AN ASSESSMENT OF THE USE OF ADMINISTRATIVE DATA IN THE IDENTIFICATION OF NEW FARMS

Johanne Denis and Jeannine Morabito, Statistics Canada
Johanne Denis, R.H. Coats Building, 3rd Floor, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

Key Words: Dual Frame, Frozen Frame, Dynamic Frame, Administrative Data

1. Introduction

The major Canadian agricultural surveys are traditionally redesigned following each quinquennial Census of Agriculture. The Census of Agriculture is enumerative and covers all farms in Canada that produce agricultural products intended for sale. It provides information on a multitude of agricultural activities. It is therefore a good opportunity to re-evaluate the Agricultural Survey Program in light of the current environment and of the anticipated future trends in the agricultural sector.

After various studies and discussions with subject matter experts, the agricultural surveys are being redesigned in 1997 based on the 1996 Census of Agriculture list of farms. The working assumptions are that we will use: (i) separate designs for crop and livestock surveys; (ii) a frozen Census list frame for most of the sample designs; and that (iii) new entrants in agriculture and farms missed by the Census (Census undercoverage) will be identified through administrative data, as opposed to the historical use of an area frame.

Going from a dual frame approach using a list and an area component to a dual frame approach using two lists only is a major change for the Agricultural Survey Program. The aim of this paper is therefore to outline the proposed methodology. In section 2, we present some background information about the dual frame approach that has been used in agricultural surveys since 1979, and why this approach is now being relinquished. In section 3, an alternative to the area frame is proposed. The methodology that is being considered to identify new entrants in agriculture as well as to account for Census undercoverage is discussed. In section 4, a description of the pilot study that was conducted in September and October 1997 is provided and some preliminary results are given.

2. Agricultural Survey Designs Over Time

Prior to 1979, the Agriculture Enumerative Survey (AES) was conducted. This multi-purpose survey was based almost entirely on an area sample producing estimates for crops, livestock and expense items for all provinces. The AES provided complete coverage of the population. However, in order to improve the efficiency of the sample design, a list frame comprised of a group of large farms (large with respect to some key items) was introduced in 1979 for three Atlantic provinces as a test of the use of dual frames. The Hartley multiple frame estimator, which includes a dual frame estimate for the overlap portion, was then used to combine list and area estimates.

The Census of Agriculture list of farms was first used as a sampling frame following the 1981 Census of Agriculture (Ingram and Davidson (1983)). Given the success of the dual frame approach in three Atlantic provinces, it was decided to continue with this approach. The multi-purpose National Farm Survey (NFS) made use of an area frame covering the whole country in combination with the frozen Census list frame. The area frame component was to target new farm entrants and Census undercoverage. A two-stage stratified sample design was used to select a sample of farm segments from this frame. A multiple frame screening estimator, which is the sum of the list frame and the non-overlap area frame estimates was then used. The Hartley multiple frame estimator was not considered for the NFS. This decision was based on a 1978 AES test which showed that it provided little gain in efficiency for most items (Armstrong (1979)). As well, it would have been more expensive to implement due to the necessity of completing a full questionnaire for all overlap area sample farms and higher respondent burden.

Following the 1991 Census of Agriculture, the area frame survey was redesigned and was called the Area Farm Survey (AFS). Technological advances allowed the move from a two-stage to a one-stage area design. This one-stage design allowed better control over the inclusion probabilities (smaller weights), thus increasing the efficiency of the sample.

