AN OVERVIEW OF THE CANADIAN LABOUR FORCE SURVEY DATA PROCESSING SYSTEM

Ian Macredie, Statistics Canada

I. INTRODUCTION

The Canadian Labour Force Survey serves as the primary source of current employment and unemployment statistics in Canada. It is a continuing, monthly survey of 56,000 households based on a multistage, stratified, probability, area sample. This sample is divided into six equal, representative panels with one panel being replaced each month after having been in the survey for six consecutive months. The computer hardware support for the survey consists of a Head Office main frame and nine mini-computer installations. One of these mini's is located on site with the main frame and serves as a communication centre for the other mini-computer installations which are located across the country in the eight Regional Offices of Statistics Canada. The regional mini-computers are linked to the Head Office mini-computer through dial-up telecommunications lines (with "scramblers" to maintain security).

II. SAMPLE CONTROL AND DATA COLLECTION

Conceptually the survey's data processing system can be said to begin with data entry of sample cluster address lists into the Regional Office mini-computers via VDU consols(1). These lists are transmitted to the Head Office main-frame which performs dwelling selection for each month's sample based on prescribed random starts and cluster sampling ratios. The results of this selection process (combined with selections performed previously) drives the creation of what is referred to as the STRUCTURE AND RESPONSE FILE. This file constitutes the basic operational data base of the data collection phase of each month's survey and contains records representing each and every one of the 62,000 dwellings selected for each month's survey sample(2).

The STRUCTURE AND RESPONSE FILE is then split according to geographic area and the sub-files are transmitted to the appropriate Regional Offices. This file is used in the Regional Office to drive the preparation of questionnaire (using line printers constituting part of the mini-computer installation) and the questionnaires are shipped to the interviewers (1,100 in total) who conduct the interviews and return to completed documents(3) to the Regional Office.

The completed questionnaires are received in the Regional Office on a daily basis where their content is converted to machine readable form using the VDU consols. The action of data entry updates the STRUCTURE AND RESPONSE FILE and the updated portions of this file are transmitted to Head Office nightly. It should be mentioned that no editing or imputation is performed prior to or during data entry. Rather, the data entry process has been expressly designed to create a record replicating the questionnaire exactly as it is received from the interviewer.

III. HEAD OFFICE DATA PROCESSING

As the updated STRUCTURE AND RESPONSE records arrive in the Head Office they are accessed on a read only basis to create a CONSOLIDATED FILE. This is a single file which serves as the data base for all subsequent processing (coding, editing, imputation, etc.). In order to maximize the timeliness of the publication of the survey's results, processing is done concurrently with data receipt. The data are released to the public three weeks after the reference week on which they are based.

Once transmission from all of the Regional Offices is complete, the STRUCTURE AND RESPONSE FILE is prepared for transmission back to the Regional Offices. This preparation consists of three basic operations. Firstly, records associated with dwellings which have been in the sample for 6 months are deleted. Secondly, skeletal records representing newly selected dwellings (replacing those deleted) are added to the file. Thirdly, the content of those records continuing in the sample is reduced to those fields which will be printed on the questionnaires in the Regional Offices in the following month. The information printed back consists of the demographic characteristics plus selected labour market activity variables.

The system used for editing and imputation of the collected data consists of a single large program using a magnetic tape, sequential file, data base. The task breakdown for processing is such that the computer system performs the edits (identifies error conditions) while a clerical group provides the corrections (imputations) by following detailed prespecified procedural instructions (expressed in decision tables). The man/machine interface is accomplished using optical character recognition (OCR) turnaround documents.

The edits are exhaustive in that they articulate all of the valid logical relationships and acceptable value ranges in the data set. The edits are strictly applied, that is, imputations must be performed until all edit conditions are satisfied.

The imputation values applied to correct edit failures are obtained from three sources. In descending order of frequency of application these are: the internal logic of the questionnaire, the previous month's response for the same person to whom the record in error refers, and the corresponding fields of similar records(4) selected using a modified hot deck routine.

The data processing system makes extensive use of the panel feature of the LFS sample design. As each record enters the processing system (by the creation of a CONSOLIDATED RECORD read from the STRUCTURE AND RESPONSE FILE) it is matched with the CONSOLIDATED RECORD, where available, from the previous month for the same respondent. Matches are found for about 80 per cent of the records. Information from the previous month finds three applications:

(i) As mentioned, it is used where appropriate as a source of imputation values in the case of edit failures.

