EDIT AND IMPUTATION IN THE LABOUR FORCE SURVEY

Anis Ashraf and Ian Macredie, Statistics Canada

1. INTRODUCTION

The objective of this paper is to describe the edit and imputation system used in the Labour Force Survey (LFS).

The LFS serves to provide monthly estimates of the three mutually exclusive categories of employed, unemployed and not in the labour force as well as a wide range of descriptive statistics relating to these categories. The data produced by the survey include not only a large and complex set of stock measures of labour force and related characteristics but also a number of flow statistics. These flow measures serve both to describe the composition of the stock estimates and to explain the net changes in these stocks from one month to the next.

The target population for the LFS is defined as the civilian, non-institutional population 15 years of age and over and resident in the ten provinces of Canada. The sample design is a stratified, multistage, probability sample of dwellings.

In the delineation of strata, a number of factors are taken into consideration. These factors include socio-economic condition, the "importance factor" on the basis of labour force participation by industry, type of dwellings, configuration, type of area, i.e. urban-rural, etc. Each of the ten provinces are considered an independent stratum. Within each province the Economic Regions (ER) are considered primary strata and further stratification is carried out within the ER boundaries. The last stage of stratification is a geographically contiguous area with the exception of "Apartment" strata. The apartment buildings of a certain description within a city constitute one or more strata. The Hospitals, remote areas, etc., constitute the special area stratum.

A dwelling is selected in two or more stages of selection depending on the type of area. The household(s) within the selected dwellings are interviewed once a month for six consecutive months. The sample is selected in a manner such that for a given month it consists of six distinct and equal groups of dwellings differing only with respect to the number of months that they have been in the sample. All household members are interviewed about their demographic characteristics and the eligible members about their historical and current labour market activities.

The LFS questionnaire design and data collection and capture method will be briefly summarized to provide a background for the description of the edit and imputation system.

2. QUESTIONNAIRE DESIGN

The survey employs two documents for data collection. The first is the HOUSEHOLD RECORD DOCKET (F03) on which is recorded all of the demographic information (age, sex, marital status, etc.) for all household members regardless of age. The second document is the LFS QUESTIONNAIRE (F05) which records the labour market activities of each household member 15 years of age and over. In other words there is one HOUSEHOLD RECORD DOCKET for the entire household but there are as many LFS QUESTIONNAIRES as there are adults in the household.

This allocation of the survey's content reflects an important distribution in the characteristics of the data; a distinction which, as will be seen, is reflected in the organization of the editing function. Specifically, five out of the six demographic items contain logical interrelationships among the household members while all of the labour force activities of each household member are considered to be logically independent of the corresponding activity data of the other household members. Accordingly, by recording all of the demographic information on all household members in a physically contiguous fashion (and collecting it in an uninterrupted block of time within the interview) the interviewer is afforded an opportunity to efficiently review the demographic information which she has recorded on the DOCKET and to detect and resolve any logic failures in the presence of the respondent.

The LFS QUESTIONNAIRE, while self-contained as far as inter-household member logical relationships are concerned, is nevertheless characterized by a detailed network of internal logical relationships. Both its highly structured quality and the network of relationships are attributable to the objective of meeting the survey's statistical goals while minimizing respondent burden. Since the primary objective is the division of the population into the mutually exclusive categories of employed, unemployed and not in the labour force, and given that these categories are established using a hierarchy of labour market activities (e.g. working takes precedence over looking for work), the questionnaire begins basically with a few pivotal questions to establish which of the relevant labour market activities the respondent is engaged in. Almost all of the subsequent questions are confined to the enumeration of the attributes of this particular activity.

3. DATA COLLECTION AND CAPTURE METHODS

The data collection and capture process for the LFS is supported by a trans-Canada mini-computer network. Specifically, each of the eight Regional Offices of Statistics Canada has a mini-computer installation which is connected via telecommunications line to the Head Office computing centre. Each installation consists of one or more CPU's, a high speed line printer and a battery of VDU consoles for data capture. Each month questionnaires (Form 03's and 05's) are prepared in the Regional Offices from files transmitted to them from the Head Office. For dwellings in the sample for the first time, this preparation consists of printing dwelling identification information only. For all other households, all of the previously collected demographic data is printed back on the Form 03 as are selected fields on the Form 05. These prepared documents are shipped to the interviewers who contact the selected household in person or by telephone. The documents are

completed by the interviewers and returned immediately to the Regional Office. In the Regional Office, the content of all of the questionnaires is converted to machine readable form through the VDU consoles and the data is transmitted to Head Office.