Although the dual frame approach using an area component is theoretically valid, several limitations affect the quality of the estimates. These are listed below:

i) A farm cannot be important for major field crops without having an important cultivated area. These farms are usually well identified by the Census of

Agriculture. If some changes happen in the intercensal period with regard to farmer / owner of the farm, the list sample is more likely to catch these changes than the AFS.

ii) The screening multiple frame estimator will provide a gain in efficiency as long as one can identify enough non-overlapping farms. New farms constitute a rare population and are more likely to be small in terms of land area, thus difficult to find. On the other hand, they might be large in terms of sales. This is especially true for some livestock farms specializing in hogs, poultry or feedlots, as well as for sod and nursery farms and fruit and vegetable farms. In addition to being small in terms of area, these farms are scattered over the country, thus not clustered in the same area segments.

iii) The probability of being missed by the AFS, that is, not listed within a segment, is higher for small farms than for large farms, in terms of land area. Since small farms are more likely to be non-overlapping farms, this is a serious limitation to the area approach.

iv) The overlap detection is not an easy task. If a farm is erroneously identified as non-overlapping with the list frame, for example, the resulting screening estimator overestimates the variables of interest.

Most of these problems could potentially be solved by decreasing the segment size (for non-response adjustment problem) and increasing the sample size of segments. Moreover, improvement in the overlap detection procedures could be studied and implemented. However, the increased availability of administrative files combined with advances in record linkage, as well as budgetary constraints, are now driving the Agricultural Survey Program to move from a dual frame approach, using a list and area component, to a dual frame approach using lists only.

3. Description of the Proposed Methodology

As already mentioned, an important limitation of the AFS is the elusive characteristic of the sub-population being targeted, namely new farm entrants (births) and Census undercoverage. The aim of the proposed methodology is thus to specifically target the sub-population of interest, before sampling.

The redesign of Canadian agricultural surveys will initially take place in the fall of 1997 for the Crops Survey and in January 1998 for the Livestock Survey. Other agricultural surveys will be redesigned throughout the year of 1998. We then propose to conduct a Farm Update Survey (FUS), once a year. Each year, a sample of potential births is to be selected from administrative sources. These sampled farms are to be contacted each April, starting in April 1998, to establish if they began to operate after 1996 Census Day (May 14). As well, a sample of potentially undercovered farms is to be selected once, in 1998, and then contacted each year following 1998.

Two major administrative sources are currently under study for the purpose of identifying potential births and potential undercoverage. These two sources are described below in terms of content/coverage, concepts/definitions and frequency/timeliness of the data.

3.1 Administrative Sources

A. Personal Income Tax File (T1 file) of *farmers* reporting for unincorporated *farms*:

Content/Coverage:
Individuals (farmers) reporting a gross farm income greater than zero or a net farm income other than zero are included on this file. Each record contains information concerning the farmer's name and address, gross farm income and net farm income.

Concepts/Definitions:
Records represent farmers. Multiple links between farmers and farms are therefore possible (for example, husband and wife linked to the same farm, partners of the same farm, farmers operating more than one farm).

Frequency/Timeliness of Data:
One file is produced per calendar year. A preliminary version of the file with good completeness is available in November of year y+1, covering the previous calendar year (year y). The final file is available in February of year y+2. In the worst scenario, a time lag of approximately two years is to be expected to identify a new entrant in agriculture. For example, a new entrant in January of year y would appear in the file available in February of year y+2.

B. Income Tax File of incorporated farms (T2 file):

Content/Coverage:
Incorporated businesses reporting agricultural specialization representing 51% or more of total activity are included on this file. Therefore,

incorporated businesses for which agriculture is not their main activity cannot be identified. Each record contains information concerning the corporation name, sales and SIC code. No contact name is available on this file.

Concepts/Definitions:
Records represent incorporated businesses with 51% or more of total activity coming from agriculture. Unlike the T1 file, multiple links between farmers and farms do not exist.

Frequency/Timeliness of Data:
Again, one file is produced per calendar year with the same delays as for the T1 file.

3.2 Identification of Potential Births Using Tax Files

Once a year, it is proposed to identify potential births by matching two subsequent years of T1 and T2 files. For example, in the spring of 1998, the 1996 and 1995 T1 files will be matched by T1 numbers; a similar approach will be used for the T2 files. Records on 1996 files not found on 1995 files will become potential births for 1996.

