A DATA COLLECTION METHODOLOGY FOR THE 2001 CANADIAN CENSUS

G. H. Choudhry, Statistics Canada
Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada.

KEY WORDS: Questionnaire Editing, Follow-up Procedures, Data Quality, Cost of Data Collection

SUMMARY

In order to better protect the confidentiality of the census information, a new data collection methodology is proposed for the 2001 Canadian Census of Population and Housing. The methodology is based on the concurrent collection and processing of census data. The clerical edits and telephone follow-up operations will be centralized into a number of District Offices and hence the term **"Centralized Edit Methodology"** for the proposed data collection methodology. Field follow-up for the nonrespondent households and the failed edit questionnaires that could not be resolved on the telephone will also be controlled from the District Offices. The questionnaires will be bar-coded and an automated questionnaire tracking system will be developed so that it will be possible to find the status of a questionnaire at any point in time.

1. BACKGROUND

The current data collection methodology requires that the mailed-back questionnaires be returned to the enumerator to complete the edit and follow-up process. This raises the issue of personal privacy which is a concern to Canadians. During the 1991 Census, 20% of all census-related correspondence directed to the Canadian Privacy Commissioner, the Chief Statistician, and the Minister responsible for Statistics Canada, dealt with issues of privacy, confidentiality and local enumeration.

The Canadian Privacy Commissioner is currently assessing complaints filed during the 1991 Census and will be recommending changes to the current data collection methodology. The recommendation might be that the respondents had to be made aware that if they did not want their census forms to go back to the local enumerator, the forms could be mailed directly to Statistics Canada in Ottawa. If such an approach is recommended, it would be extremely costly and most difficult to implement this approach under the current data collection methodology.

In 1993, a study to assess the feasibility of centralizing the collection operations into a number of D/Os across the country was conducted by Hicks *et al.* (1993). The outcome of the feasibility study was a recommendation to consider full implementation of the Centralized Edit Methodology for the 2001 Census subject to successful testing during the 1996 Census. A field test of the proposed data collection methodology involving approximately 400,000 households is planned during the 1996 Census. The proposed methodology requires that the completed questionnaires be mailed back to the D/Os of which there will be approximately 50 across the country. The clerical edits and telephone follow-up operations for failed edit questionnaires will be conducted in the supervised environment of the District Offices by a staff most likely unknown to the respondents. Field follow-up will be required only for the nonrespondent households and the failed edit questionnaires that could not be resolved on the telephone. Therefore, the Centralized Edit Methodology will be a means to address the issue of personal privacy.

The organization of the present paper is as follows: The proposed methodology is presented in section 2. Its impact on cost, data quality and timeliness is discussed in section 3. The paper also deals with the implications for the Census of Agriculture (section 4), the Data Quality Studies (section 5), and the Post-censal Surveys (section 6). Finally, section 7 contains some concluding remarks.

2. PROPOSED METHODOLOGY

The major thrusts of the proposal are to introduce mail-out/mail-back where feasible, centralize collection operations and automate document control. The questionnaires will be processed on a flow basis, and the control will be at the questionnaire level. In the current data collection methodology it is at the enumeration area (EA) level. It should be noted that mail-out of questionnaires is not an integral part of the Centralized Edit Methodology. The Centralized Edit Methodology can also be implemented with enumerator delivery of questionnaires as will be the case for rural areas.

The proposed data collection methodology is a significant departure from the current methodology. The data collection will no longer be reliant on a single individual performing all data collection activities in a defined area *i.e.* an EA. Staff in the past censuses have been generalists and were expected to perform numerous tasks. Because of the small number of units involved in each activity the

opportunity to build up expertise did not exist. Under the Centralized Edit Methodology personnel will be hired for a specific operation. They will be trained to perform only one collection activity and will build their skill level over a period of time. Assignment sizes will be larger and the number of staff will be substantially reduced. A number of activities which were previously carried out in the field will instead be carried out in the D/Os under supervision. Field activities will be carried out by field enumerators under the supervision of a Crew Leader who will have responsibility for 10-12 enumerators.

Now we will describe the pre-census day activities which focus on address list compilation and delivery of questionnaires, and the post-census day activities which include all collection activities from check-in to data capture.

2.1 Pre-Census Day Activities

The focus of pre-census day activities is to create an automated complete and accurate address list and to deliver questionnaires (pre-addressed where feasible) to all private dwellings in Canada. The control of all collection activities and processing activities is dependent upon this file.

The following is an overview of the different methodologies that will be employed for the creation of the address lists and the delivery of questionnaires in various area types.

