipino, Hawaiian, Korean, and American Indian). A mail survey was also conducted on a small sample of the cases placed on the NAT frame in order to make estimates of Hispanic-Owned firms whose owners had surnames that were not on the Hispanic surname list. The economic data (receipts, employment, etc.) were obtained from the 1967 Economic Census.

Minority-Owned Large corporations (nonsubchapter S) were identified from other government agencies, public sources, and contacts with minority development agencies. Data from these corporations were obtained from the 1967 Economic Census. No systematic or scientific/statistical methodology was implemented for making estimates for these cases (i.e., collecting data for large corporations was more of a hit-or-miss operation). The data for these large corporations were included in the published tables.

Survey Results[1,2]

Hispanic-Owned Estimates -- Estimated number of firms was 100,212 (with approximately 22,000 or 22.0% contributed from the NAT frame sample). Estimates from the NAT frame were published in a "Not allocated by ..." category for each table. Detail estimates for specific 4-digit SICs included only cases originally on the HSP frame. Hispanic-Owned firms accounted for 1.3% of all US firms.

Other Minority-Owned Estimates -- Estimated number of firms was 58,673 (no estimates of undercount were attempted from the NAT frame sample). All cases in the OTH frame were mailed survey forms. Those cases responding as minority-owned were tabbed. Although collected on the survey form, the detailed breakdowns of the Other Minority groups (i.e., Chinese, Japanese, American Indian, etc.) were not published. Other Minority-Owned firms accounted for 0.8% of all US firms.

Black-Owned Estimates -- Number of firms tabulated was 163,073 (cases were tabbed based on the race codes received from SSA). Black-Owned firms accounted for 2.2% of all US firms.

The 1972 SMOBE and the 1972 Special Report on Women-Owned Businesses

Frame development

The methodology for producing the different frames

---

1 Making comparisons of firm estimates from one survey year to another survey year is not valid. The criteria for inclusion in the SMOBE and WOB have changed from survey to survey. (Two such changes: 1) industrial coverage has increased, and 2) minimal receipts requirements have varied from survey to survey.) Please refer to the individual SMOBE or WOB publications for more detailed information. The in-scope criteria within a given survey year are consistent, so estimates for specific minority groups within a given survey year are comparable.

2 Estimates presented are original published estimates and do not include subsequent revisions/adjustments.

was similar to that used in the 1969 special study. There were two major changes/differences.

1. The SSA provided race and sex codes for each business owner's SSN. This additional information made possible the estimates of the number of Women-Owned businesses.

2. The universe of sole proprietors was matched to the sole proprietor responses from the 1969 special survey. If a firm was still active in 1972 and had responded as Hispanic or Other Minority in 1969, that firm was selected with certainty, and the race, ethnicity, and ancestry response codes from the 1969 survey were moved forward to the 1972 record (the case was not mailed a survey form, but it was included in the tabulations of the 1972 survey data). Receipts and employment data were obtained from the 1972 Economic Census.

Minority-Owned and Women-Owned Large corporations were identified and handled as they were in the 1969 survey.

Survey Results[3]

Hispanic-Owned Estimates -- Estimated number of firms was 120,108 (with approximately 34,000 or 28.3% contributed from the NAT frame sample). Estimates from the NAT frame were published in a "Not allocated by ..." category for each table. Detail estimates for specific 4-digit SICs included only cases originally on the HSP frame. Hispanic-Owned firms accounted for 1.4% of all US firms.

Other Minority-Owned Estimates -- Estimated number of firms was 66,841 (no estimates of undercount were attempted from the NAT frame sample). Other Minority-Owned firms accounted for 0.8% of all US firms.

Black-Owned Estimates -- Number of firms tabulated was 194,986 (cases were tabbed based on the race codes received from SSA). Black-Owned firms accounted for 2.2% of all US firms.

Women-Owned Estimates -- Number of firms tabulated was 402,025 (cases were tabbed based on the sex codes received from the SSA). Women-Owned firms accounted for 4.6% of all US firms.

3 See footnotes 1 and 2.

The 1977 SMOBE and WOB Survey

Frame Development

There were two major changes in the methodology for developing the frames for the 1977 SMOBE and WOB.

1. Based on responses to the 1970 Decennial Census, the Population Division of the Bureau developed two lists of Hispanic Surnames (a "short" list of about 5,000 Hispanic surnames, and a "long" list of about 8,000 Hispanic surnames). The Population Division provided these lists to the SMOBE staff. For the 1977 SMOBE, all owners were matched against the "short" list and were placed on the HSP frame if an owner's surname was on that list. Only owners living in Texas, Colorado, New Mexico, Arizona, and California were matched against the "long" list. The major difference between the "short" and "long" lists was that the "long" list contained more surnames thought to be of Mexican origin.

2. There were so many cases on the HSP frame that had a surname of "Martin" (or "Martín") that these cases were placed on a separate frame. A probability sample of these cases was selected. This was the first time that cases placed on one of the minority frames were subjected to being sampled instead of being subjected to a complete canvass.

As in 1972, prior race, ethnicity, and ancestry information on sole proprietors was used if 1) the case was still active, and 2) there was no indication of a change in ownership. Receipts and employment data were obtained from the 1977 Economic Census.

Minority-Owned and Women-Owned Large corporations were identified and handled as they were in the 1972 surveys.

Survey Results[4]

Hispanic-Owned Estimates -- Estimated number of firms was 219,355 (with approximately 107,000 or 48.8% contributed from the NAT frame sample). The 107,000 estimate from the NAT frame was made from approximately 650 Hispanic responses (out of approximately 58,000 mailed). Estimates from the NAT frame were published in a "Not allocated by ..." category for each table. Detail estimates for specific 4-digit SICs included only cases originally on the HSP

4 See footnotes 1 and 2.

frame. Hispanic-Owned firms accounted for 2.2% of all US firms.

Other Minority-Owned Estimates -- Estimated number of firms was 110,837 (no estimates of undercount were attempted from the NAT frame sample). Other Minority-Owned firms accounted for 1.1% of all US firms.

Black-Owned Estimates -- Number of firms was 231,203 (cases were tabbed based on the race codes received from SSA). Black-Owned firms accounted for 2.4% of all US firms.

Women-Owned Estimates -- Number of firms was 701,957 (cases were tabbed based on the sex codes received from the SSA). Women-Owned firms accounted for 7.1% of all US firms.

The 1982 SMOBE and WOB Survey

Frame development

Several major changes occurred in the frame development process for the 1982 SMOBE and WOB.

1. In addition to the file of businesses received from the IRS, the Bureau received a file of people (SSNs) that filed a self-employment (SE) tax form. This information was used to determine which filer on a joint 1040C (sole proprietor) tax return was the owner of the business. In the past surveys, the first filer listed was assumed to be the business owner. However, since in most instances the first filer listed was male, this caused a severe downward bias in the estimates for Women-Owned businesses. A major improvement in the WOB resulted from the additional data received from the SE tax form filers.

2. Based on results of the 1980 Decennial Census, and on results of research on the 1977 SMOBE responses, the surname list was expanded greatly. The surname list used in the 1982 SMOBE included approximately 17,000 surnames. Also, a few Asian and American Indian surnames were added to the surname list.

3. All owners were subjected to the surname match. (In past surveys, owners with a Black race code had not been subjected to the surname match.)

4. IRS rules regarding the number of owners allowed in 1120S corporations increased from 10 to 25; but, because the contract for receiving the data had already been signed, the Bureau received only the first 10 owners listed on the tax return. Also, for 1120S corporations, the Bureau received the number of shares of the company owned by each of the 10 owners. An 1120S corporation was placed on the HSP frame if at least one of the owners had a name on the Hispanic surname list. In determining minority ownership, the sum of the number of shares listed for the 10 owners received was considered 100% of the company shares. If 50% or more of the shares were owned by minority owners, then the firm was considered minority-owned.

5. A "partial match" to the surname list was performed. If the first four characters of an owner's surname matched the first four characters of a surname on the surname list, that firm was placed on a file for further review to determine if the firm should be placed on the HSP, OTH, or NAT frame. (The "long" and "short" surname lists methodology used in the 1977 SMOBE was dropped.)

6. Perhaps the change that had the greatest impact for SMOBE was not so much a change in methodology, but a change in the basic philosophy of what the surname match should accomplish. In past surveys, the surname list was used to pre-identify as many owners as possible that were almost certain to be Hispanic.

   For the 1982 SMOBE, the surname match was used to pre-identify as many owners as possible that "had a fairly good chance" of being Hispanic or Asian. This change meant that the number of firms placed in the HSP and OTH frames increased tremendously as compared to past surveys (from approximately 240,000 (+84,000 "Martin," sampled at 1 in 10) on the HSP and 113,000 on the OTH frames in 1977, to approximately 433,000 on the HSP and 304,000 on the OTH frames in 1982). Since the size of these frames increased so much, a sample of these cases was selected and mailed. The sample design took into account the geographic distribution of the Hispanic, Asian, and American Indian populations to ensure coverage of firms owned by these minorities in all sections of the country (the sample weights for these cases ranged from 1 to 10). States having sparse populations of Hispanic, Asian, or American Indian populations were sampled at a higher rate than states having larger populations of these minorities. A sample of the cases in the NAT frame was selected to make an

estimate of the undercoverage of Hispanic and Asian-Owned firms (the sample weights for these cases ranged from 100 to 200).

7. Cases originally on the HSP frame that were sampled, mailed a survey form, and responded as Asian, American Indian or Other Minority were included in the Asian, American Indian, and Other Minority tabulations. Cases originally on the OTH frame that were sampled, mailed a survey form, and responded as Hispanic were included in the Hispanic tabulations.

Minority-Owned and Women-Owned Large corporations were identified and handled as they were in the 1977 surveys.

As in past surveys, prior race, ethnicity, and ancestry information on sole proprietors was used if 1) the case was still active, and 2) there was no indication of a change in ownership. Receipts and employment data were obtained from the 1982 Economic Census.

Also, possible race codes received from the SSA changed beginning with people filing for SSNs in 1981. The new race categories were 1) Asian, Asian-American, or Pacific Islander, 2) Hispanic, 3) Black, 4) North American Indian or Alaskan Native, and 5) White. Most owners of businesses in 1982 had filed for SSNs before 1981, so very few of the newer, more detailed race codes were received. Most owners were still coded as either White, Black, or Other.

