Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

DATA COLLECTION METHODS FOR THE 2001 CENSUS OF AGRICULTURE

Claire Bradshaw[1], Joseph Duggan, Claude Julien, and Rosemary Villani

ABSTRACT

With the goal of obtaining a complete enumeration of the Canadian agricultural sector, the 2001 Census of Agriculture has been conducted using several collection methods. Challenges to the traditional drop-off and mail-back of paper forms in a household-based enumeration have led to the adoption of supplemental methods using newer technologies to maintain the coverage and content of the census. Overall, this mixed-mode data collection responds to critical needs of the census programme at various points. This paper examines these methods, several quality assessments, and future challenges to obtain a co-ordinated view of the methods' individual approaches to achieving data quality.

KEY WORDS: Drop-off/mail-back, CATI, Internet Collection, Quality Evaluation, Census development.

1. DATA COLLECTION

As with all censuses, the data collection for the 2001 Census of Agriculture was conducted with the final objective of obtaining complete and accurate data from all farming operations in Canada. Although the objective is simple, the collection process to achieve it has, by necessity, become more complex. Emerging realities in agriculture have led to modifications in the traditional census collection. New collection modes have been developed over several censuses in response to the changing situations of our respondents. Innovations and improvements in technology have provided further motivation for these changes. This first section of the paper will present the traditional data collection method, then describe the current collection realities for Canadian agriculture, and finally provide an explanation of the new methods that have been implemented for the 2001 Census of Agriculture.

1.1 The Traditional Collection Method

The 1956 census marked the beginning of the cross-Canada joint collection of the Censuses of Agriculture and Population. As the majority of agricultural operations are household-based, they can be effectively and accurately collected using the same methodology. This longstanding partnership has yielded many other benefits for the Census of Agriculture. Significant benefits include: savings for the Canadian taxpayer, a comprehensive communications program, an experienced Canada-wide collection management team, a time-tested process, and a facilitated data linkage between agriculture and population to produce socio-economic data about farm operators and their families.

In 2001, about 32,000 census representatives had the task of collecting the Census of Population for close to 12 million households, while collection for the Census of Agriculture impacted fewer than 300,000 households. The census representatives received from 25 to 30 hours of training on their basic duties for the Censuses of Population and Agriculture, and those encountering agricultural operations required an additional 4.5 hours training for editing and follow-up of agriculture questionnaires.

[1] Claire Bradshaw, Agriculture Division, Statistics Canada, 12-D8 Jean Talon Bldg., Tunney's Pasture, Ottawa ON K1A 0T6, e-mail: claire.bradshaw@statcan.ca

In the data collection phase, most households (about 98%) are enumerated by the drop-off and mail-back methodology. The census representative drops off a Census of Population questionnaire at each household in their assigned area. When the census representative confirms that a farm operator resides in the household (when contact is made with someone in the household), a Census of Agriculture questionnaire is also delivered. When contact is not made, the census representative will use their judgement and training to identify agricultural operations. Respondents are instructed to complete the questionnaires and forward them in the enclosed postage paid envelope.

The mailed-back questionnaires are returned to the census representative, who edits the forms and conducts a telephone follow-up for any questionnaires that fail edit. When dropped-off forms have not been returned by mail within a certain period of time, a non-response follow-up is required. Usually, up to three telephone contacts and four personal visits will be attempted in order to obtain a response. If no response is forthcoming, either from a direct refusal or continued no contact, a missing questionnaire form is completed. These non-respondents will be followed-up by the supervising census commissioner and in extreme cases, the requirements of law (as specified in the Statistics Act) are used to obtain participation.

Within the growing and complex Census of Population collection, Census of Agriculture has a number of tools for field staff to enhance the collection for agriculture. One of the key tools is the agricultural operator screening question on the Census of Population questionnaires. The screening question asks, "Is anyone listed (as a household member) a farm operator?" With this identification, the census representative ensures that farming households that were missed at drop-off receive agriculture questionnaires. Some of the other tools include: a farm count for each enumeration area from the previous census, special training for field staff on identifying agriculture operations which emphasises easily missed operations, and for urban and rural areas with few farms, a list of agriculture operations from the previous census. This list indicates the addresses and telephone numbers of known operators from the last census, and has been updated to account for new farming operations, operations no longer farming, and other changes as indicated from agricultural surveys during the intercensal period.