Given cost constraints, an important issue arises with regards to a gross farm income lower threshold for T1 and a sales lower threshold for T2. A preliminary study showed that out of a population of 40,000 potential T1 births (farmer level), 70% of the records had below $10,000 of gross farm income. This however makes sense since a new farmer is more likely to generate less income in the first year of operation. Since the census list portion of the sample designs will use a $1,000 threshold, it is suggested to use the same threshold for the administrative list portion. Records under that threshold will however be followed-up on subsequent years so as to include them if their gross farm incomes become greater than $1,000.

Since potential T1 births are at the farmer level, different file manipulations are subsequently performed in order to establish, as much as possible, links between farmers that might operate the same farm. The knowledge of links between two or more records is therefore used before data collection, so as to contact only one of the respondents corresponding to the records that are linked together. Matches are performed between T1 potential births and the 1996 T1 universe file by name, city and gross farm income to identify partnerships. As well, matches are performed with the 1996 Census list of farms to remove any potential births that overlap with the 1996 Census of Agriculture.

3.3 Identification of Potential Census Undercoverage Using Tax Files

Once a year, the universe of T1 and T2 records involved in agriculture, as defined in section 3.1, is matched with the Farm Register (FR), which is our repository of farms. Census farms are included on the FR as well as other farms. These other farms might be, for example, farms that were enumerated during the 1991 Census of Agriculture and for which no information was obtained about their business status. This match between tax files and the FR is performed to establish links that could further be used to validate tax data using survey data for the purpose of the Tax Data Program. An approximate 65% match rate of tax records is obtained from that exercise. Unmatched tax records are not further examined. These unmatched records could be due to Census undercoverage, links that couldn't be established because of bad information ("bad links") or births. The sub-universe of births is therefore included in the unmatched tax records. Diagram 1 presents the various sub-universes of interest.

In order to account for Census undercoverage, it is proposed to add, in addition to a birth sample (①), a sample of unmatched tax records to the FR (② excluding ①) (from the 1996 tax record match with the FR) and of matched records with the FR not covered by the 1996 Census of Agriculture (③). This sample of farms potentially missed by the 1996 Census is to be contacted in the spring of 1998 to establish if they were really missed. Once classified as having been missed, they will further be included in various commodity and financial surveys throughout the intercensal period.

3.4 Sample Design of the Administrative Portion

Redesign working assumptions for livestock and crop surveys involve two separate designs with a static frame approach. The 1996 Census list of farms is therefore independently stratified for both surveys based on Census data. These stratifications remain static during the intercensal period. Deaths identified through surveys are dealt with by using domain estimation. However, if one doesn't take into account the dynamic nature of agriculture, list estimates deteriorate during the intercensal period.

A sample of potential births from tax files and of farms potentially missed by the 1996 Census is therefore to be