It should be pointed out that virtually no editing or imputation is performed prior to or during the data capture process. The objective of data entry is to create a mirror image of the questionnaires exactly as received from the interviewers.

4. THE EDIT RULES

The edit rules used in the data processing system are divided into the three sequentially applied groups of: record structure edits, demographic data edits, and LFS Questionnaire edits. These edit groups correspond to three distinct portions of the record each of which will be described subsequently in all three groups. The edits cover virtually all of the discernable logical relationships. With a very few exceptions, these edit rules are strictly applied to the records in processing, that is, the edits cannot be overruled or suspended for any given record or groups of records. The result is that imputations must be performed until a given record will pass all of the relevant edits before that record can be labelled 'clean'.

As mentioned these three edit groups are applied sequentially and within these groups, smaller groups or even edits involving logical relationships within a pair of fields are also applied sequentially. The sequential nature of edit application means that particular edit failures will halt a record's progress through the system and until the failure is resolved by imputation, no further edits are performed. Finally, at various stages in the edit system, edited and successfully imputed fields in the record are isolated or 'sealed off' and further imputations cannot be applied to these fields. In subsequent edits, if reference is made to these isolated fields, only those fields still accessible for imputation can be modified in order to resolve the edit failure. This may constitute something of a compromise in the statistical properties of the imputations by constraining the values which some of the imputations can assume. However, it has the extremely desirable property in a very tightly scheduled processing environment of ensuring that records do not regress back to editing and imputation stages already completed.

4.1 Record Structure Edits

These edits ensure that the basic record structure is correct according to the following decision table. (This is an extremely simplified version with the actual logic running to 18 pages of decision tables). In reading this decision table it should be remembered that the file contains records for a dwelling selected for the sample, all previously responding household members and all currently responding household members.

DECISION TABLE 1 - RECORD STRUCTURE EDITS

(Schematic representation)

MEMBER OF RESPONDING HOUSEHOLD THIS MONTH	Y	Y	Y	Y	Y	N	N	N
AGE ≥ 15	Y	Y	Y	N	N	Y	Y	N
LFS QUESTIONNAIRE DATA PRESENT	Y	N	N	N	Y	-	-	-
RESPONDENT LAST MONTH		Y	N	-	-	Y	N	-
. CREATE LFS QUESTIONNAIRE DATA BY HOT DECK *			X					
. CREATE LFS QUESTIONNAIRE DATA FROM LAST MONTH RESPONSE **		X				X		
. DO LFS QUESTIONNAIRE EDITS	X							
. DO DEMOGRAPHIC EDITS	X			X				
. LEAVE AS NON-RESPONSE							X	X
. RESOLVE STRUCTURAL ERROR ***					X			

* A record from a HOT DECK is defined as one drawn essentially at random from a population of records of the same Primary Sampling Unit, age, sex, and marital status as the record in error.

** Current month's data is a replication of previous month's data with updates to time specific variables.

*** Resolution is accomplished by changing the age or deleting the LFS QUESTIONNAIRE entries.

4.2 Demographic Edits

These edits cover the demographic data fields consisting of age, sex, marital status, family identifier relationship to the head of the family, and educational attainment. The unique feature of this set of edits is that they are the only set which involve the search for inter-record consistency. This edit group can be further divided into three sequentially applied sub-groups:

(i) Field validation edits, that is, each field in each record must conform to the valid value range specified for that field. For example, the sex field can assume only two values viz. 1 or 2 corresponding to male and female. Values outside of this range, including blanks, constitute edit failures.

(ii) Record specific relationship edits, that is, fields within a record may conform independently to their respective valid value ranges but still fail edits in this subgroup. For example, a respondent cannot be coded as 'spouse' in one field and 'single' in another.

(iii) Family relationship edits, that is, edits which ensure that the family composition is logically correct. For example, each family must have exactly one 'head', parents must be older than children, etc.