- **Precanvass Areas (60% of the dwellings)** - These are the larger urban centres with a population of 50,000 and over for which an automated address list is available (Swain *et al.* ; 1992). The dwellings in these areas are geocoded to the block-face level through a Street Network File (SNF), and hence these areas are known as SNF urban areas. Enumerators will canvass the areas covered by their assignments and add to, and/or delete addresses listed in the register based on their observations. The address file will be updated on the basis of precanvass operations. The address file will also be periodically updated until Census Day with a Point-of-Call data base from Canada Post. Using the updated file, questionnaires will be addressed and bar-coded, and then delivered to each dwelling by Canada Post.
- **Prelist Areas (20% of the dwellings)** - In the smaller urban centres with a population between 5,000 and 50,000, enumerators will canvass their assignments and list all valid dwellings. These areas will be called non-SNF urban (or Prelist) areas. The addresses will be captured to create the automated address list for the prelist areas. The address file will also be updated through the Point-of-Call data base, questionnaires will be addressed and bar-coded, and delivered by Canada Post. If SNF coverage is extended to urban centres with population 5,000 , the prelist operation will not be required for the 2001 Census.
- **List/Leave Areas (18% of the dwellings)** - These are typically the rural areas for which an adequate address file for mail-out purposes cannot be created at this time. Without names, the addresses, for the most part, are not usable for mailing purposes. Therefore, the same procedure as is currently used for the delivery of questionnaires will be followed. Enumerators will canvass their enumeration area, list each valid dwelling, and drop-off census questionnaire.
- **List/Enumerate Areas (2% of the dwellings)** - List/Enumerate areas correspond to the remote areas and most Indian Reserves. The approach for these areas will also be the same as in the current collection methodology. The enumerator will visit each dwelling in the assignment area, list it, and enumerate respondents through a personal interview.

2.2 Post-Census Day Activities

The questionnaires will be mailed-back directly to the D/Os by the respondents. The focus of post-census day activities will be to ensure that questionnaires are completed accurately before being shipped to data capture.

An automated collection control system will be implemented to control all the operations within the D/Os. The Collection Control File (CCF) which will be a file of all dwelling identification numbers will maintain the status of each dwelling for D/O operations. Questionnaires will be received from Precanvass, Prelist and List/Leave areas, and the check-in of mail returns will begin as soon as the first questionnaire is received. The questionnaires for Precanvass and Prelist areas are bar-coded on the front page at the time of printing and the receipt of these questionnaires will be registered using these bar-codes. Questionnaires received from the List/Leave areas will be bar-coded in the D/Os. From this point on, List/Leave questionnaires will be treated in the same manner as the Precanvass and Prelist questionnaires. The questionnaires will be edited in the D/Os and follow-up action will be taken as required before being shipped to data capture. The bar-coded labels for the List/Enumerate questionnaires will also be generated in the D/Os and the questionnaires will be shipped directly to data

capture. No edit or telephone follow-up action will be taken as it is unlikely that respondents in these remote areas can be contacted by telephone.

The following is a brief description of the edit and follow-up activities in the D/Os.

- **Field Follow-Up for Nonresponse** - Each District Office will generate Nonresponse Field Follow-Up (NRFU) listings by EA from the CCF for Precanvass, Prelist and List/Leave areas. Enumerators will visit each nonresponse dwelling to determine the status on Census Day. The statuses are:
  - the unit was occupied on Census Day and the enumerator will complete a questionnaire by interview;
  - the unit was unoccupied on Census Day and the enumerator will complete a questionnaire with a vacant dwelling status;
  - the unit was out of scope (commercial dwelling, demolished, *etc.*) and the enumerator will assign a delete status;
  - the unit was occupied on Census Day but the enumerator cannot make contact with the household (household on vacation, *etc.*). The enumerator will complete a special questionnaire indicating this status.