Survey Results[5]

Hispanic-Owned Estimates -- Estimated number of firms was 298,177. This estimate consisted of 248,141 estimated from the HSP and OTH frames, and 50,036 estimated from the NAT frame. The 50,036 NAT frame estimate (16.8% of the total) was not included in the main tables of the publication; it was presented in an appendix table as an estimate of the undercount. Hispanic-Owned firms accounted for 2.8% of all US firms.

Other Minority-Owned Estimates -- Estimated number of firms was 294,493. This estimate consisted of 255,642 estimated from the HSP and OTH frames, and 38,851 estimated from the NAT frame. The 38,851 NAT frame estimate (13.2% of the total) was not included in the main tables of the publication; it was presented in an appendix table as an estimate of the undercount. Other Minority-Owned firms accounted for 2.4% of all US firms.

Black-Owned Estimates -- Number of firms was 399,239 (cases were tabbed based on the race codes received from SSA). Black-Owned firms accounted for 3.3% of all US firms.

Women-Owned Estimates -- Number of firms was 2,884,450 (cases were tabbed based on the sex codes received from the SSA and for sole proprietors, based on the additional ownership information obtained from the Self-employment form match). Of the 2,884,450 firms, 1,794,000 (62.2%) would not have been identified as women-owned had the 1977 methodology of assigning ownership to the first filer on the 1040C tax return been used. Women-Owned firms accounted for 23.9% of all US firms.

The 1987 SMOBE and WOB Survey

Frame Development

There were only minor changes in the methodology for developing the frames for the 1987 SMOBE and WOB when compared to that used for the 1982 surveys.

The number of possible owners of an 1120S corporation increased from 25 to 35, but the Bureau still received only the first 10 owners listed on the 1120S tax form. Also, for 1120S corporations, the Bureau no longer received the number of shares owned by each owner listed. However, research conducted on the 1982 SMOBE responses, and research conducted after the 1987 SMOBE was completed indicated that neither of these developments had a noticeable impact on SMOBE estimates.

The "partial match" of the surname list, which was performed in the 1982 SMOBE, was dropped for several reasons: 1) classification of the cases that were "partial matches" was an intensive clerical operation, and clerical resources were scarce, 2) most of the surnames that were partial matches in the 1982 SMOBE had been assigned an Hispanic, Asian, or Nonminority classification based on 1982 responses, 3) few "new" partially matching surnames would be expected for the 1987 SMOBE, and 4) research indicated that very few additional minority-owned firms were identified by this procedure.

The surname list was expanded to include

5 See footnotes 1 and 2.

nonminority surnames (i.e., non-Hispanic and non-Asian). These surnames were added to the surname list and used in the surname match process, not so much to pre-identify non-Hispanic and non-Asian owners, but to place minority/nonminority codes on each owner for use in an imputation for nonrespondents.

As in past surveys, prior race, ethnicity, and ancestry information on sole proprietors was used if 1) the case was still active, and 2) there was no indication of a change in ownership. Receipts and employment data were obtained from the 1987 Economic Census.

No estimates were made for Minority-Owned and Women-Owned large corporations.

Survey Results[6]

Hispanic-Owned Estimates -- Estimated number of firms was 489,973. This estimate consisted of 422,373 estimated from the HSP and OTH frames, and 67,600 estimated from the NAT frame. The 67,600 NAT frame estimate (13.8% of the total) was not included in the main table of the publication; it was presented in an appendix table as an estimate of the undercount. Hispanic-Owned firms accounted for 3.6% of all US firms.

Other Minority-Owned Estimates -- Estimated number of firms was 439,271. This estimate consisted of 376,711 estimated from the HSP and OTH frames, and 62,560 estimate for the NAT frame. The 62,560 NAT frame estimate (14.5% of the total) was not included in the main tables of the publication; it was presented in an appendix table as an estimate of the undercount. Other Minority-Owned firms accounted for 3.2% of all US firms.

Black-Owned Estimates -- Number of firms was 424,165 (cases were tabbed based on the race codes received from SSA). Black-Owned firms accounted for 3.1% of all US firms.

Women-Owned Estimates -- Number of firms was 4,114,787 (cases were tabbed based on the sex codes received from the SSA and for sole proprietors, based on the additional ownership information obtained from the Self-employment form match). Women-Owned firms accounted for 30.0% of all US firms.

6 See footnotes 1 and 2.

Conclusions

Since the first special study conducted in 1969, many changes and improvements have been made to the SMOBE and WOB frame creation process. From implementing the use of the Self Employment file from the IRS to help identify women business owners, to expanding the names on the surname lists to better identify potential minority owners, these changes have led to better coverage and more accurate estimates.

We are currently in the planning stages for developing the frames for the 1992 SMOBE and WOB surveys. As with all surveys of people or businesses, the population of interest is constantly changing. In planning for future surveys, we must anticipate and recognize these changes so that we continue to provide complete and accurate estimates of all minority-owned and women-owned firms.

Acknowledgments

Many thanks go to Edwin Robison and Donna McCutcheon, not only for their comments and suggestions on this paper, but also for their hard work and dedication to SMOBE and WOB. Thanks also to Ruth Runyan, Mitch Trager, and Margaret Allen for their suggestions.

References

Brawner, Hilda - Follow-up Procedures on Delinquent Reports, 77EC-SS, 3A2, (1977 Economic Censuses Manual) (1978)

Robison, Ed - Estimation Procedure for the Census of Minority-Owned Business Enterprises - Spanish Publication, 77EC-SS, 7H18, Complete Revision, (1977 Economic Censuses Manual) (1980)

Robison, Ed - Survey of Minority-Owned Business Enterprises Overview for the Statistical Methods Branch, (Internal Memorandum, 1989)

Tupek, Alan, and Margarita Perez - Sampling Specifications for the Spanish Mailing for the Census of Minority-Owned Business Enterprises, 77EC-SS, 7H1, (1977 Economic Censuses Manual) (1978)

Tupek, Alan - Mailing a Sample of the S3 Spanish File in the Census of Minority-Owned Business Enterprises, 77EC-SS, 7H4, (1977 Economic Censuses Manual) (1979)

USING PROFESSIONAL ASSOCIATIONS OF INDIVIDUALS FOR ESTABLISHMENT FRAMES

Carl Ramirez, United States General Accounting Office
U.S. GAO, 441 G Street, N.W., Room 3660, Washington, D.C. 20548

KEY WORDS: Sampling frame, mail survey, finance

Introduction

A critical task in any survey is the adoption of an appropriate frame -- a complete listing of the elements comprising the population of interest that will allow the identification and selection of unique elements for further study. Typically, surveys of establishments make use of area maps or plans, business registers, directories, or other complete censuses to create a list of organizations (Federal Committee on Statistical Methodology, 1988; Deming, 1960; Yates, 1949). One alternative to this type of frame is a listing of *individuals*, whose organizational affiliations define the population, and who serve as informants for those establishments. Such a listing could be derived from the membership rolls of a professional association affiliated with the industry in question.

Professional associations are voluntary, usually nonprofit organizations with members from a particular profession or industry. They often serve as forums for issues concerning the profession, centers for collectively sponsored research, agents for political advocacy, monitors of professional standards and ethics, arenas for the conferral of awards and professional recognition, and clearinghouses for news, business contacts, and employment or educational opportunities.

The *Encyclopedia of Associations* (Gale Research, 1993) contains over 22,000 entries, including thousands of trade, business, and commercial organizations, and hundreds of governmental, public administration, military and legal associations. There are many other national associations that cover health, scientific, engineering, public affairs, union, and agricultural professions. A recent series of General Social Surveys of samples of U.S. adults revealed that an average of 14.5 percent of this population claim to be members of professional or academic societies (National Opinion Research Center, 1991). It is clear that the number and scope of such associations is great; any number of establishment populations of interest to the survey researcher are represented by such associations. Association lists also have applications in multiframe designs, in supplementing or correcting frames, and in surveys of the individual members themselves.

This paper describes the use of a frame defined by the membership of such an association in a national survey of U.S. state and local governments and their financial entities. While this case study is only meant to illustrate one application of this method, and cannot be used to draw inferences about the preferability of this method compared to others, the reasons why this approach was taken in this instance may be instructive for future frame choices.

The GAO Survey of State and Local Government Financial Entities

In response to a request by the House Energy and Commerce Committee and the Senate Banking, Housing and Urban Affairs Committee of the U.S. Congress, the U.S. General Accounting Office (GAO) recently surveyed state and local governments, public employee retirement systems, special districts, and other governmental financial entities (which can be considered establishments) on their use of financial derivative products -- complex financial contracts whose value depends on the values of other underlying instruments or assets.[1] In some government jurisdictions and entities, the use of such devices -- which can include futures, forwards, options and swaps -- is prohibited by law; in others, derivatives are widely used to lower the cost of financing, or as mechanisms to hedge the risks inherent in the investments of the assets used to finance their missions. From this survey, GAO hopes to learn more about the nature and extent of the use in the public sector of these controversial instruments.

GAO's task was complicated by a number of factors. First, the ideal survey universe would include not only governments, but also state and local employee pension plans, and even more varied financial entities such as highway authorities, school districts, and public utility commissions. It seemed unlikely that one ready-made list would suffice, and the resources to develop an exhaustive enumeration of all such public sector financial entities were not available.

Second, the complex and technical data GAO planned to request from these entities required an informant with specific knowledge. The most likely candidate was hypothesized to be the official often titled the "Director of Finance." With the great variety of

government structures, however, the functions of this person could be assumed by other personnel -- comptrollers, treasurers, or accountants. A previous survey of state and local financial officers we reviewed confirmed this difficulty (Petersen *et al.*, 1986).

Third, the questionnaire requested some potentially sensitive material: credit ratings, reasons for the use of derivatives, and losses suffered from using them. Therefore, in addition to the level of the informant's knowledge, the authority and responsibility of the chosen informant could determine the quality of the response. These two factors -- functional role and level of authority -- were proposed by Edwards and Cantor (1991) to be the main components of respondent selection that affect establishment survey response. Respondent selection, as shall be seen later, is a crucial element of this and many other establishment surveys.