All records in a family must pass the field validation edits before any of the records can pass to the record specific relationship edits, and then all records must pass the latter sub-group, before any of the records in the family can pass on to the family relationship edits.

Each time an edit failure occurs and an imputation is applied, all of the records in the family must again pass through the entire set of demographic edits. In this way an imputation applied in response to a family relationship edit failure cannot create an illogical record specific relationship.

4.3 LFS Questionnaire Edits

Conceptually the LFS Questionnaire edits can be grouped into the following three classes although operationally the three types are mixed.

(i) Question sequence edits

The LFS questionnaire is a highly structured document containing six basic paths (and a multitude of secondary paths within this basic six). The paths are identified by reference to a number of pivotal questions whose values determine which path is to be followed through any given record. Edits of this type search for valid question completion sequences and as soon as a departure from a valid path is encountered editing ceases and imputations are called for to restore the appropriate valid path.

(ii) Valid value edits

Associated with each field in the LFS Questionnaire portion of the record a range of valid values where validity is defined by either response code list values (e.g. values 0-3 for 'activity prior to job search') or in a few cases by considerations of reasonableness (e.g. the hours worked per week cannot exceed 126 hours).

(iii) Logical relationship edits

Within a valid path consisting of questions containing valid values, these edits ensure that the logical relationships among the questions are observed. For example, if 'going to school' is given as a reason for currently working part-time then it is required that the question addressed to educational activity indicate that the respondent is presently 'attending school'.

5. IMPUTATION METHOD

The imputation method consists of a Decision Table Analysis of each edit failure for respondents. This analysis determines the sources of data, internal to the record or external, current or past that may be used to resolve the inconsistency. An imputation value may be determined on the basis of logical relationship that is defined between two or more sequences or segments of questions and these are mostly quantitative in nature. On the other hand, if the nature of the inconsistency is such that the information on hand is not sufficient and/or leads to more than one imputation value, then recourse is made to the previous month's data or a hot deck approach.

The diagram on the next page is a schematic representation of the imputation methods. Each step in turn is described in the following section with an illustrative example.

5.1 Decision Table Analysis Using Internal Data

The edit conflict is analysed with respect to all possible questions and their responses that have bearing on the conflict. The simplest case of this kind will be where a unique response is determined on the basis of other responses. The majority of the edit conflict and imputation values in this case are of quantitative nature, i.e., either a straightforward numeric entry or a time update. As an illustration of this type consider the following example.

Example 1

A respondent who worked during the reference week, is asked, among other questions, about having more than one job, the usual hours worked at his main and other job, the actual hours worked during the reference week and the change of employer. A number of edit relationships are defined between the above questions. One of them is that if a person has one job during the reference week then the response to questions on hours of work for other job (not main job) should be blank. Consider the edit conflict where

Q11 = 2 and Q13b ≠ ( )

i.e. Q11 = Did ... have more than one job last week?

YES = 1 NO = 2

Q13 = How many hours per week does ... usually work at his

a - main job ☐☐

b - other job ☐☐

The relationship between Q11 and Q13 is defined as follows:

If	Q11=1	If	Q11=2
then	Q13(a)=01-99 or	then	Q13(a)=01-99
and	Q13(b)=01-99	and	Q13(b)=( ) i.e. blank.

In the event that the above relationship is found not to hold by the edit, then the decision table analyses is carried out. The imputation rules under each condition are spelled out and specify the value for imputation.