  Completed questionnaires will be returned daily or every second day to the Crew Leaders who will review them for completeness and return them to the D/Os.
- **Clerical Edits** - Clerical edit of all completed questionnaires will take place in the District Offices. The questionnaires will be assigned to edit clerks in the form of batches. After each assignment has been edited and has undergone quality control, accepted questionnaires will be shipped to data capture and rejected questionnaires will be assigned to telephone follow-up.
- **Telephone Follow-Up for Failed Edit** - The failed edit questionnaires will be assigned to telephone follow-up. After telephone follow-up has been completed or attempted and the completed work has undergone quality control, the accepted questionnaires will be shipped to data capture. Unresolved mail return questionnaires (*e.g.* no contact, wrong telephone number, no telephone available, refused to complete by telephone, *etc.*) will be assigned for field follow-up. Unresolved nonresponse follow-up questionnaires will **not** be assigned for field follow-up because these questionnaires were completed by enumerator interview and were verified by the Crew Leader. Therefore, it is very unlikely that more information can be obtained by further personal contact by a field enumerator. Additional telephone follow-ups for these cases will be attempted by more experienced telephone operators.
- **Field Follow-Up for Failed Edit** - Enumerators will conduct field follow-up for failed edit mail returns which could not be resolved by telephone follow-up. This will be the main component of follow-up during failed edit field follow-up. In addition, field follow-up will be conducted for:
  - **residual nonresponse** - nonresponse cases not resolved during nonresponse field follow-up;
  - **vacant/delete check** - dwellings which were classified vacant or delete during nonresponse field follow-up to ensure that they were classified correctly.

The field follow-up for failed edit will not be combined with the nonresponse field follow-up because it would delay the nonresponse field follow-up operation which in turn will have a negative impact on the response rate. Moreover, a separate failed edit field follow-up could also serve the purpose of a second check for the dwellings determined to be vacant or deleted during nonresponse field follow-up. Therefore, field follow-up for nonresponse will be completed before field follow-up for failed edit questionnaires can begin.

3. IMPACT ON COST, DATA QUALITY, AND TIMELINESS

3.1 Cost

The results of the Centralized Edit feasibility study show that adopting the proposed data collection methodology for the 1991 Census would have resulted in an additional requirement of approximately $3.4 million in 1991 dollars (Hicks *et al.* ; 1993) which is approximately 3% increase in the cost of data collection. The increase in the cost is mainly due to higher costs for questionnaire production and processing. The differential impact for the data quality studies is not included which would be an additional $2.0 - $2.5 million depending on the option to be implemented. The various options for the data quality studies are discussed in section 5. However, it is expected that the Centralized Edit Methodology may become less costly than the current methodology by 2001 due to potential for automation for the Centralized Edit (*e.g.* use of OCR technology for capturing short forms).

3.2 Data quality

Except for the field follow-up operations, all the collection activities will take place in the controlled environment of a D/O. A questionnaire will be handled by different persons at various stages of collection and processing, *e.g.* editing of questionnaires will be done by edit clerks and the follow-up will be conducted by telephone or field interviewers whereas all the tasks are performed by a Census Representative (CR) under the current methodology. There will be two main advantages of this approach: (1) each person will be responsible only for a particular task for a longer period of time, and therefore becoming more proficient at it. For example, under the current methodology, among other tasks a CR edits between 50 and 60 long questionnaires, whereas under the Centralized Edit Methodology, an edit clerk will be editing these many long questionnaires every day for a period of about 6 weeks, (2) since a questionnaire will be handled by several persons at different phases of collection and processing, there will be a greater chance of discovering and correcting errors. Another major advantage from the data quality point of view would be that the adverse impact on the response rate due to the local enumerator issue which relates to confidentiality will be virtually eliminated. Therefore, the quality of the data will improve under the Centralized Edit Methodology.

The coverage will also improve due to quality control checks during creation of mailing lists for mail-out/mail-back areas *i.e.* precanvass and prelist areas. There will also be up to four matches with Canada Post Point-of-Call database for these areas, the last one at the time of mail-out of questionnaires, resulting in additional improvement in the coverage. These areas account for 80% of the total dwellings to be enumerated during the census. The remaining 20% of the dwellings are in the rural areas, Indian Reserves, the collectives, and canvasser areas. The methodology for the delivery of questionnaires for these areas will be the same as the current methodology, and the coverage will be roughly the same as during the 1991 Census. Therefore, the overall coverage of the dwellings and hence that of the population should improve under the Centralized Edit Methodology as compared with the coverage levels achieved under the current list/drop-off methodology.

3.3 Timeliness

Under the Centralized Edit methodology the control is at the questionnaire level whereas it is at the EA level under the current methodology. Control by questionnaire would allow for concurrent collection and processing activities. Under the current methodology an EA assignment must have completed all edits, follow-up and quality checks before it can go to processing operations. The earliest date for start of processing is seven to eight weeks after the Census Day. Under the Centralized Edit methodology questionnaires will be sent to processing operations immediately after being "accepted" by the edit operation. Control of collection operations at the questionnaire level and concurrent processing activities will make it possible to release the Census data earlier than the present release date.

