Recensement Census

NATIONAL CENSUS TEST

AUTOMATED CODING

REPORT #7

RECENSEMENT
96
CENSUS

Statistique Canada Statistics Canada

Canada

NATIONAL CENSUS TEST

AUTOMATED CODING

REPORT #7

Prepared by: Raymond Roy
Date: June 30, 1994
File # 19961-5-50

Executive Summary

The automated coding project for the National Census Test (NCT) was completed successfully. This was demonstrated by the working group meeting the delivery dates and overall objectives for all aspects of the project.

The batch coding rates achieved during the NCT coding were slightly lower than those achieved during the 1991 Census as the coding was done using the standard Automated Coding by Text Recognition v1.06 software with default parsing strategies. The overall batch coding rate achieved was 91.1% compared to 92.2% achieved during the 1991 Census. The intent of this coding operation was to assess the impact of changes to the question wording for the NCT *not* to replicate a production system.

Automated coding was performed on the Edit Sample Study and National Census Test responses comprising of 10,145 and 44,804 person records respectively. A total of 11 questions were coded using the automated system. The two new variables to be autocoded for the 1996 Census, namely, Relationship to Person One and Place of Work were not coded as part of the NCT testing.

Open-ended questions such as the NCT Ethnic Origin contained multiple responses on a single entry line. If this type of question format is retained for the 1996 Census, resolution of multiple responses will necessitate program changes to the automated coding production system. This will permit an efficient and consistent approach when dealing with this type of problem.

Introduction

As part of the National Census Test (NCT), an automated coding project was established to code the written responses from the NCT questionnaires. An initial meeting was held on August 30, 1993 to discuss the requirements of the project, the responsibilities of the working group members and the coding system to be used.

The working group was comprised of representatives from Subject Matter, Operations, Special Surveys and Systems and Integration. Operations was responsible for coordinating the automated coding process, the coding (both batch and manual resolution), and the outputs (coding reports and coding result files). The batch coding was performed by staff from Systems and Integration section who also developed an input screen for the manual resolution codes. Subject Matter performed the manual resolution coding and the analysis of the results. Special Surveys were responsible for providing raw data for coding and for linking the coded files to the database.

The following table contains a list of team members, their area of responsibility and the Subject Matter variables coded:

Table 1 - List of Participants

Member	Area	SM Variables
Tina Chui	Subject Matter (HFSS)	Place of Birth, Citizenship, Immigration, Ethnic Origin, Indian Band/First Nation
Mike Crew	Subject Matter (HFSS)	Place of Birth, Citizenship, Immigration, Ethnic Origin, Indian Band/First Nation
Ginette Dussault	Subject Matter (DEMOL)	Language, Mother Tongue
Jamie Erskine	Systems and Integration	
Brad Hawkes	Subject Matter (COD)	Place of Work
Patricia Murrell	Systems and Integration	
Michel Pouliot	Subject Matter (DEM)	Mobility - 5 Years Ago
Neelam Prakash	Special Surveys	
Raymond Roy	Operations	
Paul Ruscher	Systems and Integration	
Lorie Shinder	Special Surveys	
Luc St. Amour	Subject Matter (DEMOL)	Language, Mother Tongue

Working group meetings were conducted every second week during the NCT automated coding project. A total of 12 meetings were held between August 30, 1993 and March 10, 1994 to discuss progress and address problems associated with coding of the NCT written responses.

Coding

The group took on the added responsibility of coding the Edit Sample Study (ESS) which was comprised of 10,145 person records. The coding of the ESS allowed the participants a chance to learn and adapt to the system prior to coding the National Census Test written responses. There was a total of 44,804 person records captured for the National Census Test.

Automated coding consisted of two parts. The first coding was done through a batch coding system and the residual was coded in a manual resolution system (interactive coding). The batch system used the Automated Coding by Text Recognition (ACTR) software, version 1.06. It should be noted that this was the shelf version of the software with the default parsing strategy, and not the 1991 Production version.