Choosing a Frame for the Survey

GAO was presented with a number of alternatives in designing its frame. Comprehensive lists of counties, municipalities and other incorporated places could be derived from a number of Census Bureau products, including surveys of state and local government finances, closely related to GAO's survey topic (Bureau of the Census; 1988, 1990, 1991). A number of other enumerations of government entities also exist (see table 1), but none cover all the subpopulations of interest and provide detailed respondent identification or mailing information.

Table 1 also displays some possible alternatives: associations of individuals affiliated with professions relevant to the GAO survey. Hypothetically, a listing of government entities could be derived from individual representatives who are association members. Some of these associations also have as members the finance professionals who would be our preferred informants.

While a multiframe survey could have been designed to incorporate portions of these various lists, GAO ultimately chose to use only the membership lists of the Government Finance Officers Association of the United States and Canada (GFOA) as the basis for its frame. After removing members who were not finance officers, GAO selected one informant from every governmental entity represented in the GFOA, stratifying them as described in table 2.

Table 1: Possible Frames of State and Local Financial Entities

Enumerations of government entities:

- Census products -- reports on *State and Local Governmental Finances, County Government Finances, Directory of Governments* (Name and Address File), *Census of Governments*, Summary Tape Files broken down by incorporated place, etc.
- Directories of governments and agencies -- *Municipal/County Executive Directory, Braddock's Federal-State-Local Government Directory, American City and County Directory of Administrative Services, Directory of City Policy Officials*, etc.
- Financial rating agency reviews of governments -- *Moody's Municipal and Government Manual.*

Professional associations of individuals:

- Government Finance Officers Association of the United States and Canada (GFOA)
- National Association of County Treasurers and Finance Officers (NACTFO)
- National Association of State Accountants, Comptrollers and Treasurers (NASACT)
- State Association of Accountants, Auditors, and Business Administrators (SAAABA)
- Assorted associations of finance officers within selected states

In April, 1993, GAO mailed questionnaires to a single informant at all of these governmental entities, using the same mailing lists that GFOA uses to correspond with its membership. GAO expects to publish a report on derivative products later this year.

Table 2: GFOA Membership Surveyed

SURVEY STRATA	NUMBER SURVEYED
Municipalities (towns, cities, villages)	2,790
Special Authorities and Districts	1,037
Counties	539
Local Retirement Systems	98
State-level Entities (Departments of Finance, etc.)	64
State Retirement Systems	55
State Governments	52
TOTAL ENTITIES SURVEYED	4,635

Advantages of the Professional Association Frame

There were a number of characteristics of the GFOA membership frame that made it particularly useful for the GAO survey of state and local government entities. The reasons for using it touch on several major issues of survey design and implementation.

A Single Frame Most Closely Fitting the Desired Population

The GFOA is a voluntary organization, and does not cover every U.S. state and local governmental entity in its membership rolls. Therefore, the coverage this frame affords will not allow expansion of the survey results to this universe. However, the frame provides a close fit to a uniquely defined *type* of entity. GAO was interested in gathering financial data from various government agencies and other public and quasi-public entities, and GFOA's membership was drawn from across all of the components of this diverse population. The capability to project results to the artificial but meaningful population of governments represented in GFOA is sufficient for the purposes of this study: the objective was to describe patterns and examples of derivative product usage rather than to create point estimates of some universally applicable variable.

GAO's frame design sacrifices the advantages of complete coverage of a large but nebulous population with the expectation of higher data quality (as shall be seen below) obtainable from a well-described and tailored sub-population. This approach enabled GAO to use a single frame for this study, which is parsimonious and easy to report. It also eliminates the need for a multiframe design, which might involve overlapping entries, administrative problems, and other sources of error. (Federal Committee on Statistical Methodology, 1988). In addition, creating this specific universe allowed GAO to conduct a census of member entities, allowing statements to be made about small subgroups (entities recently using derivative products, for example) while simplifying sampling error considerations.

Identification of the Proper Informant

Respondent selection is especially problematic in establishment surveys, and it plays a major role in determining measurement error in these surveys (Dutka and Frankel, 1991; Edwards and Cantor, 1991).

The GFOA frame began as an individual-based listing, so names, titles, and telephone numbers were readily available. This enabled us to more accurately target the informant whose knowledge and level of authority were most appropriate. The alternative, choosing a generic title without a name, was not as attractive.

A very important advantage of the individual-based frame is that it allows personalization of the survey process. Dillman (1978) emphasized in his exposition of the Total Design Method for surveys that personalization is critical. One study reports that initial telephone screening to identify respondents by name for a subsequent mail survey of establishments significantly increased response rates in a number of surveys (Van Liere *et al.*, 1991). The use of a professional association frame may garner similar benefits.

Because several finance officers from one town, county or state agency can choose to join GFOA, we were faced with the task of selecting only one to represent that entity. This selection is akin to removing duplicate elements to create unique entries of the true elements, which are governmental entities. The positive aspect of this was that we could learn more about the structure of a government entity and choose our respondent accordingly. For example, many financial agencies are represented in GFOA by three members: a director of finance, a comptroller, and a treasurer. We learned that our respondent of choice in those cases would be the director. However, our choice might be different when presented with a different lineup, which might include other titles, such as financial analyst or senior accountant. A set of rules for respondent selection was devised after a series of face-to-face pretests of the survey, and consultation with government finance experts.

The name-based system also offers administrative advantages; it provides us with a single contact point for nonresponse follow-up or further surveys. Nevertheless, the survey instrument GAO used was designed to be passed on to a more appropriate respondent within the organization if necessary, and that person would be identified on the questionnaire, allowing us to correct our master lists.

Finally, establishment informants obtained through a professional association may be more interested in the issues facing their industry and more knowledgeable about the information requested by the survey.

List Quality

No sampling plan can succeed unless selected elements are actually reached by the survey. If a large proportion of mail questionnaires are undeliverable, biases can intrude. While entire governments rarely go

out of business -- birth and death rates are relatively low -- or even move to new addresses, smaller government agencies may undergo more changes, and elected or appointed government officers can have a very high rate of turnover. Thus, another advantage of the professional association frame is its relatively high quality -- because the GFOA membership roster is used as a mailing list for several periodicals (the GFOA newsletter, the Government Finance Review, and other frequent notices), the names and addresses of its membership must be accurate. In addition, periodic membership renewal updates and verifies such information.

Because GFOA's membership database was frequently used to produce mailings, GAO was able to procure a file that contained names and addresses already in label format. In fact, other computerized member information may be available, and with the permission of the professional association, this administrative data might be appended to a respondent's survey answers, thus eliminating the need to ask for it in the survey. Programmers maintaining the GFOA membership database could separate members by the type of organization they represent, allowing us to stratify our sample, as in table 2.

Positive Auspices of an Endorsing Association

Early in the study, GAO obtained the cooperation of GFOA in surveying its membership. Besides furnishing GAO with its mailing list, GFOA provided an endorsement letter to be enclosed in the mail-out package. Because respondents will be familiar with and favorable toward the association, the survey may gain added legitimacy, possibly increasing response rates and data quality. Dutka and Frankel (1991) reported salutary effects on candor and cooperation in a survey of car dealers for which the National Association of Automobile Dealers was recruited as a cosponsor.

For the derivatives survey, GFOA also publicized and encouraged participation in the upcoming survey in its periodicals and through announcements at its conferences. GFOA's good offices also extended to soliciting willing members for in-person pretests.

Administrative Assistance from the Association

In addition to the aforementioned assistance, GFOA also provided expert consultation on questionnaire development issues. GFOA maintains a number of special groups, such as the Government Finance Research Center. The staff in this group include economists and accountants who are familiar with the financial practices of their membership. This research center has also conducted surveys of GFOA membership on other subjects and made a number of recommendations for the GAO survey.

Disadvantages of the Professional Association Frame

Incomplete or Inappropriate Coverage of the Population

The most severe restriction on the use of the membership frame is that not all establishments in the universe of interest may be represented by a member. Most professional associations are voluntary, and most require fees and a self-initiated application process. Some associations may even restrict membership to a subclass of the population. This can lead to coverage errors and other biases in results, as will be seen in the next section. Unfortunately, this suggests that many associations form specialty populations; generalizing to all establishments of a certain type may be impossible.

Voluntary associations are usually open to most interested parties. This can result in the presence of foreign elements in the frame. The GFOA membership lists contained librarians, reporters, and other "industry observers" who were interested in developments in government finance but who were not finance officers. These individuals could be easily removed by the GFOA's mailing list database program, however.

As noted previously, individual-based lists must be converted to establishment frames by choosing one element to represent each establishment. This is particularly important if probability sampling of establishments is planned. This selection process can require much work if the judgment process for making these choices is difficult.

Response Bias

There are two possible sources of response bias from professional association frames. First, since members usually self-select themselves into the association, those members picked as establishment survey informants may be markedly different from those who did not choose to join the association, for reasons well known to social scientists. If these differences occur along dimensions measured by the survey instrument, bias can result. Although most establishment surveys request factual data about the organization, and not opinions, an informant may still influence the reporting process in unknown ways.

Second, associations themselves may have political agendas. Even if their communication with informant members during the survey is neutral or nonexistent, the

likelihood that members join to affirm an association's previously asserted political aims may lead to the first source of bias mentioned previously.

Ceding Control to the Cosponsor

No matter how helpful the assistance of the association, the survey researcher should always be cautious when introducing a third party into the researcher/respondent relationship. Collaboration on a survey often means that the cosponsor will expect to introduce material into the questionnaire, obtain data that might be confidential, and determine who is surveyed.

Although it is difficult to strike such a balance, the survey researcher must benefit from the cosponsorship while maintaining an arm's-length relationship. These issues, however, are not unique to professional association involvement -- relationships with clients who are stakeholders has long been a difficult problem for survey researchers.

Discussion

While the ultimate results of the GAO study of state and local government usage of derivative products are not yet known, we believe that using an individual-based frame derived from the membership of a professional association of government finance officers was a rewarding decision. This frame did have drawbacks, to be sure -- undercoverage of the ideal universe, for example -- but we believe the advantages of frame quality, scope, and the various benefits to survey administration outweighed the negatives.