4. IMPACT ON CENSUS OF AGRICULTURE

Questionnaire delivery and collection for the Census of Population and Census of Agriculture are currently integrated. Under the Centralized Edit Methodology for the Census of Population, the following three options are considered for the Census of Agriculture:

Option 1: Conduct the Census of Agriculture completely separate from the Census of Population.

Option 2: Conduct the Census of Agriculture as a post-censal survey, *i.e.* a fall mail-out census based on the identification of agriculture operators from the completed population questionnaires.

Option 3: Consider full integration of the Census of Agriculture with the Census of Population under Centralized Edit Methodology. In the urban areas, the agriculture questionnaires will be mailed-out from a Farm Register, and additional farm operators will be identified from the completed population questionnaires. These operators will be contacted post-censally and agriculture questionnaires completed. In the rural areas, the agriculture questionnaires will be delivered by census enumerators with the population questionnaires. Editing of agriculture questionnaires will be centralized and field follow-up will be coordinated with follow-up of population questionnaires.

The above three options are assessed in terms of data quality, historical continuity, timeliness, cost and agriculture-population linkage. Option 1 is rejected on the basis of cost implications. Option 2 cannot be accepted either, primarily due to the fall reference date. There will be a break in the historical continuity and also the release of data will be delayed. There is also the risk of increased undercoverage due to self identification of agriculture operators only through population questionnaires. Under option 3 *i.e.* full

integration with Census of Population using Centralized Edit Methodology, the quality of data at the micro level would improve due to more (potential) automation of a process which is currently manual, more specialization of staff, standardization of procedures and more control of collection activities. For these reasons, the preferred option would be full integration with the Census of Population under Centralized Edit Methodology. Adoption of this option would increase costs over the current collection methodology, primarily due to new equipment and facility costs. Moreover, new operational and data quality procedures would be required to match each agriculture operator to the appropriate person enumerated in the Census of Population.

5. IMPACT ON DATA QUALITY STUDIES

In 1991, the data from the reverse record check (RRC) study was used to estimate the undercoverage rates of persons, households and families for the census. The estimates of the undercoverage rates for persons were obtained for broad age/sex categories and a number of other characteristics at the national level, and for the provinces and the territories. The RRC sample in the 10 provinces of Canada was a sample of individuals from a number of non-overlapping frames including the Previous Census Frame. In the two territories (Yukon and N.W.T.), the sample of persons was selected from Health Records.

The sample size for the 1991 RRC study was 56,000 persons out of which 45,000 were selected from the 1986 Census. The sample design for the Previous Census Frame was a stratified two stage sample design. A sample of enumeration areas (EAs) was selected at the first stage, and at the second stage 10 persons were selected from each of the selected EAs. The persons selected for the RRC sample were matched with those enumerated during the census, and the records corresponding to the unmatched persons were traced in the field. The unmatched records were re-matched at other addresses where these persons could have been enumerated. The procedures for matching and search operations are given by Boudreau and Germain (1990). As a result of searching and tracing operations, sample cases were classified into one of the following six categories:

1. enumerated,
2. not enumerated,
3. deceased prior to census,
4. emigrated or abroad prior to census,
5. out-of-scope,
6. unresolved.

The proportion of records that is not resolved is usually small (it was 4.8% for the 1991 Census). More details about the coverage error measurement programme are given by Germain and Julien (1993).

In 1996, the RRC study will also be used to estimate the overcoverage replacing the current overcoverage study. The whole sample will be traced in the field to obtain all the potential addresses where the sampled persons might have been enumerated. These addresses will then be matched with the Census questionnaires to find out whether the persons have been correctly enumerated, over-enumerated, not enumerated and why they were not enumerated (missed, deceased, out-of-scope *etc.*).

For matching with the Census questionnaires, the EA corresponding to the address is identified and its box is pulled out and a search is done through the questionnaires to identify the questionnaires corresponding to the address and to check whether the person has been enumerated at this address. In 1996, the questionnaires of the Centralized Edit test site will be sorted by EA in order to conduct the RRC Study. The sorting of questionnaires by EA will take place before these are sent to data capture. If the Centralized Edit Methodology is implemented at the national level for the 2001 Census, the sorting operation will be very costly and time consuming. Moreover, the sorting operation could not take place before data capture because it would slow down the processing operations, thus further delaying the matching operation. Alternatively, ways could be found to improve the match rate of selected persons with the census data base, thus diminishing the need to access individual questionnaires. For example, the names and addresses for the entire population could be captured on the census data base. The addresses for the urban areas are already on the CCF and only the names and the household IDs will have to be captured. For the rural areas, both the names and the addresses will have to be data captured. Under this option, the search operations can be automated because the individual names would be on the census data base. The quality of the searching will have to be evaluated specially in the rural areas where the addresses are not very precise.