1.2 Changes in Agriculture

When planning the data collection, the evolving nature of agriculture needed to be considered. While certain trends had been detected in previous censuses, planners had to bear in mind that changes over the five-year gap between censuses could be more significant than expected. Some of the changes in agriculture that were considered for 2001 are outlined below.

- Large corporate farms represent a growing proportion of all farms in Canada. Many of these corporate farms have widespread land holdings, often extending over many provinces, as well as complex operating arrangements. Contact needs to be maintained through a corporate head office because these operations cannot be identified during door-to-door enumeration.
- More farm operators also work off the farm, making contact at the door less likely. Similarly, more farm operators do not live on their operations. In these cases, census representatives cannot confirm that an agriculture questionnaire is required.
- A declining number of farms, a declining percentage of rural population, and an increasing number of non-traditional operations are involved in farming, all of which increase the difficulty in identifying operations.
- There is increasing turnover in farm ownership between censuses; any list of agricultural operations is rendered a less-effective coverage tool.

1.3 Additional Collection Modes

The Census of Agriculture first initiated a change in collection mode as a result of the corporate farms trend. A special collection, with contact at the corporate business address, was formally implemented in 1986. (In the 1981 Census, an ad hoc process collected Census information for some operations missed in the traditional collection.) The special collection has since evolved to a unit that continually updates and maintains profiles on large corporate farms for Agriculture Division's regular survey program. In a census year, the operations are contacted and profiled prior to Census Day and arrangements are made to collect census information as appropriate. This operation has grown from about twenty-five businesses in 1986 to over three hundred corporate contacts in the 2001 Census.

In the 1996 census, another change to agriculture collection resulted from changes to collection dates and a new reference date implemented by the Census of Population. The traditional first Tuesday in June reference date was moved to the second Tuesday in May. By moving the date forward, it allowed the delivery and completion of the questionnaires within the same month, so those households that move (usually at the end or beginning of a month) would not be as easily missed.

In response, the Census of Agriculture implemented the Progress of Seeding Survey. The change in Census Day had an impact on the field crop areas reported by farm operators. Much of the field crop seeding across Canada typically occurs before the first of June. Depending on the spring weather, a large portion of the crop may not be seeded when respondents complete their forms in May. In years where seeding is incomplete by mid-May, crops seeded after Census Day could differ from those planned and reported on the census. A progress of seeding survey was first implemented in the 1996 census to verify or update crop data. It involved following up with operators who reported less than 90% of their field crops seeded when they completed their forms. In 2001, about 44,700 farmers across the country were contacted for this survey by Statistics Canada regional offices. The resulting updates from this computer assisted telephone survey were later re-integrated to the database.

The Farm Coverage Follow-up Survey is another collection mode to address coverage challenges for the Census of Agriculture. This survey is a computer assisted telephone interview of a list of larger farms deemed to be missing once census returns are received from the regions. Incoming records are matched to a list of known farms extracted from a farm register. The first Farm Coverage Follow-up survey was conducted as an ad hoc collection during the processing phase in the 1996 census. In 2001, a planned follow-up process for all provinces was used to locate any missed, but active farms. Their updated information was then integrated into the census data before publication.

Completing Census of Agriculture data over the Internet was the newest collection mode in the 2001 Census. The Canadian government strategic direction to provide services on line will require that an Internet option be available for completing census questionnaires in the 2006 Census. In advance of full implementation, this option was offered in two select areas (one predominately urban and the other, rural) during the 2001 Census. The objectives were:

- to obtain a measure of public acceptance of an electronic alternative,
- to evaluate data quality provided via Internet versus paper,
- to obtain some measure of IT infrastructure requirements and
- to assist in the development of a standardised approach to the development of web-based data collection for household surveys.

Households in the two sites received an insert with their 2001 Census questionnaire packages informing them of the option to complete their questionnaire via the Internet. This option included the Census of Population long and short forms as well as the Census of Agriculture form. The insert provided the web site address and instruction on how to access the questionnaires. Once the appropriate questionnaire was selected, the access code that was provided on the insert was keyed into the electronic questionnaire. This action authenticated the user, allowed them to access the application, and then notified the field collection operation once a questionnaire had been received from that household.

A single portal was used for both agriculture and population to ensure a common look and feel. The Internet questionnaire was designed to resemble the paper version to control the effect on data quality. It also included navigational aids, drop-down menus, and on-line edits based on traditional census representative paper edits.