Whether or not this method is applicable to other surveys depends upon the tradeoff between likely increases in frame (coverage) error and possible decreases in nonresponse and measurement error. In particular, the researcher must consider the availability of suitable associations that closely match the universe of establishments in question, that maintain useful membership records, and that are amenable to a cooperative arrangement on the researcher's terms. In any case, association membership lists can be used to test the comprehensiveness of other frames, and have been successfully incorporated into multiframe designs.

A useful step towards evaluating this type of frame would be the more rigorous study of how the factors mentioned above (and any others) enter into the determination of survey error.

Notes

1. For a discussion of financial derivative products, several recently published finance textbooks, such as Brigham & Gapenski's (1985) *Financial Management: Theory and Practice* explain the theory behind them. In addition, comprehensive articles on derivatives and other aspects of risk management have recently appeared in financial and business journals, such as the *Economist* (April 10, 1993).

References

Brigham, E. and L. Gapenski (1985) *Financial Management: Theory and Practice.* Chicago: Dryden.

Deming, W. (1960) *Sample Design in Business Research.* New York: Wiley

Dillman, D. (1978) *Mail and Telephone Surveys: The Total Design Method.* New York: Wiley.

Dutka, S. and L. Frankel (1991) "Measurement Errors in Business Surveys." In P. Biemer *et al.* (eds.) *Measurement Errors in Surveys.* New York: Wiley.

Edwards, W. Sherman and David Cantor (1991) "Toward a Response Model in Establishment Surveys." In P. Biemer *et al.* (eds.) *Measurement Errors in Surveys.* New York: Wiley.

Federal Committee on Statistical Methodology (1988) *Quality in Establishment Surveys (Statistical Working Paper 15).* Washington, DC: U.S. Office of Management and Budget.

Gale Research Co. (1992) *Encyclopedia of Associations, 1993.* Detroit: Gale Research.

National Opinion Research Center (1991) *General Social Surveys, 1972-1991: Cumulative Codebook.* Chicago: NORC.

Petersen, J., P. Watt and P. Zorn (1986) *Organization and Compensation in Local Government Finance.* Chicago: Government Finance Research Center.

U.S. Bureau of the Census (1988) *Local Government Finances in Major County Areas: 1985-1986 (GF-86-6).* Washington, DC: Government Printing Office.

U.S. Bureau of the Census (1990) *State Government Finances in 1989 (GF-89-3).* Washington, DC: Government Printing Office.

U.S. Bureau of the Census (1991) *Directory of Governments, 1988: Name and Address File.* Washington, DC: U.S. Bureau of the Census.

U.S. General Accounting Office (1990) *Early Childhood Education: What Are the Costs of High-Quality Programs? (HRD-90-43BR)* Washington, DC: U.S. GAO.

U.S. General Accounting Office (1992a) *Depository Institutions: Contracting Practices with Data Processing Servicers (GGD-92-19).* Washington, DC: U.S. GAO.

U.S. General Accounting Office (1992b) *Securities Arbitration: How Investors Fare (GGD-92-74).* Washington, DC: U.S. GAO.

Van Liere, K., R. Baumgartner, P. Rathbun and B. Tannenbaum (1991) "Factors Affecting Response Rates in Surveys of Businesses and Organizations." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, 1991, Scottsadale, AZ.

Yates, F. (1949) *Sampling Methods for Censuses and Surveys.* New York: Griffin.

STRATEGIES FOR THE REDESIGN OF A MAJOR BUSINESS SURVEY

Bob Anderson and Richard Vincent, Statistics Canada
Bob Anderson, 7-A Jean Talon Bldg., Ottawa, Ontario K1A OT6

KEY WORDS: Survey design, strategy, planning, administrative data

1. Introduction

The Survey of Employment, Payrolls and Hours (SEPH) is the largest monthly establishment survey in Canada -- about 60,000 establishments are now surveyed. SEPH was founded in 1983, although there were less comprehensive predecessors. SEPH is designed to measure levels and trends of monthly payrolls employment, average weekly and hourly earnings, and average weekly hours at the Canada by detailed industry level and at the province by industry division (e.g., Manitoba, Construction) level.

These estimates are used to help formulate labour income, the most important component of Canada's gross domestic product; to project output estimates, to aid calculation of productivity, to track labour market conditions, and to index contract prices and legislated entitlements/benefits. SEPH's greatest strength is the industrial detail of its employment and earnings estimates. Its employment estimates compare well at the three digit industry level with the estimates produced by the annual surveys of industry. Its earnings estimates, together with derived fixed-weighted earnings estimates, are a measure of wage inflation. SEPH total earnings are used to project labour income estimates and to industrially allocate annual labour income benchmarks.

Considering its short history, SEPH has been relatively successful. The number and importance of its many users have attested to this. The Survey has had some bumpy periods. A switch in frames led to a data break in January 1987. Later that year potential data breaks, caused by frame size code updates, were made transparent to users at a heavy resource cost and at a cost to the accuracy of data levels. Users of detailed industry estimates noted data fluctuations. Sometimes these fluctuations were within targetted co-efficients of variation. At other times a new level lasting a year suggested that rotation (i.e., monthly sample selection) was the culprit. The ultimate cause of these rotation breaks was often inaccurate size coding on the frame. From January 1989 to October 1990 SEPH suffered from undercoverage due to births not being processed on the frame. Because of its many underlying influences, the "average earnings" variable was difficult to explain and the explanations impossible to quantify. Finally aggregate payrolls employment estimates were compared unfavourably with household survey paid worker estimates. The lack of benchmarks diminished SEPH's credibility especially vis-a-vis benchmarked surveys.

The cumulation of these problems led to a perception that SEPH data were not as good as they could be. There was a perception that SEPH processes/systems and methodology were black boxes in which complex, unexplainable computations occurred. Indeed some people felt that SEPH was over designed and trying to do too much.

These perceptions were heightened by conclusions from the program evaluation (Andersen et al., 1991) that SEPH suffered from under reporting, particularly among respondents reporting on computer listings, and, lacked data dimensionality. This program evaluation recommended that SEPH use administrative data sources (particularly as a benchmark source for employment estimates), evaluate continuously respondents' ability to report, and develop a supplementary survey capacity.

Fortunately in the last few years SEPH has been able to incorporate births missing from January 1989 to October 1990, release a continuous time series back to 1983, produce new outputs, introduce a revised questionnaire, improve reporting, and enhance its coverage of large firms.

Nevertheless the problems and perceptions resulted in the need to redesign or re-engineer SEPH. As a first step SEPH managers together with a few senior Statistics Canada managers held a symposium to chart the course for the SEPH of the future.

2. The Future SEPH Sketched Out

At this symposium lessons from SEPH were elaborated. The lessons (Anderson, 1991) can be summarised as follows:

(a) keep methodology, systems and operations

simple;
(b) insulate the survey from frame irregularities and impacts;
(c) maximize the use of administrative data ;
(d) make efficiencies in expensive follow-up and editing processes;
(e) understand and communicate well with the respondent;
(f) consult widely with users and focus on product quality;
(g) collect core information monthly and less time sensitive data periodically.

Throughout this discussion it became evident that SEPH should evolve from a survey to a program of information on employment and compensation. The SEPH infrastructure should be the vehicle for this program. Goals for a future SEPH were elaborated as:
(a) improve data quality;
(b) cut costs;
(c) reduce response burden, particularly on small business;
(d) enhance data dimensionality.

Goals (b) and (c) were imperatives imposed by Statistics Canada's senior management.

These goals, especially the first three, are fairly standard. But what means would be used to achieve these goals? How would the lessons from SEPH be incorporated in the redesign? In retrospect, it is clear that to incorporate the lessons from SEPH a paradigm shift was necessary. The shift can be characterised as a shift from:
. project planning to program planning
. processes to outputs
. data production to information provision
. a survey (SEPH) to an employment and compensation program.

Five strategies or means of marshalling resources were chosen:
(a) focus on client-supplier relations;
(b) develop a supplementary survey capacity;
(c) incorporate administrative data;
(d) use simplified and flexible processing algorithms/systems;
(e) install simplified and more efficient sampling.

Resources were deployed and organisational structures were created to implement these strategies. A User Consultation Committee was established. A Special Surveys Unit was set up. A Benchmarking and Administrative Data Section was charged with incorporating administrative data. Redesign co-ordination committees were organised. Simplicity and flexibility became watchwords for all these groups.

The steps in implementing strategies (a) to (c) are explained in some detail while (d) is briefly summarized in the sections which follow . More detail on (d) can be found in Anderson and Vincent (1993), an expanded version of this paper. Strategy (e) is described in Dolson (1993). The first two strategies will be discussed in Section 3. Section 4 discusses the future use of administrative data in SEPH. In Section 5 a brief overview of a flexible processing algorithm is given. We will start by putting users and respondents first.

3. Responsiveness to Users and Respondents

As was noted in the first Section, an evaluation of Statistics Canada's labour statistics program determined that the data obtained through SEPH lacked dimensionality or depth, and that more comprehensive measures of compensation were needed (Andersen et al., 1991). Program managers conceived of the program as one that separated out the core monthly SEPH survey of conventional employment, payrolls and hours data, from the more complex forms of wage and non-wage compensation and other labour market data. The essential idea is that monthly core data would be collected through the SEPH questionnaire more or less as it currently exists, while more complex information would be collected via supplements to the core survey, on a periodic basis. As well, the program will include the development of a cost-recoverable, special survey capacity. It is the program managers' intention to make the core, supplementary and special survey vehicles and resultant data banks the sources of choice when it comes to establishment-based labour market information in Canada.

Core data requirements were worked out in consultation with a wide range of types of users. Essentially these are similar to present objectives: to produce estimates of the total number of payroll employees, average weekly wages and salaries, average hourly earnings and other variables for each province at the industry division level, and, to produce national estimates for each three digit SIC (detailed Standard Industrial Classification standard) for the same variables. The core data refer to conventional establishment-based labour statistics

and are relatively easy to collect.