Another option for estimating the undercoverage and overcoverage as an alternative to RRC study would be a Post-Enumeration Survey (PES) similar to the one conducted by the US Bureau of the Census (Hogan; 1991, 1992). The PES is based on an area sample design, and matching and searching operations will be facilitated due to the clustered sample design. The primary sampling unit (PSU) for the PES will be a cluster of dwellings, *e.g.* a block or an EA. In the

urban areas, a sample of blocks can be selected (very small blocks will be grouped with neighbouring blocks). A separate sample will be selected from the highrise apartment buildings (> 4 storeys). In the rural areas, a sample of EAs can be selected, and the sampled EAs can be divided into clusters of dwellings and a sample of two clusters can be selected from each of the sampled EAs. The sample size for the PES will be larger than the RRC sample size due to a more clustered sample design for the PES.

These various options for the data quality studies are being investigated and will be tested as part of the Centralized Edit test during the 1996 Census.

6. IMPACT ON POST-CENSAL SURVEYS

The two post-censal surveys which were conducted in 1991 by selecting samples on the basis of responses obtained during the census are: (i) Health and Activity Limitation Survey (HALS), and (ii) Aboriginal Peoples Survey (APS). The HALS is conducted to obtain more information about Canada's disabled population, and similarly the APS is conducted to obtain information about the aboriginal people. The sample sizes for the HALS and the APS are 150,000 and 180,000 persons respectively. The sample design for both the surveys is a stratified two-stage sample design where EA (or group of EAs) is the primary sampling unit (PSU), and the individuals are selected from the selected EAs on the basis of their responses to certain questions on the census long questionnaire.

The EAs are selected on the basis of information from the previous census. The long census questionnaires for the selected EAs are checked individually to select the sample of individuals with the desired characteristics on the basis of their responses to the selected questions.

Between the two surveys *i.e.* HALS and the APS, long questionnaires are checked for more than half of all the EAs in Canada to select the sample of individuals for these surveys. The data collection starts immediately after the sample of individuals has been selected from a given EA. Under the Centralized Edit Methodology, it will not be operationally feasible to select the sample in the District Offices as the control is at the questionnaire level and there are no EA boxes from which to select the sample. The sample will have to be selected after all the long questionnaires have been data captured. Therefore, there will be some delay in selecting samples for the post-censal surveys. The main advantage for the post-censal surveys will be that the sample selection operation can be automated instead of manual which is more costly and also error prone. The sample will also be more efficient because the selection of EAs can be based on the more up-to-date information from the current census instead of the previous census as is the case under the current methodology. On the other hand, the disadvantage would be that due to the delay in sampling, data collection will be delayed and the census addresses would have become out-of-date which will make tracing more difficult and costly. Alternatively, Poisson sampling which can be implemented concurrently with the data capture operations could be used.

7. CONCLUDING REMARKS

The Centralized Edit methodology will deal very effectively with the problem of local enumeration which relates to the issue of personal privacy and confidentiality of the information. Questionnaires will no longer be returned to the local enumerator. Field enumerators will only deal with nonresponse cases and the cases which could not be resolved through telephone follow-up by the District Office staff. Moreover, the proposed methodology will result in improvement in data quality and timeliness. The methodology also has the potential to benefit from future technological advances, *e.g.* OCR technology.

ACKNOWLEDGEMENTS

I would like to thank my colleagues Jocelyn Tourigny and Michael Hidiroglou for reviewing the paper.

REFERENCES

Boudreau, J.-R. and Germain, M.-F. (1990), "User's Guide to the Quality of 1986 Census Data: Coverage",Statistics Canada Publication 99-135.

Germain, M.-F. and Julien, C. (1993), "Results of the 1991 Census Coverage Error Measurement Program", *Proceedings of the 1993 Annual Research Conference, USBC*, pp. 55-70.

Hicks, D., Choudhry, G.H., Faguy, R., Roberts, K. and Zelenbaba, M. (1993), "Centralized Edit Feasibility Study", Statistics Canada Report.

Hogan, H. (1991), "The 1990 Post-Enumeration Survey: Operations and Results", *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 1-10.

Hogan, H. (1992), "The 1990 Post-Enumeration Survey: An overview", *The American Statistician*, Vol. 46, pp. 261-269.

Swain, L., Drew, J.D., Lafrance, B. and Lance, K. (1992), "The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census", *Survey Methodology*, Vol. 18, No. 1, pp. 127-141.