DECISION TABLE 2

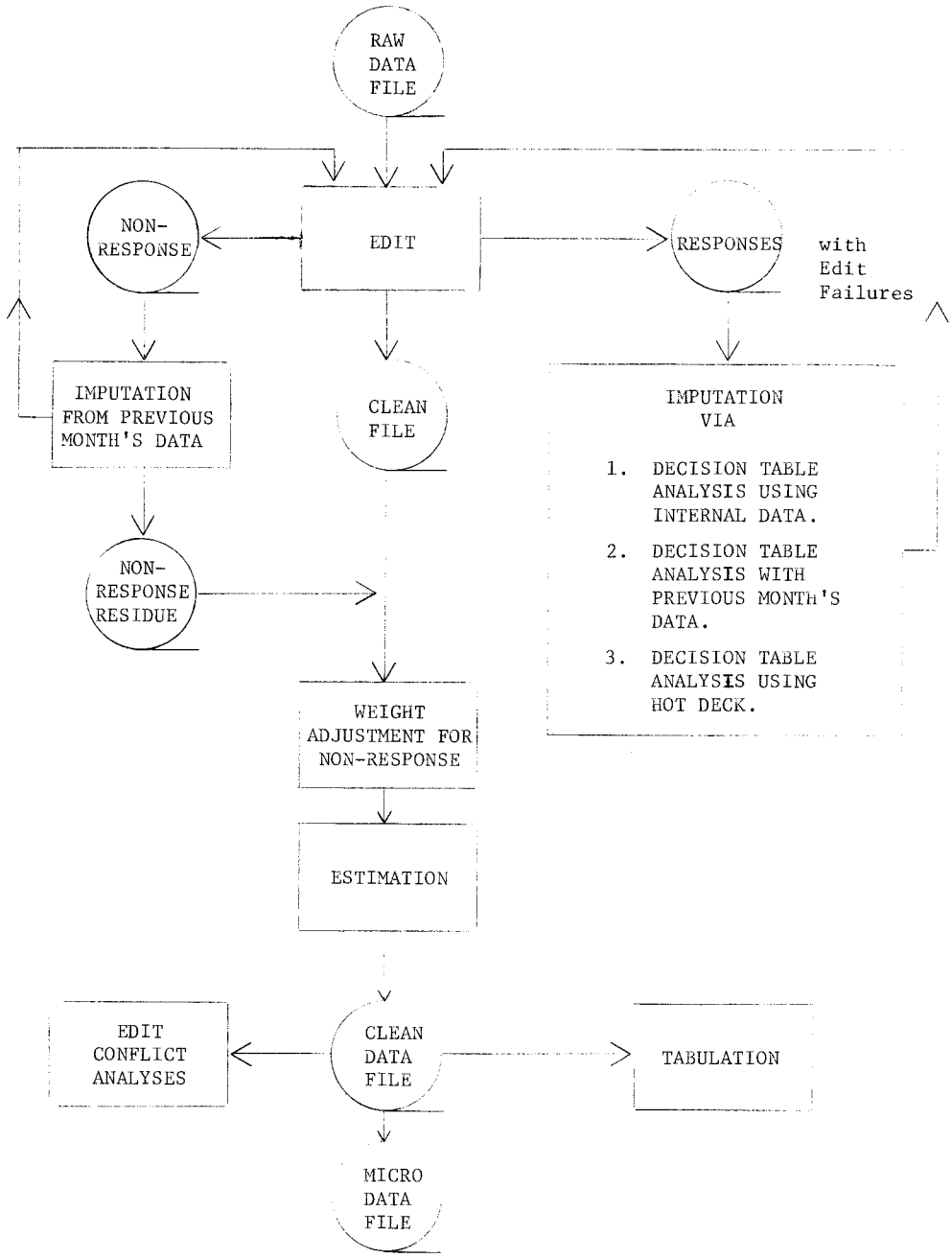
Condition Statement	Imputation Rule 1	2	3
Q11 = Had more than one job last week?	N	N	N
Q13b = Hours usually worked at other job = 01-99?	Y	Y	Y
Q18b = Hours actually worked at other job = 01-99?	Y	Y	N
Q71 = Has changed employer since last month?	Y	N	
Then Q11 should be	Y	Y	
Q12 = was this a result of changing employer last week	Y	N	
Q13b =			blank

The principle behind the above analyses is to determine a single imputation value for a given edit conflict on the basis of all relevant information. This principle is developed on the overall philosophy of minimum change to the existing responses. Let us look at each imputation rule given above.

Rule 1. If the respondent usually (and actually) works between one and ninety nine hours at the other job, not main job and has changed employer, then in fact he did have more than one job during the reference week. Therefore, in this sequence, the "no" to Q11 is an erroneous entry. The correct entry is "yes" and in this case the "no" is changed to "yes".