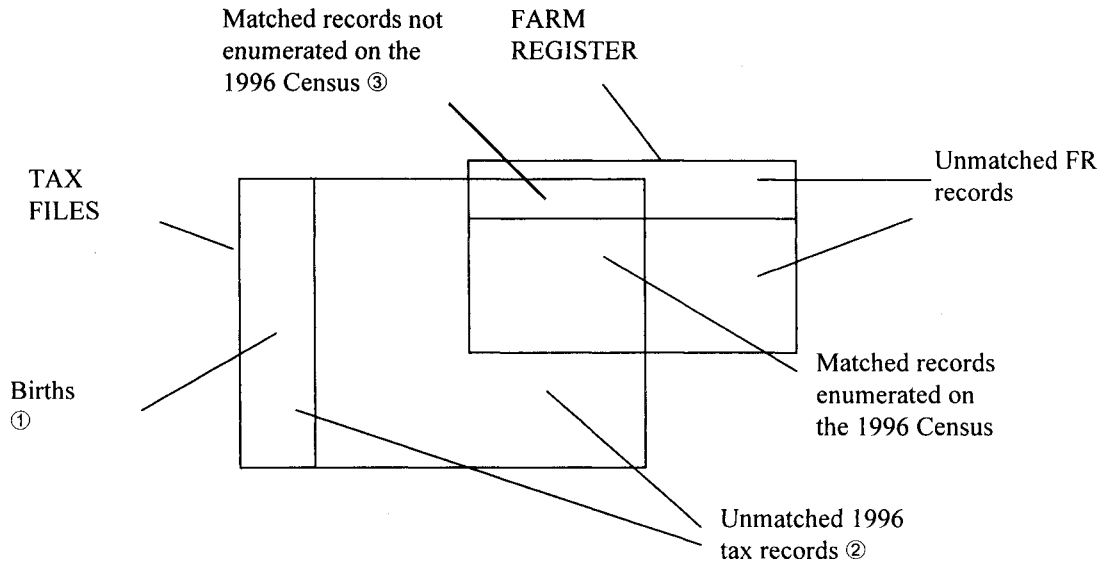

Diagram 1: Sub-universes defined by matching T1 and T2 files to the Farm Register

selected. Stratification is performed within each province, by assigning the records to one of the three sub-populations of interest (births, matched to FR but not enumerated in the 1996 Census, unmatched to the FR but not potential births), by gross farm income / sales thresholds and by SIC (for T2 records only). Allocation parameters and methods are yet to be examined, once the results of the pilot survey are available. As stated in section 3.3, the sample of farms potentially missed by the 1996 Census is to be selected once from the 1996 tax record match with the FR.

Sampled units are then contacted to establish their status as of Census Day and to obtain information about the farm level structure (partners, farm name(s)). Information about agricultural activities is also collected in order to update the FR for surveys that use a dynamic frame approach. An overlap detection activity is then conducted with the 1996 Census list of farms. Links between farms and farmers are established in order to adjust sampling weights to represent farms. The weight share method is to be used. This method, initially developed in the context of longitudinal surveys (Ernst (1989)), has been generalized to situations where a population of interest is sampled through the use of a frame which refers to a different population, but linked somehow to the first one; in this case, farmers that are linked to farms (Lavallée (1995)).

A second contact is then to be performed with non-overlapping farms to collect data pertaining to each survey of interest. As for the AFS, a multiple frame screening estimator is then to be used.

4. Pilot Survey

A pilot survey was conducted in September 1997 with the following objectives:

i) to assess the proposed methodology in identifying births as well as any potential problems with the kind of information requested from the farmers;
ii) to develop an automated approach for overlap detection and to develop procedures for manual resolution;
iii) to develop procedures for the establishment of links between partners of the same farm; and
iv) to evaluate what sampling rate is required to reach the target sub-population of interest.

The pilot survey was conducted in two provinces: Quebec and Manitoba.

4.1 T1 units

Table 1 shows the number of units in the original 1995 T1 file by gross farm income (GFI) threshold. A file of potential births was then created by comparing the 1995 and 1994 tax files.

Table 2 provides the counts of potential births identified from the 1995 tax file, by gross farm income threshold. Only potential births with a gross farm income of more than $1,000 were retained. Potential births with an unknown gross farm income were kept in case the gross farm income amounted to more than $1,000. Notably, 40 % of potential births have a gross farm income between $1,000 and $10,000 while

according to table 1, about 24 % of the Quebec and Manitoba population falls within this interval. This fact is not surprising since new farms tend to be small.