The supplementary surveys will add dimensionality to the core information. An example of dimensionality is part-time / full-time employment status, and unionized / non-unionized status. The main limitation to the types of questions which can be asked in surveys supplementary to the core survey is primarily operational. The questions must be related to the contact person's ability to report. Essentially this means the payroll departments serving statistical establishments.

The special surveys differ from the supplementary surveys in that they are not necessarily limited to SEPH's current statistical structures for employment reporting. They can be legal entities or operational entities quite different from the conventional employment reporting unit. They could range from fairly straightforward conventional survey data gathering to complex case studies to collect in-depth labour market information. The major step taken along this line so far has been to conduct a major study into the feasibility of collecting wage and non-wage compensation data and other labour-related information from employers under federal jurisdiction. This includes assessing their capacity to report, and the feasibility of using the current SEPH statistical reporting unit as the "entry point" for the proper contact persons (Beauregard, 1993).

The program evaluation (Andersen et al., 1991) also pointed out that respondents must be continuously educated. As part of meeting this recommendation we have adopted a policy of continuous questionnaire design improvement. The main outcome has been a simplified form for smaller employers, and a clarification of definitions and reporting instructions, particularly in relation to using the terminology of the employer community. Qualitative (e.g., focus groups) and cognitive research have been very useful in reworking SEPH collection materials (Gower and Nangundkar, 1991).

The SEPH program now ensures that long term respondents are informed of any changes to concepts, definitions, etc. This problem arises due to turnover in the employers' designated contact person. In the near future, every contact name change will trigger a follow up to ensure the person has all the relevant materials such as instructions. In addition, visits to respondents are given more emphasis.

More responsiveness to respondents and clients should result in improved data quality. Respondent responsiveness will reduce response error. Clients will provide feedback on the estimates, their relevance, and new information needs. Respondents and clients are key to undertaking special surveys and improving SEPH's dimensionality; respondents to provide data and clients to pay for these surveys.

4. The Incorporation of Administrative Data

In recent years the SEPH program has been researching the use of administrative data, especially for annual benchmarking purposes (Vincent, 1992). For many years senior managers at Statistics Canada have been discussing adding two questions to the Revenue Canada monthly small remitter form. This is the form used by small employers remitting taxes deducted at source; Revenue Canada is the Canadian federal agency entrusted with collecting federal (and some provincial) taxes. In the Spring of 1992 Revenue Canada senior management agreed to add the two questions -- one on employment and the other on payrolls. The implications on SEPH were enormous. Suddenly SEPH would have what amounts to monthly benchmarks for two core variables for small business. SEPH would have to be rethought. The opportunity arose to improve frame currency, coverage, and reporting.

Frame currency refers to how well current economic statistics reflect economic reality when the differences result from frame processing activities. Improved currency comes from the use of an administrative list rather than a statistical frame for the identification of structures. In the case of SEPH, a comparison of units on the administrative list versus units on the statistical frame reveals that four months will be gained by using an administrative list for the small business universe. The administrative list is continuously updated, never "frozen", and not directly used for collection of the two new statistical variables. In contrast, structures on the statistical frame are "frozen" well in advance to allow sample selection, new entrant initiation, corrections, and mail out.

Improved coverage comes from including units not registered on statistical frames. For efficiency reasons, statistical frames only maintain active records; often tiny or sporadic activity records are excluded. Some of these excluded records represent economic activity. Because these excluded records are never eligible for survey selection, SEPH has missed this economic activity. Other records

excluded from statistical frames are births awaiting classification. The SEPH redesign team, remembering the lessons of SEPH, decided that the unclassified records must be included in the small remitter (PD7A) records to be tabulated.

Improved reporting is expected on the PD7A forms because the questions are simple and remitters want to be accurate. The questions were designed to accord with Revenue Canada definitions (especially "gross monthly payrolls") and record keeping practices (especially "number of employees in the last pay period"). The questions were tested in focus group and executive interviews (D.R. Harley Consultants Limited, 1992) and modified accordingly. Most respondents want to report accurately, but the fear of penalty is an added incentive for completers of administrative forms.

Administrative sources have disadvantages. Generally only a few simple questions on administrative forms can be used for statistical purposes. The reference period may not be standardised enough for some statistical purposes. For example, "gross monthly payrolls" is based on the payroll month which may vary from 4 weeks to 6 weeks. These shortcomings obligate statistical agencies to continue some surveying -- though modelling and other techniques can limit the size of the survey sample and thus the response burden.

Another drawback of administrative data is the lack of follow-up. Outlier and imputation techniques, by themselves, must be able to correct for data abnormalities.

Administrative frames do not provide classification, profiling and data integration capabilities. Statistical frames remain essential to statistical programs.

A final disadvantage of administrative data is its exposure to legislative and regulatory change (Jabine and Scheuren, 1987). We must always be vigilant of these changes and their effects on statistical programs.

SEPH's incorporation of small remitter information necessitates the use of a multi-frame approach -- an administrative portion for PD7A (small businesses) and a statistical portion for large and medium establishments. In fact the need for more information from small businesses makes essential the drawing of two populations from the administrative list of PD7A accounts. The first population will be used for administrative estimation of employment and payrolls. The second population (or more precisely, sub-population) will be used for estimation of the full range of SEPH variables (from an overlap survey of 2,500 accounts on a monthly basis). Ratios and co-efficients derived from the overlap survey estimates will be used to mass impute the other variables on the administrative records; then expansion estimation will be used to derive the full range of SEPH variables. Details of SEPH's plans for sampling and estimation in the new multi-frame environment can be found in Dolson (1993).

Administrative data offers room for considerable data improvement. To achieve improved coverage, survey managers must rid themselves of notions of excluding semi-active units from the frame. In fact a paradigm shift which could be characterised as from exclusion to inclusion must be made. It took some SEPH staff a considerable time to understand that in the PD7A portion accounts are being surveyed, not statistical establishments. Other advantages of using administrative data are more evident -- cost and response burden reductions. The use of administrative data will go a long way towards achieving SEPH redesign goals.

5. A Flexible Processing Algorithm

Editing, follow-up and imputation are generally conceived of as steps leading to the preparation of survey records for estimation. They are the "adjustment" steps between what the respondent has (or has not) reported, and the preparation of weighted estimates to represent the universe. The typical approach is to conceive of the respondent's information as containing errors which must be corrected before estimation. This approach usually leads to stringent editing rules, extensive follow-up and significant levels of imputation. Often no error is too small to warrant a follow-up to the respondent. Often a single relationship error between two variables is considered an indication that the entire record should be replaced by an imputed substitute in line with the "average" response. This approach results in these processes consuming substantial portions of a survey's budget, sometimes to the detriment of other survey processes and to the detriment of the exploitation of the data results.

To illustrate the stringent and complex editing rules required to meet the objective of detecting and correcting respondent provided information, some numbers relating to the editing of the current SEPH

survey are revealing. There are as many as 101 active edit rules assessed for each record. The results of these edits require about 3,800 telephone calls per month to resolve, sometimes involving more than one telephone call per reporting unit. The calls are to resolve suspected errors in about 25% of the total reporting units.

There is an approach that allows survey managers more flexibility in the expenditure of survey resources. That approach is to conceive of editing, follow-up, and imputation as processes within the quality assurance paradigm, rather than as processes within the error/correction paradigm of "cleaning" the data for estimation. The preferred approach places the editing, follow-up and imputation steps upon an even level with other survey processes. Editing data thus becomes one of the choices that a survey manager can make in terms of what resources to expend to assure a certain level of quality, or fitness for use. This is not to be misconstrued as "abandoning" editing , follow-up, and imputation. Rather it is to be understood as investing in these processes only to the extent that value is added at an acceptable cost. In practical terms this means prioritizing records in terms of the impact that failure to resolve errors would have on the fitness for use of the estimates. It means automating the typical decisions editors must make along the decision path to decide whether a record probably has a problem or is just exhibiting abnormal but real economic behaviour. It means using as much of a respondent's data as possible, rather than giving in to the temptation to replace a flawed record by an imputed substitute.

Compared to the quantity of editing rules assessed and the number of telephone calls made in the current SEPH, this approach will reduce the number of edit rules from 101 to 37, and should reduce the number of telephone calls per month from 3,800 to 1,000 - 1,500, for the same number of respondents.

This approach is more flexible than the error/correction paradigm because it places editing, follow-up and imputation within the quality control paradigm -- ensuring the fitness for use of the end product -- rather than concentrating on specific, expensive processes.

The redesigned survey processes will use the flexible paradigm of editing, follow-up and imputation as quality assurance processes. Data quality will be at least as good, while costs will be significantly less.

6. Conclusion

The shift in SEPH from processes to products means that there has been a greater appreciation of the need for data quality and dimensionality. One definition of quality is user satisfaction. This definition alone forces providers of statistical data to have closer contact with users of statistical data. One must also continuously monitor the ability to report of respondents. The need for new data to add dimensionality to SEPH necessitates knowing the respondent's capacity to report as well as developing a special/supplementary survey capacity. The need to find sponsors for special surveys is an added reason to keep in contact with users who are also potential clients. Data quality, especially coverage considerations, argues for the use of administrative data. Administrative data cannot meet all statistical program needs, but it can provide simple to collect, key variables. Finally the emphasis on the product means looking at editing as one of the quality assurance activities: errors with little data impact can be ignored while errors with large impacts can be ranked and followed up or imputed.

In general strategies often are not as clearly laid out as is remembered and opportunism may supplant previously defined strategies. Yet in the case of the SEPH redesign the strategies were laid out and documented almost as presented in this paper. It is true that the opportunity was seized to add two statistical questions to the Revenue Canada Taxation small remitter form rather then to benchmark SEPH annually as was earlier envisaged. However the small remitter statistical initiative fits within the strategy of incorporating administrative data and the new data basically constitute monthly employment and payrolls benchmarks for small businesses. Execution of the strategies is important; the strategies were supported by appropriate resources.

Paradigm shifts are difficult to implement. For example, staff tend to apply old ways of doing things to the new paradigm -- people go with what they know. One must educate the staff about why change is necessary, what is different in the new paradigm, and the new model. Without this education, the old formulas will creep into revisions to the programs/systems, operations, and methodology.