2. ASSESSING THE QUALITY OF DATA COLLECTION ACTIVITIES

As described in the preceding section, the collection activities for the Census of Agriculture span a considerable portion of both time and space. All of this is done with a network of collection vehicles and procedures, which are intended to supply a complete enumeration of the nation's agricultural sector. The procedures for evaluating the quality of this collection are themselves a collection of diverse activities. This section of the paper will attempt an overview of the primary methods for assessing quality, with particular attention given to the more quantitative measures planned to provide input to the development of the next Census of Agriculture.

2.1 Methods for Assessing Quality

Of the various methods for assessing quality, it is difficult to choose one that is the most important. Every collection method ultimately has an impact on the quality. And in return, every quality assessment ultimately has an impact on the Census. The systematised quality assurance and quality control procedures that are followed in collection provide the most immediate results and improve the quality of the current Census collection. Beyond these, there is another set of evaluations that are focussed on achieving quality in the following Census cycle. The earliest assessment has traditionally been the Census Observer Program for Head Office Staff, quickly followed by the Census Representative Debriefing Study. Later evaluation includes two separate quantitative studies, the Edit Sample Study and the Coverage Evaluation Study.

The Census Observer Program allows staff from the Head Office to visit the field during collection and to observe the field training and collection activities first hand. The organisation gains from the individual observer's experience through a written report that is due shortly after the observation. During the 2001 collection period, this program was restricted in rural areas due to concerns over the spread of Foot and Mouth Disease (a serious infection for livestock that is transportable by humans) in response to outbreaks reported from overseas.

The second major source of information from the field concerning data quality was obtained from the Census Representative Debriefing. A sample of three hundred agricultural enumerators was selected to participate by completing a debriefing questionnaire, a booklet with few questions but plenty of space for comment. The first part sought the enumerators' observations concerning the training and the census material provided, whether it was clear which agricultural operations were to be enumerated, and possible improvements to collection procedures. The second part dealt entirely with the Census questionnaire, the layout, wording, and opportunities for other changes. A matrix format was used to obtain information on problems and respondent reaction for each of the twenty-seven steps in the questionnaire. Again, ample space was provided to elaborate on any of the responses to the debriefing questions.

A major initiative for the 2001 Census quality evaluation is the implementation of the Coverage Evaluation Study, to objectively measure the level of coverage achieved through the combined efforts of all collection activities. This measure will be defined strictly in terms of the number of agricultural activities in operation and not on a commodity, or otherwise weighted, basis. The machinery for the study is based on that from the Farm Coverage Follow-up, described in section 1.3 above. However, the study is solely an evaluation and will not be done within a time-frame that would allow its results to be integrated into the census results. Essentially, the study will use several sources to confirm the existence of agricultural operations in a way that allows the determination of whether these operations have, in fact, been missed in the enumeration.

The approach permits the calculation of estimated rates of undercoverage that are methodologically valid - similar to the overall rates produced by the Edit Sample Study, which is treated separately below.

2.1.1 The Edit Sample Study

The Edit Sample Study (ESS), as carried out for the 1991 and 1996 Censuses of Agriculture, provided valuable information concerning the level of quality of the questionnaires received by mail. Originally conceived as a study of unedited responses of a sample of questionnaires completed by respondents, the study was later extended to compare these original responses to the edited and followed-up responses captured in census processing. The study was designed to achieve these two objectives for the 'core' traditional methodology as well as for the complete sets of both the Internet responses and the data provided from the CATI interviews in the Farm Coverage Follow-up. However, for simplicity of explanation, the study will be described initially in terms of the traditional paper-based mode of collection.

The 2001 ESS will provide estimates of edit failure rates, questionnaire follow-up rates, and response rates for every step and for selected questions on the questionnaire. The results will be used in assessing the questionnaire design, instructions, and concepts used in the census. These estimates can be calculated in three study phases to cover three points in collection: phase I examines the original state of questionnaires returned by respondents, phase II looks at the questionnaires after editing and follow-up, and phase III can evaluate the effects of data capture.

To blend with the operational procedures of the census, a two-stage sampling plan was used to select questionnaires for the study. The primary sampling unit was the Census Commissioner District (CCDs) and the secondary units were questionnaires sampled from selected CCDs. At the first stage, 150 CCDs were selected systematically with probability proportional to size (number of farms from previous census) from each region in the study according to a power allocation. In the second stage, four questionnaires were randomly selected on each of the first five days of mail returns in every selected CCD.