Rule 2. If the respondent usually (and actually) works at another job between one and ninety nine hours at the other job, in addition to his main job, then he did have more than one job during the reference week. Since he did not change employer, therefore the response to reason for more than one job is not due to change of employer. The correct value for Q11 is therefore yes and is so recorded.

DIAGRAM 1: SCHEMATIC REPRESENTATION OF IMPUTATION PROCEDURES



Rule 3. If the respondent has reported that he usually works at another job a certain number of hours but did not work at the other job during the reference week, then the number of hours reported for usual work at the other job is changed to zero.

5.2 Decision Table Analysis Using Previous Month's Data

The use of previous month's data is largely justified on the basis of high serial correlation that one might intuitively expect for certain selected characteristics, such as industry, type of worker, hours worked, etc. These are the only "external" data that are used for imputation in the LFS, if they can be so considered. Particularly, in view of the fact that in case of no change in information from the previous month, all the information is carried forward to the current month. The information provided earlier by the respondent is updated if there has been a change since the last interview. The following example is illustrative of the edit failure and the use of the previous month's data for imputation.

Example 2

Consider a respondent who reported having more than one job during the reference week, but did not report, or for whom the usual hours worked on the job, other than the main job are not reported. The edit conflict of this nature requires an imputation value that will either change the respondent to having one job or give the number of hours worked at the other job. The hours of work can be determined from either the question on actual hours of work for workers (employed) or from usual hours of work from those on lay off, etc. Failing these two sources there is no information that can be used. If the respondent has been in the survey for more than one month and has responded in the previous month, then assuming a high month to month correlation, previous month's information can be used for imputation. The following decision table analyses the edit conflict and specifies the imputation values using previous month's data.

DECISION TABLE 3

Condition Statement	Imputation Rule			
	1	2	3	4
Q18(b) = Hours actually worked at other job = 01-99?	N	Y	Y	Y
PM Q13(b) = Hours usually worked at other job = 01-99?		Y	N	N
PM Q35(b) = Hours usually worked at other job = 01-99?			Y	N
Then Q11 should be	N			
Q12, change of employer	b			
Q13(b)	b	*	**	Q18b

where PM = Previous Month
b = Blank
* = Previous Month's Q13(b)
** = Previous Month's Q35(b)

Rule 1. If the actual hours worked at the other job and the usual hours worked at the other job are both not reported, then impute a "no" response for Q11, i.e. did not have more than one job.

Rule 2. If the actual hours worked at other job are reported and from previous month's the usual hours worked are also available, then impute the usual hours from previous month to the current month.

Rule 3. This rule is similar to Rule 2 except that Q35(b) refers to persons on lay off, etc. If that value is reported, then it is imputed for current month.

Rule 4. If the actual hours worked are the only information available, then this is imputed for Q13(b) in the current month.

5.3 Decision Table Analysis Using Hot Deck

The imputation value for edit failures, for which neither the current month's data nor the data from previous month(s) could be used, is obtained by a hot deck approach.

In general, a "hot deck" is defined as all the records pertaining to the respondents in the same Primary Sampling Unit (PSU), that have the same path and age-sex group as the respondent for whom the (imputation) value is required.

The "hot deck" approach is used in relatively fewer occasions and is totally dependent on the survey methodology, specific labour market activity of the respondent and the logical sequence of the activity. The particular aspect of the survey methodology in this case is the stratification. The strata are defined in such a way that their population is homogeneous with respect to certain related characteristics. The Primary Sampling Units (PSU) within a stratum are also defined with respect to the same characteristics though to a lesser degree. The constraints of path and age-sex group are not design-dependent but it is assumed that given homogenity of population in a stratum the correlation between persons of same age-sex group and path in an area defined as above is fairly high. From all available records that are defined to constitute a hot deck, the system chooses the first record that meets the criterion. This method of record selection is considered a close approximation of random selection, since the records are processed upon arrival and there is no particular order of their receipt. It should be noted here that the composition of hot deck is dependent on the edit failures and not on the respondent or the household. Therefore, it is quite conceivable that if two imputation values are to be searched for a given respondent, then two different hot decks could be defined. Furthermore, if several members of the household required imputation for different reasons altogether different hot decks may be defined for each case. In certain instances additional conditions are added to the general definition of "hot deck". This is illustrated in example 3 given below. On the other hand, if no record meets the criterion of a particular hot deck, the overall constraint is relaxed. For example, if age-sex group was defined to be males in the 25-34 age

group, and no record met this criterion, then age-group will be expanded to 25-39 and so on.