The redesign of the Survey of Employment, Payrolls and Hours will result in a fundamentally changed statistical program which could be dubbed the

employment compensation program. This program infrastructure will include administrative data, core survey, and special/supplementary survey components. Data quality and dimensionality will be improved. Flexibility will be built in. Editing will be more intelligent and efficient. Canada's employment compensation program will be ready to enter the twenty-first century.

Bibliography

Andersen, Peter, Morley Gunderson, and Robin Rose (1991). Statistics Canada Labour Statistics Program Evaluation Study. Ottawa: Statistics Canada working document.

Anderson, Bob (1991). Redesigning SEPH: The Employment and Compensation Program. Ottawa: Statistics Canada Labour Division internal working document.

Anderson, Bob and Richard Vincent (1993). Strategies for the Redesign of a Major Business Survey (expanded version). Labour Division internal working document. Ottawa: Statistics Canada.

Beauregard, Jack (1993). Federal Jurisdiction Feasibility Study. Ottawa: Statistics Canada Special Surveys/Labour Division report.

Berthelot, Jean-Marie and Michel Latouche (1990). "Follow-up Strategy for Economic Surveys" in Proceedings of Statistics Canada Symposium 90. Ottawa: Statistics Canada.

Chinnappa, N., R. Collins, J-F Gosselin, et al. (1990). Report on editing. Ottawa: Statistics Canada Methods and Standards Committee.

Dolson, Dave (1993). "On Redesigning Canada's Establishment Based Employment Survey". Ottawa: Statistics Canada. A paper presented at the International Conference on Establishment Statistics, Buffalo, New York, June 27-30, 1993.

D. R. Harley Consultants (1992). Report of the Assessment of a New Employer Remittance Form Conducted on Behalf of Revenue Canada, Taxation and Statistics Canada. Ottawa: Revenue Canada Taxation.

Gower, Allen and Mukud S. Nargundkar (1991). "Cognitive Aspects of Questionnaire Design: Business Surveys versus Household Surveys" in Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

Jabine, Thomas B. and Fritz Scheuren (1987). "Statistical Uses of Administrative Records in the United States: Where are we and where are we going" in Proceedings -- Statistical Uses of Administrative Data, an International Symposium. Ottawa: Statistics Canada.

Vincent, Richard (1992). "The use of Administrative Data to Improve Survey Estimates" in Proceedings of the U.S. Bureau of the Census 1992 Annual Research Conference. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

ON REDESIGNING CANADA'S ESTABLISHMENT BASED EMPLOYMENT SURVEY

David Dolson, Statistics Canada
11-J R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6

KEY WORDS: Administrative data, regression estimation, two-phase sampling

1. The Current Survey

Canada's monthly establishment survey to measure the volume of employment was established in 1918. Its most recent redesign, implemented in 1983, is called the Survey of Employment, Payroll, and Hours (SEPH). It collects data on payroll employment, weekly earnings, and weekly paid hours. The primary objectives currently include:

> to provide monthly estimates of the total number of paid employees, average weekly earnings, average weekly hours and other related variables at the industry division by province level.
>
> to provide these estimates for Canada at the three digit Standard Industrial Classification (SIC) level

The list of establishments SEPH uses as its frame is derived from Statistics Canada's business register (BR). For each monthly survey cycle the frame is updated for births, deaths etc. as reflected on the BR. The primary source of information for maintenance of the BR is the Payroll Deduction (PD) accounts each employer has with Revenue Canada. A group of establishments linked together by ownership or control is called an enterprise. On the BR, each PD account is linked to the enterprise to which it belongs. It is primarily through the births, deaths etc. of these PD accounts that the BR is maintained. A more detailed discussion of the BR is given by Cuthill (1989).

SEPH covers all industries except agriculture, fishing and trapping, private household services, religious organizations, and military services. It is designed as a stratified sample of establishments with stratification by industry division (16), province or territory (12), and employment size group (4). Each stratum is further subdivided into sub-strata by 3 digit SIC called cells. The sampling within each cell is simple random without replacement.

The required precision of the estimate of total employment is specified at the industry division by province level. To achieve this, a sample of about 60,000 establishments is selected from the population of about 800,000. Of these, about 27,000 are self-representing; these are primarily establishments belonging to enterprises having 200 employees or more. The remaining take-some sample is allocated to strata in proportion to the estimated number of employees in the take-some population of each stratum. Within each stratum the sample is further allocated to cells in proportion to the population size in each cell. The take-some sample is rotated at the cell level. Sampled units remain in the sample for at least a year, except for sampled births which generally remain in the sample for fewer occasions. Units which rotate out of the sample are kept out for at least a year.

Dead units detected by independent sources as well as from SEPH are removed from both the sample and the frame in order to simplify their treatment operationally. To maintain a nearly unbiased estimator, an estimate of dead units in the population, called the death adjustment factor, is used when computing weights. Schiopu-Kratina and Srinath (1991) have shown that the resulting estimator performs better, conditionally, than other more traditional estimators of totals.

Estimation of totals and variances is done at the cell level and these estimates are aggregated to the desired level. The basic structure for the SEPH estimator of total is $\hat{N}\,\overline{X}$ where $\hat{N}$ is the estimated number of live units in the population and $\overline{X}$ is the mean of the variable. $\hat{N}$ is not allowed to exceed the actual number of units in the population which may include some unknown deaths. Although this estimator is not unbiased it has a smaller mean squared error than the unbiased estimator in which $\hat{N}$ is unconstrained.

A more extensive description of the current SEPH methodology is given by Schiopu-Kratina and Srinath (1991).

2. Redesign Considerations

A secondary objective of the current SEPH is to produce estimates at the three digit SIC-province level annually. This as well as the second primary objective led to the choice of a detailed stratification using 214 three digit SIC industries in the current design. In fact, SEPH produces three digit SIC-province estimates

monthly. However, detailed estimates such as these are now more clearly viewed as having much less importance than larger aggregates.

The primary objectives of the redesigned survey will include production of good quality estimates of total payroll and total employment each month at the national level by three digit SIC and provincial level by industry division. At a more detailed level emphasis will be placed only on a few "important" three digit SICs in each province. The estimates of total payroll are especially important for estimation of monthly labour income.

Hence there will be a reduced need for detailed industrial strata. "Important" industries may be identified as design strata, while other industries will not. Estimates for these latter industries will likely be less precise than those for the "important" industries.

A result of the detailed stratification in the current design is many cells with small populations (and small employment) but large sampling fractions. The sample rotation methodology developed to cope with this is complex, as is the computer system which implements it; simpler methods are now available and a less detailed stratification will alleviate the problem.

There was a cost to use of the death adjustment factor (daf). Neither the methodology nor its computer systems implementation are simple and it has been problematic in survey maintenance. So, at the cost of a small loss of efficiency, it is now desired to simplify the methodology and systems related to the treatment of deaths by removal of the daf and moving to a more standard treatment of deaths.

Like many surveys of its generation SEPH overedits its data. A large part of the survey budget is spent in this operation. SEPH's editing will be improved and its cost reduced by using newer methodology and systems. In particular, selective editing methods proposed by Hidiroglou and Berthelot (1986) and by Latouche and Berthelot (1992) will be used. These are being done using Statistics Canada's generalized DC2 system for data collection, capture and edit.

Generalized survey processing software was not readily available when SEPH was being designed and SEPH specific systems were developed. Some of these systems have turned out to be excessively rigid and hard to maintain. There are several new developments in computing hardware and software that are being used to develop generalized systems at Statistics Canada which will be helpful in a redesigned SEPH. Computer Assisted Interviewing methods can reduce costs while improving data quality and timeliness; this methodology will be used for some of the data collection in the redesigned survey. The generalized software for survey processing being developed at Statistics Canada that will be used for SEPH are: DC2 as noted above; GEIS, the Generalized Edit and Imputation System; and GES, the Generalized Estimation System.

Finally, and most important, some new administrative data are now available from the payroll deduction data source. Those employers who are to remit payroll deductions monthly to Revenue Canada are now asked to report on the PD7 form which accompanies their payment, the *total payroll for the month* and the *total number of employees* for the last pay period of the month. These monthly remitters are generally smaller employers. Larger employers, who make remittances more often, are not currently required to provide these data on their remittance forms.

Because these new data are available for a large fraction of SEPH's target population a substantial reduction in the SEPH sample size, and hence in the cost of the survey, will be possible. This factor in particular, has provided the impetus to redesign the survey. All necessary redesign activities are scheduled for completion so that the reduced sample size can be implemented for the survey with January 1994 reference month.

3. The Redesigned Survey

Because the new administrative data are available only for the smaller businesses which are required to supply the two new variables on their PD7 forms, SEPH will use two frames - the ESTABLISHMENT frame consisting of a list of establishments, and the ADMIN frame consisting of a list of PD accounts. They are derived from the Business Register and the list of all PD accounts.

Any enterprise which has at least one PD account for which the new data are not required has all of its establishments placed in the ESTABLISHMENT frame. In addition, all establishments belonging to enterprises having more than one establishment or more than 99 employees are also included. All PD accounts for such enterprises, whether the new data are required or not, are therefore excluded from the ADMIN frame. The ESTABLISHMENT frame will include about 100,000 establishments accounting for

about 70% of total employment. A monthly survey similar to the current SEPH will be designed for this frame.

The ADMIN frame includes all remaining PD accounts required to supply the new variables. It will comprise about 800,000 PD accounts, accounting for about 30% of total employment. In the short term (two to three years) it is too costly to capture the new data for all accounts every month. (In the longer term Revenue Canada will automatically capture the data for all of these accounts and provide them to Statistics Canada). Consequently a two-phase sample will be selected. The first phase sample of PD accounts, for which data will be captured from the PD accounts, is called the ADMIN sample. From this sample, a subsample will be selected to collect data for the other SEPH variables not available on the PD7 forms.

3.1 The ESTABLISHMENT Survey

The ESTABLISHMENT frame will be stratified by province (12), industry set and employment size. The industrial stratification will be province specific and oriented towards "important" industries within the province -- generally those with large employment. SEPH subject matter experts initially identified 740 such industry province combinations. Each of these may constitute an industry set for the given province; those with much of their employment or much of their variance coming from the non-self-representing part of the ESTABLISHMENT frame will be retained as industry sets. Remaining three digit SICs will be aggregated to one or more industry sets defined at higher levels of industrial classification so as to balance the need for adequate homogeneity in these strata with the need to constrain the number of them to a reasonable level. At the time of writing, the number of industry sets per province ranges from a low of 13 to a high of 54 for a total of 360. (This compares to 1863 in the current design.)