If fewer than four questionnaires arrived in one day at the CCD office, additional questionnaires were selected from the following day's mail until the required number was selected from each CCD. If needed, an additional day could be used to make up for the shortage. A control form was used to identify and record the date of selection of each questionnaire participating in the ESS, and to record the number of questionnaires that arrived each day at the office. The content of these questionnaires was then transcribed onto blank questionnaires for use in the study while the originals were put back into the collection processing stream after a 100% quality check by the census commissioner. The transcribed questionnaires and control forms were forwarded to Head Office, were captured, and will be used as the basis for estimates from phase I of the study.

The questionnaire identifiers from the control forms link the phase I study questionnaires with those received from the field through the usual course of collection and now resident on the Census database. Although these are the 'original' questionnaires, they have in some cases undergone significant change (through editing, follow-up and capture) in the collection process. The data for these questionnaires are referred to as the phase III data. Because these data were captured by an automated system, (combining Intelligent Character Recognition and Key-from-Image) there exists an image of the questionnaire prior to capture. This image can be used as a reference to remove the effects of capture through verification between the image and the phase III data. The resulting restored data then reflects what is recorded in the image and is referred to as the phase II data.

The response rates and edit failure rates as determined from the phase II data provide measures of the quality of the data after editing. Also, in comparison to the results from phase I, they indicate the improvement to quality resulting from the field editing and follow-up. The comparison of the calculated quality measures from phase II and phase III indicates the impact of data capture (including quality assurance) on the overall quality of the data. It can reasonably be expected that the effects of capture on response rates and edit failure rates will be rather small, particularly in relation to the standard errors of the

estimators of the differences between any two rates. Alternatively, measures incorporating the size and direction of data capture errors can be devised and weighted using the study sample and weights.

If it were not for the interest in the automated capture system being used, it would be possible (for efficiency's sake) to bypass phase II of the study. In this scenario, phase III data would be compared with those from phase I to arrive at a sufficiently accurate measure of the effects of editing and follow-up – ignoring the effects of capture. Implicit in this understanding is that the vast majority of the field edits centre on the presence of data, or really the lack of it, and do not involve the actual reported values.

There are several opportunities related to the inclusion, on a census basis, of the Internet and CATI data into the ESS. An important point is that all of this data is available to the study without the extra expense (and error) of transcription, capture, and linkage. This reduces the need for sampling and other intervention and so does not add to the operational complexity of the study. The opportunities for analysis are many. The measures of data quality will be available for these two populations separately, and these can be compared to the results from the traditional collection method. Because the editing and capture for the new modes occur simultaneously, it will be impossible to separate these effects. In other words, there will only be one phase of data for the Internet and CATI records. The results for these records are comparable to phase I since they represent the new level of 'initial' data. However, they are also comparable to phase II or III because they have also been refined through the collection process and represent the level of data quality to be expected heading into Head Office processing.

The 2001 Edit Sample Study will then be in a position to provide the first examination of mode effects for the Census of Agriculture. These effects have traditionally been ignored because of the extremely small percentage of the population that they have represented. Ultimately, the consequences of not having alternate modes of collection would have been far more damaging to the Census. It is expected that the use of these modes will increase in coming censuses, and the information that will be obtained from this study will play an important role in the development of the 2006 collection methodology. Further, the knowledge that is gained concerning these mode effects will have an impact on how these data are treated in down-stream Census processing and in subsequent data validation.

3. CHALLENGES FOR THE FUTURE – THE 2006 CENSUS OF AGRICULTURE

3.1 Major change to methodology

The current drop-off and mail-back methodology for conducting the Census was introduced in 1971 and, in each subsequent census, several developments have been made to improve quality. However, increasing public concern over privacy of personal information and the physical security of completed questionnaires make it necessary for Statistics Canada to change the way it collects census data. Farm operators have frequently objected to providing information on their farm businesses to a local census representative, (and a potential competitor) that lives "down the road." Furthermore, completed questionnaires were kept for several weeks in the homes of the census representatives while they were verified for completeness. The census representatives' work assignment (or completed questionnaires) could not be delivered to the census office until collection was completed. This situation presented the possibility of completed questionnaires being stolen, lost or inadvertently seen by someone other than census staff.