Example 3

The following example is of a respondent who has a job but he was not at work during the reference week and his reason for not being at work was neither lay off nor a future start. The question completion sequence is assumed to be as follows:

Q31	= 1	last week had a job to start at a definite date in the future
Q32	= 2	will start in two weeks
Q33	= 1	was absent due to illness
Q34	= 1	had more than one job last week
Q35a	= 20	worked 20 hours at main job last week
Q35b	= 9	worked 9 hours at other job last week
Q36	= ( )	no reason for working less than 30 hours
Q37	= 1	was absent from job for one week
..		
Q80	= 2	last week didn't attend school.

In the above example, the conflict lies with edit conditions on the number of hours worked. The edit condition is defined that if a respondent has worked less than a total of 30 hours in the reference week, then a reason should be provided for it. In case of hours worked exceeding 30 hours no reasons are asked. In other words,

If Q35a + Q35b < 30 then Q36 = 1-6
and if Q35a + Q35b ≥ 30 then Q35 = blank,

otherwise it is considered a conflict, such as the case above where

Q35a + Q35b = 29 (i.e. < 30) and Q36 = 1-6.

The exact method of resolving this conflict is given in the following decision table.

DECISION TABLE 4

Condition Statement	Imputation Rule				
	1	2	3	4	5
Q35a + Q35b ≥ 30	Y	N	N	N	N
Q35a + Q35b < 30		Y	Y	Y	Y
Q36 = 1, 6 for previous month		Y	N	N	N
Q14 = 1, 6 for previous month			Y	N	N
Q80 = 1				Y	N
Imputation for Q36 equals in the current month	b	*	**	3	HD

where b = Blank
* = Previous Month's Q36
** = Previous Month's Q14
HD = Hot Deck

Discussion of Imputation Rules

1. If the number of hours worked are equal to or greater than 30, then question 36 for the current month is coded blank.
2. If the hours worked are less than 30, then the response from the previous month is looked up for Q36 and whatever that value may be, it is imputed for this month.
3. If the response from previous month's Q36 is not available but from Q14 is available, then this value is imputed for the current month. This situation arises for respondents who worked in the reference week when selected in the sample but subsequently have not been working in the reference week or equivalently did not work during the reference week for this month but did in the previous month.
4. This rule applies, that when other information on the questionnaire can be used such as question 80, attendance of school, etc. during the reference week. The assumption is that although there is no definite proof, the reason for working less than 30 hours most probably is attending school. Obviously this rule will not apply to this respondent, since he has told us that he was not going to school.
5. This rule calls for finding an imputation value using hot deck. The hot deck is defined as the same PSU, age-sex group and path that has question 36 ≠ blank. Suppose that the record that satisfies the conditions in the hot deck has Q36 = 4 i.e. the respondent could only find part-time work. The conflict is thus resolved by imputing a code 4 for Q36.

5.4 Imputation for Complete Non-Response

A household is considered to be non-respondent if it could not be interviewed. The reason for non-interview could be refusal on the part of the household or that the household was temporarily absent during the interview week or not at home at the times when the interviewer made a call and so on. The non-respondent households for a given month are of two kinds: those for which previous data is available and the remainder for whom no data is available. The former are imputed for by transferring the past month's data to the current month. The information is carried forward to the current month with the necessary updating of the time dependent characteristics. For example, if a person, in the previous month had been looking for a job for 16 weeks, then, in the current month, the number of weeks are updated to 20 weeks.

The remainder of the non-respondent households are those for whom it is the first month in the sample or more but have not responded. The imputation of all such households is carried out by distributing the sampling weight of the non-respondents to the respondent households. The net effect of this procedure is that an average value of respondents in a given area is imputed for non-respondents in that area. This adjustment is done for each type of area within a PSU.

REFERENCE

1. Statistics Canada, Methodology of the Canadian Labour Force Survey, Catalogue 71-526, Occasional.