(ii) It is used to correct for total record non-response. In specific situations,

current records are created for non-respondents by bringing forward last month's record and adjusting certain time specific variables(5).

(iii) It is used to augment the data set in such way as to reduce both respondent burden, and, response and coding variance. Specifically, for persons not currently employed, certain data elements refer to 'historical' events which generally remain unchanged in the in-sample period (for example, the date when the respondent last worked, reason for leaving last job, etc.). Responses to these variables are generally collected only in the first interview and carried forward from one month to the next by the Head Office processing system. Similarly, for the employed, industry and occupation coding is done the first time a record enters the system. Unless a job change occurs, these codes are carried forward from one month to the next.

As mentioned, the STRUCTURE AND RESPONSE file is accessed by the processing system on a 'read only' basis, that is, the file is not edited prior to being turned around to the Regional Offices. However, once editing and imputation have been completed, detailed reports on the edit failures are transmitted via the mini-computer network (as a separate transmission) to the Regional Offices who in turn forward them to the interviewers. The purpose of course, is to inform the interviewers of their recent errors with the expectation that this will help them avoid the replication of these erroneous practices in future surveys.

IV. IMPROVEMENTS TO THE SYSTEM

Perhaps the most significant improvement to the system, and one for which acceptance testing is currently underway, is the introduction of automated industry and occupation coding. Verbal descriptive information consisting of the name of the employer, the type of business and the nature of the work performed are available on the CONSOLIDATED FILE records. At present these verbal descriptions are examined by coding clerks who assign the appropriate standardized industry and occupation codes. With the automated coding system, these verbal descriptions will be matched against previously coded descriptions stored in a 'look-up' library and a substantial proportion of the records will be coded without recourse to coding clerks.

The processing system as it presently exists has functioned more than satisfactorily since its implementation three years ago. The man/machine interface of the system plus the exchange of error condition information with the Regional Office staff and the interviewers has given the subject matter logic of the edits wide critical exposure to individuals ideally suited to evaluate its relevance and reasonableness. In addition a body of data can now be derived on the statistical properties of our imputation procedures. Given this accumulation of data and experience we are now in a position to refine the edit logic and the imputation algorithms and to automate many of the imputation routines presently performed clerically.

The mini-computer installations offer the potential for a number of data processing enhancements although financial constraints will preclude their implementation for some time. Specifically, if telephone interviewing were done in the Regional Offices, one could eliminate the use of paper questionnaires altogether and have the responses entered directly in the system through the VDU consols. Such developments could be taken one step further with on-line editing and error resolution taking place in the course of the actual interview. Computer assisted telephone interviewing (CATI) projects conducted in the United States have already proven the feasibility of this technique.

V. CONCLUSION

The Canadian Labour Force Survey data processing system has been found to be an effective support mechanism for this large-scale, continuing survey. While enhancements to this system are both possible and desirable, the existing system has already shown itself to be sufficiently robust that the basic structure is likely to remain in place for some time to come.

FOOTNOTES

(1) The cluster is the penultimate stage in the sampling process and generally consists of one city block in urban areas. For a detailed description of the sample see: Methodology of the Canadian Labour Force Survey 1976, (Statistics Canada Catalogue No. 71-526 Occasional).

(2) The count of dwellings selected (62,000) exceeds the count of households in the sample (56,000) due to the selection of what prove to be vacant dwellings, dwellings occupied by persons outside of the survey universe, etc.

(3) All questionnaires are returned, either containing the information collected through interviews or containing documentation on the reason for non-response. Accordingly, at this and every other stage of the survey's data processing system, a total accounting is kept of all dwellings selected for the survey.

(4) The characteristics vector used to define 'similarity' depends on the field in the record requiring imputation. In defining the population of donor records (the hot deck) extensive use is made of the various stages of stratification used in the sample selection process.

(5) This process of "carrying forward" previous month's data is performed only when the previous month's information consists of actual response vectors for that month. It is further restricted to those cases where the interviewers non-response reports indicate that the same persons previously interviewed are still resident in the same dwelling.