These pressures, along with the federal government initiative to provide all respondents with the option of submitting their completed questionnaires electronically, have required the Censuses of Population and Agriculture to move to a more centralised and automated approach for the 2006 Census. The new methodology will involve mailing-out census questionnaires to a majority of Canadian households based on a comprehensive and up-to-date address register. In these areas, the questionnaire will be addressed and delivered through Canada Post. Where street addresses do not exist or the address does not identify the dwelling, households will receive their questionnaires from a drop-off method similar to that used in 2001. About eighty percent of farm households are in these rural areas which will rely on the drop-off process for

coverage of both censuses. Regardless of delivery method, all questionnaires will be pre-printed with a barcode number containing a geographic identifier *and* all questionnaires will be returned to a central data capture centre for processing.

The collection activities used for the 2001 census were controlled by enumerator assignment. As mentioned previously, the questionnaires could not be sent for capture until the collection activities for all dwellings and agricultural operations in the enumerator assignment were completed. In 2006, the control will be at the dwelling or questionnaire level, so that data capture could begin as soon as completed questionnaires were returned. This should shorten the processing period, potentially leading to an earlier release of data.

Data from the paper questionnaires will again be captured using an automated data capture system and subsequently be merged with the electronic responses. An automated edit for completeness will be applied and follow-up with respondents will be conducted from a central location using computer assisted telephone interviews (CATI). A small number of specially trained census representatives will still be required to visit non-responding households. Overall, direct mail combined with centralised processing will significantly reduce the requirement for local census representatives as well as address the privacy and security concerns.

At the core of this new methodology will be an automated control system ensuring that all questionnaires have been received. Paper questionnaires will be registered by means of a barcode, while the receipt of Internet questionnaires are confirmed electronically within the system. The control system will determine which dwellings require follow-up, either for non-response or for failed edit. The system will also track the questionnaires or records as they move from process to process in the data capture centre.

3.2 Use of technology

In 2006, integrated automated systems will be used on a much larger scale than ever before. The roles of the Internet and CATI collections will increase with respect to the traditional paper-based process. All of these systems need to be developed and tested and timely integration becomes critical. As seen in 2001, a distinct collection phase no longer exists. Collection activities will continue to be scattered throughout the processing stages of the Census of Agriculture. These collection activities must be developed and carried out in a coherent and co-ordinated fashion to avoid coverage errors.

The increased use of automated systems will allow the census to process the data faster and smarter. Automated data capture systems allow edits to be applied immediately. On-line edits can be built into the Internet questionnaire and CATI applications, also leading to better quality data. Automation will ensure that all edits are uniformly applied to all records. Automated thresholds can be applied to specific variables, resulting in better selection of the cases for follow-up.

Automation will further reduce the enumerator workforce and their responsibilities. They will no longer be required to verify the questionnaires for completeness, thereby addressing privacy concerns. Automation will require that processes be conducted from central locations with the benefit of a specialised, highly trained, and closely supervised staff, improving the data quality. In addition, greater physical security is afforded from processing in a central location.

In the 2001 census, about one percent of farm households in the Internet option areas completed their questionnaires electronically. Wider access to high-speed connections in the future would make the option of completing the census forms over the Internet more attractive. A full Internet option, with the electronic receipt of questionnaires, requires a robust and efficient control and field notification system. The new methodology, with its centralised approach, also requires a very fast notification system to inform geographically dispersed census representatives of the households to follow-up for non-response. The key to collecting both paper and electronic returns in a seamless fashion will be to provide an integrated and automated control system.

4. CONCLUSION

The traditional collection methodology served the Census of Agriculture well for fifty years. Alternate data collection modes have been adopted to supplement the basic collection, and have allowed us to maintain a consistent level of quality and coverage. The 2006 move to a more centralised and automated approach is a significant change for the Census of Agriculture. The new methodology will provide opportunities to address key privacy concerns, make editing and follow-up more efficient, reduce processing time, and improve data quality, although it *will* provide new coverage challenges. The basic task for the 2006 census will be to ensure that the various collection and data capture activities are integrated (without leaving gaps, or overlaps, in coverage) and are timely in order to maintain a high level of data quality.

REFERENCES

Chartrand, E. (2001), "Respondent Perspectives on Census Data Collection via the Internet", To appear in *Proceedings of Statistics Canada Symposium 2001*.

Duggan, J. (1999), "Methodology Overview Document: 2001 Edit Sample Study (ESS)", unpublished report, Ottawa, Canada: Statistics Canada.

Statistics Canada (2001), "Census Facts – 2001 Census of Canada", unpublished internal document, Ottawa, Canada.

Statistics Canada (2001), "Collection Specifications Book", unpublished internal document, Ottawa, Canada.