Manual resolution coding was performed by Subject Matter using data entry screens created by Systems and Integration. These screens were created using the FOXPRO software which allowed Subject Matter to enter the data more easily while storing the coded records in a small database for transfer to Special Surveys. Manual resolution coding was done on personal computers in the respective Subject Matter areas.

For certain variables, it was necessary to look at other household information in order to assign the proper codes. As the 1991 production system was not used for coding, it was necessary to run SAS jobs in order to ascertain other household information.

The reference files which were used to code these questions were the same as the reference files used to code the 1991 Census. In some instances, refinements to the reference files were made prior to coding to allow for modifications to the question format and/or cultural changes. An example of these refinements was the Place of Birth variable where Russia was split into many smaller republics since the last census.

The following two tables depict the batch coding results for the Edit Sample Study and the National Census Test by question. The results should not be compared to those achieved during the 1991 Census as the production system and detailed parsing strategies were not used for NCT coding.

Table 2 - Edit Sample Study Batch Coding Results

Question	Total Records	Records Matched	Match Rate
Q.9 - Language	1,028	958	93.19%
Q.11 - Home Language	883	858	97.17%
Q.12 - Place of Birth	950	906	95.37%
Q.13 - Citizenship	299	253	84.62%
Q.16 - Ethnic Origin	13,434	12,649	94.16%
Q.18 - Race	98	65	66.33%
Q.19 - Indian Band\First Nation	46	24	52.17%
Q.22 - Mobility - Inside Canada	1,540	782	50.78%
Q.22 - Mobility - Outside Canada	179	166	92.74%
Q.24 - Language of Education	59	36	61.02%
Q.41 - Language of Work	78	61	78.21%

Table 3 - National Census Test Batch Coding Results

Question	Total Records	Records Matched	Match Rate
Q.9 - Language	6,147	5,718	93.02%
Q.11 - Home Language	5,285	4,969	94.02%
Q.12 - Place of Birth	5,151	4,829	93.75%
Q.13 - Citizenship	1,868	1,536	82.23%
Q.16 - Ethnic Origin	45,691	42,267	92.51%
Q.18 - Race	595	416	69.92%
Q.19 - Indian Band\First Nation	551	358	64.97%
Q.22 - Mobility - Inside Canada	4,871	3,801	78.03%
Q.22 - Mobility - Outside Canada	1,072	1,017	94.87%
Q.24 - Language of Education	113	76	67.26%
Q.41 - Language of Work	395	369	93.42%

The quality of the coding performed was not measured due to time constraints and an in-house version of the coding software. It is believed that the level of quality achieved during NCT coding was slightly lower than the level achieved during the 1991 Census as a result of these two factors.

Also, the Place of Work variable (Q.42) which originally formed part of the NCT coding was not completed because of difficulties converting the two geographic reference points. The NCT used geography linked to the labour force whereas the Census uses geographic references to the province/federal electoral district and Enumeration areas.

The Place of Work manual resolution coding was to be conducted using the prototype developed during the Research and Testing phase for Census processing. On October 8, 1993, it was decided that the Place of Work variable would not be coded during the NCT as a result of difficulties converting the two types of geographic references.

The Relationship to Person One variable (Q.2) was not coded using the automated system as the reference files were created for the 1991 Census question format. There was a significant change to the NCT question format and it was deemed that the reference files were not suitable for this new format.

Schedule of Activities

A schedule of activities was created to provide milestone dates for NCT coding. The working group was able to meet all of the milestone dates in order to complete the project on time.

The following table indicates the list of activities and dates for the automated coding project.

Table 4 - Automated Coding Activities and Dates

Activities	Completion Dates
Delivery of Reference Files	November 26, 1993
Delivery of Edit Sample Study Responses	December 13, 1993
Edit Sample Study Batch Coding	December 20, 1993
Delivery of National Census Test Responses	January 28, 1994
National Census Test Batch Coding	February 4, 1994
National Census Test Manual Resolution Coding	February 21, 1994
Output to Special Surveys	March 2, 1994

Problem Areas Associated with Coding

As with any coding operation, minor problems occurred with some variables and they were dealt with in order to complete the coding according to schedule. The fixes which were performed were suited for the NCT coding but they may not be suitable for 1996 Census coding.