There will continue to be four levels of size stratification, uniform for all provinces and industries. All establishments, regardless of size, belonging to enterprises having 300 employees or more will be self-representing. This boundary of 300 employees is a compromise between a number of factors. First, for the purpose of allocation of estimates of labour income to industry and province, data are required from SEPH for complex structured enterprises operating in more than one three digit SIC or province. This boundary will include with certainty enterprises accounting for the large majority of earnings of complex structured enterprises. Secondly, it is also a compromise between the needs of generally smaller industries and provinces where a lower take-all boundary would be more optimal and those of bigger industries and provinces where a higher boundary would be better. The total sample of 31,000 for this frame will consist of about 21,300 self-representing establishments plus about 9,700 establishments selected from the non-self-representing population of about 78,700.

A more efficient approach would implement a design with industry-province specific employment size stratification, including the take-all boundary. However, to meet the January 1994 implementation date we are constrained to simple modifications of our existing system which requires that the same employment size stratification be used in all provinces and industries. A province specific approach may be implemented at a later date.

Sample allocation will be determined via an approach which initially specifies a target coefficient of variation for estimated total employment for Canada. Then, this will be translated to a CV target for estimated total employment for each province; these targets will vary to a limited extent between provinces. Within each province a CV target is then derived for each industry set. Finally this translates into a CV target for the ESTABLISHMENT portion by adjusting for the CV of estimated total employment for the ADMIN portion. This approach is described in a more generalized context by Latouche (1988).

Sample selection and rotation will continue as it does currently with one exception. It will be simplified by removal of the death adjustment factor. Instead, dead units detected by the survey will be retained in the sample until they would normally rotate out. In the longer term SEPH plans to move to a newer and simpler sample rotation method like the modified collocated sampling strategy described by Srinath and Carpenter (1993).

In general, estimation of totals will continue to use the expansion type estimator currently used by SEPH. However, when estimates are needed for industries not separately identified as strata, post-stratification will be used. The use of a sample size dependent estimator is also being considered for small domains. Several of these are described by Srinath and Hidiroglou (1985).

3.2 The ADMIN Survey

The ADMIN sample (10% of the frame in the three largest provinces, 100% in the two territories, 20% elsewhere) is manually selected each month and is a systematic sample of PD account numbers. This sample has been in place since January 1993. Although deaths are deleted and births added, no sample rotation takes place. From this sample, data are available for total employment and total payroll (these being the two new variables added to the PD accounts) but not for the full range of SEPH variables. Starting in January 1994, a subsample of 7,500 will be selected from those accounts on the frame which are potentially alive and classified for both industry and province to collect data for these other variables.

3.2.1 The ADMIN Sample

A first step in the processing of the ADMIN sample is its treatment for missing data. In any given month it is expected that no PD7 form will be received for about 30% of accounts. A large fraction of this is accounts for which there are no employees in the month due to temporary or seasonal closure; such employers are requested not to send in their PD7 forms. For very many of these units, it is known a priori that no remittance is expected and codes are maintained indicating this; imputation of zero employment and payroll for such units is easy. Employers who do have employees but for whom the PD7 form is not received in time and those who send in their PD7 forms but fail to indicate either or both of total employees and total payroll will be considered as non-respondents. Finally many deaths may (initially) be indistinguishable from non-response by a live unit. For these latter two groups, deterministic imputation is done when information is available for the same units from the previous month and using averages and trends for imputation groups (generally two digit SIC by province group combination). When such information is not available, a weighting adjustment is made.

3.2.2 The ADMIN Subsample

From the ADMIN sample, data will be available for total employment and total payroll but not for the full range of SEPH variables. To collect data for these other variables a subsample of 2,500 will be contacted each month. They will be selected from those accounts on the frame potentially alive and classified for both industry and province. However, this very small monthly sample which our budget and response burden considerations allows us is not considered to be adequate and it is planned to "borrow strength" temporally to improve the estimates. Although more sophisticated methods are available, it is planned to adopt a relatively simple one, as follows.

A subsample of 7,500 PD accounts will be selected. Rotation as well as updates for births and deaths will take place every month with each sampled unit being kept in the sample for at least one year followed by at least one year out of sample. It will be split up into three portions of 2,500, each representative by industry and province. One portion will be surveyed each month and each portion will be resurveyed quarterly. At the estimation stage each month, data for the full sample of 7,500 will be used by combining the sample for three consecutive months, centred at the month in question.

Like the ESTABLISHMENT sample, the ADMIN subsample will be stratified by province, industry and employment size group. Again, the industrial stratification will be oriented towards "important" industries. Because of the very small sample size, the stratification may have to be at a more aggregated level than that for the ESTABLISHMENT frame.

Where possible, the ADMIN subsample will be stratified by employment size group. This stratification will likely have at most two levels - 0-19 employees and 20 or more. Only one level will be used in situations where the population or expected sample size is too small. The small units covered by the ADMIN frame have very dynamic employment levels. Thus more levels of employment size stratification will likely be avoided in order to minimize difficulties with stratum jumpers.

The purpose of the ADMIN subsample is for estimation of total hours and the allocation of hours, earnings and employment to categories of employee (paid by the hour, salaried, other). Sample allocation will be oriented towards maximizing the efficiency of estimates of total hours.

3.2.3 Estimation for the ADMIN Frame

For total employment and total payroll for the month, estimation can proceed directly, using the sample of PD accounts. For all other variables, a model assisted approach will use information from both the sample and the subsample. For an excellent discussion of model assisted methods, see Särndal, Swensson and Wretman (1992).

Model groups consisting of sets of strata from the subsample will be defined. Normally a model group will consist of a number of industry sets within a province. In some cases where subsample sizes will are too small a model group may cover more than one province for its industry set(s). Regression estimation will be done at the level of the model group using total employees and total payroll for the month as the independent variables. For each model group, estimates for these two variables are controlled to be equal to the direct estimates from the ADMIN sample.

In the near term, a specific model assisted method described in section 7.12 of Särndal, Swensson and Wretman (1992) will be used; observed values are used for units in the subsample and predicted values for remaining units. This estimator, called a cosmetic estimator, is also discussed by Särndal and Wright (1984). It will be implemented via mass regression estimation. Within the context of a broader discussion of imputation, this procedure is discussed by Kovar and Whitridge (1993). Regression parameters will be estimated for each model group using the subsample data. Values for all of the other SEPH variables will be imputed for each PD account in the ADMIN sample but not in the subsample, model group by model group, using the appropriate estimated regression parameters. Although this procedure is unbiased for model groups, it is potentially biased for domains below the model group level if the model fails. This procedure also has the property that estimates of the other SEPH variables for small domains which are not represented in the subsample will be synthetic. In order to minimize the risk or frequency of negative imputed values that may occasionally arise, model groups will have to be sufficiently large as to ensure an adequate sample size while not so large as to be non-homogeneous with respect to the assisting model. Variance estimates will be available for total employees and for total payroll for the month, but not for the other SEPH variables due the use of mass imputation.

In the longer term, it is hoped to implement estimation via a modified version of the generalized regression estimator using the Generalized Estimation System (GES) being developed at Statistics Canada. Model groups will define the level at which the regression is carried out. The ADMIN data from the sample will be linked to specific model sub-groups and computation of g-weights will account for these data at this level.

The frames for a given reference month m, are first constructed in m-1 and are based upon information as of the end of m-4. The ADMIN sample, which is selected and captured in m+1, will include not only accounts on the frame but also new accounts from m-3, m-2, m-1, and m. The frames will be updated to include these units whether sampled or not. Those new accounts belonging to enterprises covered by the ESTABLISHMENT frame will be dropped while the remaining ones will be added to the ADMIN frame.

For reference month m, preliminary estimates are published in m+2 with revised estimates in m+3. At this time the ADMIN frame consists of three sets of accounts -- those which were eligible for selection to the subsample, newly classified units (both new and old), and unclassified units (both new and old).

For the unclassified units, all that can be done is to estimate their total employment and total payroll. The other SEPH variables cannot be estimated since these units are not represented in the subsample in any way.

The newly classified units were not eligible for inclusion in the subsample. However, they will be included in their appropriate industrial strata for estimation purposes as if they had been eligible. This is not a problem for estimation of total employment and total payroll where the data come from the sample. However, for estimation of the other SEPH variables it assumes that the relations between variables are not different from those for units which were eligible for selection into the subsample. SEPH subject matter experts believe this to be a reasonable assumption. Further, it is believed that even if not true, the bias will be small and acceptable since it would affect only a small fraction of the population and only in the distribution of estimated total employment, total payroll and total hours to various categories.

Estimation for reference month m will be carried out using data collected for reference months m-1, m, and m+1. From a collection point of view, although data collection will be slower for the larger units from the ESTABLISHMENT frame - for whom collection is primarily by mail - it is expected that the CATI collection for the ADMIN subsample will provide m+1 data early enough to be usable for estimation for reference month m. From the estimation point of view, this procedure assumes temporal stability - over three months - of the assisting model. In a few highly seasonal industries this is believed to be a poor assumption. In these cases there is a trade-off between variance on one hand and model bias on the other.

Bias can be reduced by a procedure in which reduced "weight" is given to the data from m-1 and m+1 at the cost of increased variance due to a reduced effective sample size. It is important to note that this does not affect estimation of the primary variables, total employment and total payroll, and is applied only in the ADMIN frame, affecting on average estimates covering about 30% of employment.

A final stage of estimation is combining estimated totals from the ESTABLISHMENT portion and the ADMIN portion to produce estimated totals for the entire target population. At this point ratios such as average weekly earnings, average hourly earnings etc. can be computed.