GFI ($)	Quebec	Manitoba
0 - 1,000	2,711 (5 %)	2,034 (5 %)
1,001-10,000	12,064 (23 %)	11,561 (27 %)
10,001-100,000	19,266 (37 %)	18,429 (42 %)
Greater than 100,000	15,816 (30 %)	9,717 (22 %)
unknown	2,798 (5 %)	1,928 (4 %)
TOTAL	52,655	43,669

Table 1: Counts of T1 units in the 1995 tax file

GFI ($)	Quebec	Manitoba
1,001-10,000	1,388 (40 %)	752 (40 %)
10,001-100,000	887 (26 %)	456 (24 %)
greater than 100,000	414 (12 %)	187 (10 %)
unknown	755 (22 %)	484 (26 %)
TOTAL	3,444	1,879

Table 2: Counts of T1 units in the 1995 tax file that were identified to be potential births for the year 1995

Finally, a sample of the T1 potential births was contacted using a Computer Assisted Telephone Interview (CATI) collection methodology. Table 3 presents counts of units selected, by gross farm income threshold.

GFI ($)	Quebec	Manitoba
1,001-10,000	635 (35 %)	752 (40 %)
10,001-100,000	508 (28 %)	456 (24 %)
Greater than 100,000	301 (17 %)	187 (10 %)
unknown	348 (19 %)	484 (26 %)
TOTAL	1,792	1,879

Table 3: Counts of T1 units contacted using CATI

The sample consisted of a census in Manitoba and an approximately 50 % sample in Quebec. To date, based on the answers provided during the interview, approximately 10 % of the T1 potential births contacted are true births for 1995. However, the most interesting result is that about 37 % of T1 potential births were found not to be involved in agriculture. We thus need to modify the methodology in order to identify these cases before sampling and improve the success rate expressed as the number of true births out of the number of cases contacted. The overlap detection with the Census list of farms is on-going as is the establishment of links between farmers and farms.

4.2 T2 units

Of the 7,108 units in the 1995 T2 tax files of Quebec and Manitoba, 325 were identified to be potential births. A sample of 75 potential births was selected to be contacted. In contrast to the pilot survey for the T1 sample, T2 units were contacted by an interviewer at Head Office who used a paper questionnaire. A telephone number and the name of a person associated with a T2 unit was usually available to us (from other administrative files or telephone books) but we did not know the relationship of this person to the corporation. We decided that it would be better if an interviewer at Head Office, who was experienced with dealing with larger farms and corporations, traced the appropriate person with whom to speak, and conducted the interview. Also, because the tracing procedure could be long and involved, the programming required to accommodate it in CATI would be too complicated and/or require too much time.

Of the 50 T2 potential births contacted to date, just one is a true birth based on the answers provided during the interview. This result might indicate that true births of corporations which are at least partly agricultural are rare. The overlap detection with the Census list of farms is on-going.

5. Conclusion

There is no doubt that in order to produce accurate statistics of the agricultural sector, one has to take into account its dynamic nature. An updating mechanism has thus to be put in place. Statistics Canada already draws a sample of farmers for whom data from financial statements and balance sheets are collected. There are conceptual differences between income tax data and the Census of Agriculture, so to date the information has not been reconciled on a production

basis. This is therefore a big challenge, especially on the conceptual side.

Other administrative sources are being examined, especially to overcome the long time lag of the T1 and T2 tax files. The Goods and Services Tax file is one of these sources. This administrative source became available to Statistics Canada on a monthly basis, in December 1996. Many diverse investigations are presently taking place on this file. Once we know more about it, this file could potentially become our main source of updates, given its (potential) timeliness.

6. References

Armstrong, B. (1979). *Test of Multiple Frame Sampling Techniques for Agriculture Surveys: New Brunswick, 1978.* American Statistical Association, Proceedings of the Section on Survey Research Methods, 295-300.

Ernst, L. (1989).*Weighting Issues for Longitudinal Household and Family Estimates.* Panel Surveys. New York: John Wiley and Sons, 139-159.

Ingram, S. and Davidson, G. (1983). *Methods Used in Designing the National Farm Survey.* American Statistical Association, Proceedings of the Section on Survey Research Methods, 220-225.

Lavallée, P. (1995). *Cross-Sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method.* Survey Methodology, Volume 21, 1, 25-32.