A meeting was held on March 10, 1994 with the intention of documenting special cases that arose during NCT coding in order to eliminate them for the next NCT and to identify any short comings in the coding system in preparation of the development phase for the 1996 Census Automated Coding System.

Open-ended Question Format

The problem which caused the most difficulty for coding the NCT responses was the change in question format on the NCT questionnaire which allowed the respondent to answer open-ended questions. This was particularly the case for Ethnic Origin where the questions only allowed for written responses.

The question was intended to have the respondent answer their Ethnic Origin (up to three separate ethnic backgrounds) listed on three separate lines; however, in many cases the respondent filled all three backgrounds on one line; thus creating a multiple response.

The system was designed to handle only one entry per line. Numerous occurrences of this situation has caused manual intervention by Subject Matter in order to rectify the situation. A flat file was created and Special Surveys staff had to split the entries contained in line 1 and place them in the areas designated for entries 2 and 3.

An example of this would be:

English/French

contained on line 1

After manipulation of the data by special surveys, the entry on the coded file looked like:

English
French

The decision was taken at the start of NCT coding to allow only three entries to be coded per write-in variable. Subject Matter noted that for many cases more than three entries were received from the respondent. A decision will have to be made as to the upper threshold for entries and how to handle situations which exceed that set limit.

Special Surveys found it difficult to link the codes assigned as part of a multiple response. They felt it would have been much easier to link these codes at the end of processing once Subject Matter had an opportunity to look at all of the codes before deciding on the appropriate course of action.

Resolution of multiple responses could be handled by assigning a unique code (ie. 888) during initial coding and these codes could be corrected at the end of processing.

The multiple response problem will continue to be problematic with the use of open-ended questions. If this line of questioning is pursued for the 1996 Census or the next NCT, further system development will be required to handle the resolution of these cases. With system development, the automated coding system should provide an efficient method for coding open-ended questions.

Data Entry Screens

The FOXPRO data entry screens which were developed by SIS because ACTR does not have the capacity to perform this entry. Special Surveys had originally created a sequential flat file to perform this function. SAS runs were requested by Subject Matter in order to facilitate their coding as other household information was sometimes required.

This issue will not be a problem for the 1996 Census as the same (or improved) screens which were available in 1991 for coding and validation of the Socio-Cultural and Mobility variables will be available including access to other household data via the system surrounding ACTR.

Direct Matching Thresholds

Another issue which will have to be investigated further for the 1996 Automated Coding production system is setting the proper threshold for Direct Matching. The standard upper and lower thresholds of 1/3/10 caused records to fail the matching process. Subject Matter intervention was required to resolve these cases; whereas, these records were system coded when the threshold was made more flexible.

Manual Resolution Coding

Manual resolution coding was complicated due to the length of the sequence identifier (22 bytes long) assigned by special surveys to each record. This could be simplified by assigning a unique identifier consisting of only 7 bytes. This would make it less susceptible to transcription or data entry errors.

For the purposes of NCT coding, codes should not be changed after they have been transferred to Special Surveys as the linking program needs a stable environment in order to function properly. These cases could be rectified for Census processing by the use of Global Fixes at the end of processing once sufficient cases had been identified as needing correction.

Conclusions

The automated coding project for the National Census Test was a success due to a team effort to achieve the objectives and milestones of the project.

The problem areas addressed in this report are exclusive to the NCT and will not have an impact on the 1996 Census, with one exception. Open-ended questions created multiple responses and these were handled with programs specifically designed for the NCT.

If this type of question format is retained for the 1996 Census, resolution of multiple responses will necessitate program changes to the automated coding production system. This will permit an efficient and consistent approach when dealing with this type of problem.