4. Concluding Remarks

The new data available from the administrative source allows for a significant improvement in the estimates for total employment and for total payroll while reducing the respondent burden amongst small businesses. Because the frame can be updated to include the most recent births, SEPH estimates will reflect a more current population than the current survey. The survey design will be more efficiently oriented to industries which are most important. Although a thorough discussion is out of scope for this paper, SEPH is making major improvements to its data collection, capture, edit and imputation procedures which will reduce the survey's costs and help improve its data quality. In the medium term SEPH will also be simplifying its sample selection and rotation procedures, possibly using Statistics Canada's Generalized Sampling System, GSAM that is under development. As well, some aspects of estimation are being implemented using the Generalized Estimation System, GES. The new survey will be less costly, more efficient, more flexible, easier to maintain and produce improved data quality.

This paper describes SEPH redesign plans as of July 1993, but since the redesign is still under way these plans remain subject to change.

ACKNOWLEDGEMENTS

I would like to thank B. Armstrong, B.N. Chinnappa, D. Binder, M. Latouche, P. Lys, and K.P. Srinath for their helpful comments.

REFERENCES

Hidiroglou, M.A. and Berthelot, J.-M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys," *Survey Methodology,* 12, 73-83.

Cuthill, I.M. (1989), "The Statistics Canada Business Register," *Proceedings of the Fifth Annual Research Conference,* U.S. Bureau of the Census, Washington, DC, 69-86.

Kovar, J.G. and Whitridge, P.J. (1994), "Imputation of Establishment Survey Data," in B.G. Cox et al (eds.) *Survey Methods for Businesses, Farms and Institutions,* New York: Wiley.

Latouche, M. (1988), "Sample size determination, allocation and selection," *Methodology Branch Working Paper BSMD-88-021E/F,* Ottawa: Statistics Canada.

Latouche, M. and Berthelot, J.-M. (1992), "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys," *Journal of Official Statistics,* 8, 389-400.

Särndal, C., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling,* New York: Springer-Verlag.

Särndal, C. and Wright, R. (1984), "Cosmetic form of estimators in survey sampling," *Scandinavian Journal of Statistics*, 11, 146-156.

Schiopu-Kratina, I. and Srinath, K.P. (1991), "Sample rotation and estimation in the Survey of Employment, Payroll and Hours," *Survey Methodology*, 17, 79-90.

Srinath, K.P. and Carpenter, R. (1994), "Sampling methods for repeated business surveys," in B.G. Cox et al (eds.) *Survey Methods for Businesses, Farms and Institutions,* New York: Wiley.

Srinath, K.P. and Hidiroglou, M.A. (1985), "Sample Size-Dependent Estimators for Small Areas With Applications to Business Data," in R. Platek and M.P. Singh (eds.) *Small Area Statistics: An International Symposium,* 118-134, Ottawa: Carleton University Press.

USING THE CLUSTER ANALYSIS FOR STRATIFICATION PURPOSES AN EXAMPLE FROM AGRICULTURE

Hans-Theo Speth, Federal Statistical Office
65180 Wiesbaden, Germany

KEY WORDS: Stratification, Cluster analysis.

In the following, it is shown that the cluster analysis is a method for finding meaningful strata in stratified sampling with several survey variables.

Stratifying the population to be sampled is a method often applied for random sampling to reduce the risk of selecting an unfavourable sample. Before sampling, the population is divided into nonoverlapping subpopulations. These subpopulation are called strata. Then a simple random sample is drawn from each of these strata.

Determining the strata is a crucial factor for increasing the precision as compared with unstratified simple random sampling. The only case to be examined here is using the stratification exclusively to increase the precision of estimates of characteristics of the whole population without aiming at estimates of parts of the population. For a single survey variable, Dalenius set up equations in 1957 which permit to optimize stratification. This means that - with a pre-set number of strata, a pre-set sample size and a pre-set method of allocating the sample to the strata - these equations allow to find the strata that minimize the standard error of the estimate. Unfortunately, there is no corresponding method in situations where there are several survey variables.

The decisive factor is that the estimate of the population total or the population mean of a variable from a stratified sample is made up of the respective estimates for the individual strata. This means that also the error variance of the estimate is made up of the error variances of the estimates of the strata. Thus it is meaningful to define the strata in such a way as to make them as homogeneous as possible, i.e. there should be as little difference as possible between the variable values of the sampling units within the individual strata. As a consequence, the estimates in the strata are comparatively exact.

This task, however, is exactly the task of the cluster analysis. Generally speaking, it is to allocate items - that are characterized by values of a multitude of variables - to classes whose elements are as similar as possible with respect to a specific criterion. Thus it would be an obvious choice to perform the stratification of sampling units - for samples with several survey variables - by means of a suitable cluster analysis method.

If $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{Y})^2$ is the variance of a variable Y in a population of the size N, with y_i indicating the individual attribute of a variable and $\overline{Y}$ indicating the mean of the variable, then there is for the stratified population with k strata:

$$(N-1)S^2 = \sum_{h=1}^{k} \sum_{i=1}^{N_h} (y_{hi} - \overline{Y}_h)^2 + \sum_{h=1}^{k} N_h (\overline{Y}_h - \overline{Y})^2$$

It can now be shown that - for stratified samples - it is mainly the first term on the right-hand side of the equals sign which is responsible for the error variance of an estimated value. Since the left-hand side is a constant, it is thus possible and meaningful with regard to the stratification to determine the strata C_h, h=1,...,k, in a way as to get a minimum value for the following function:

$$Z_l(C_1,...,C_k) = \sum_{h=i}^{k} \sum_{i \varepsilon C_h} (y_{hi} - \overline{Y}_h)^2$$

If Y is not an individual variable but a variable vector, the following equation applies:

$$\sum_{i=1}^{N} \|y_i - \overline{Y}\|^2 = \sum_{h=1}^{k} \sum_{i \varepsilon C_h} \|y_i - \overline{Y}_h\|^2 + \sum_{h=1}^{k} N_h \|\overline{Y}_h - \overline{Y}\|^2$$

$\|x\| = \sqrt{x'x}$ indicates the L_2- norm of the vector x.

The above situation then suggests in this case to minimize the target function

$$Z_2(C_1,\ldots,C_k) = \sum_{h=1}^{k} \sum_{i \varepsilon C_h} \left\| y_{hi} - \overline{Y}_h \right\|^2$$

The aim thus is to find the strata or clusters $C_1,\ldots,C_k$ in such a way as to minimize the sum of the sums of the squared Euclidean distances between the cluster members (or stratum members) and the relevant centre of the cluster.

Since minimizing the above target function is equivalent to maximizing the function

$$Z_3(C_1,\ldots,C_k) = \sum_{h=1}^{k} N_h \left\| \overline{Y}_h - \overline{Y} \right\|^2$$

this criterion provides strata with great homogeneity with regard to the Euclidean distance within the strata and great heterogeneity between the strata.

An example of the successful application of cluster analysis for stratification purposes is the stratification pattern of the representative vegetable cultivation survey for the federal Land of Rhineland-Palatinate in Germany. This is a cluster sample survey of agricultural holdings cultivating vegetables. The sampling units are communities. Within the selected communities, all survey units are questioned on principle.

The most important survey variable is the total area of vegetable cultivation. However, the areas of individual positions are covered, too. As the correlation between the area of an individual position and the total area of vegetable cultivation is not very high at the community level, stratification with regard to the vegetable cultivation area for the individual positions would not lead to a substantial increase in precision of the results as compared with simple random sampling. Since, on the other hand, for the major individual positions, the correlation between the areas under cultivation at different times is high at the community level, it was on principle considered to take these individual positions themselves as stratification variables, i.e. it could be attempted to find strata by means of a cluster analysis which would reflect something like the cultivation structures of several types of vegetables.

The cluster algorithm applied was the KMEANS algorithm. This is a (partitioning) method where it is required to preset both the number of clusters to be delimited and an initial partition of the population. KMEANS then minimizes as far as possible the above target function by means of a purposive successive exchange of units.

The variables used for the cluster analysis were the eight types of vegetables that are the most important in terms of cultivation area and the variable "area of the other types of vegetables together".

The cluster analysis resulted in some few clusters of considerable size and numerous clusters consisting of just one or two communities. This means that, in the sense of the cluster analysis, many communities had to be considered as extreme with regard to the cultivation structure. Since these communities were also the biggest in terms of area under cultivation, they were grouped together in the so-called total stratum, i.e. a stratum with sampling fraction 1. In addition to the clusters of the total stratum, there were 5 other clusters. Three of these clusters could be identified as clusters of communities with an intensive cultivation of asparagus. This was plausible insofar as in the observed region asparagus must be regarded as an isolated special culture as far as the soil conditions are concerned. The other two clusters differed only by the orders of magnitude of the vegetable cultivation areas. Thus, the conclusion to be drawn from the cluster analysis with regard to the stratification of the communities was that not only strata of communities with respect to the vegetable cultivation area should be set up, but also strata with regard to the special culture of asparagus.

The old sample design provided for four strata. Here, too, there was a total stratum of similar size. The only stratification variable was the total area of vegetable cultivation. The construction of strata was performed according to the optimum principle of Dalenius.

In the following table, the relative standard errors of the most important types of vegetables and of the total area of vegetable cultivation are indicated for the old and the new sample design with a sample size of 96 out of 256 communities. In both cases, the allocation of the sample to the strata not totally covered was performed according to the optimum allocation (allocation according to Neyman-Tschuprow) with regard to the total area of vegetable cultivation.

Cultivation area of	Relative standard error in %	
	old stratification pattern	new stratification pattern
Vegetables, total	0.5	0.4
Carrots	1.6	1.2
Lettuce	1.3	1.0
Cauliflower	0.9	0.8
Onions	1.6	0.8
Red radishes	0.1	0.1
Spinach	0.6	0.1
Asparagus	8.0	3.9
Large radishes	1.1	0.9

It is not surprising that by applying the new stratification pattern a considerable higher precision of the result could be achieved especially for asparagus.

But also for estimating the cultivation areas of the other vegetables examined, the stratification on the basis of the cluster analysis results led to an increase in precision. Thus the example illustrates the general advantage that this method can offer for surveys with several survey variables.

References:

Späth,H.: Cluster-Analyse-Algorithmen, R.Oldenbourg Verlag, München, 1977.

Dalenius, T.: Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice, Uppsala, 1957.