A REPORT ON THE APPLICATION OF A SYSTEMATIC METHOD OF AUTOMATIC EDIT AND IMPUTATION TO THE 1976 CANADIAN CENSUS

Christopher J. Hill, Statistics Canada

1. INTRODUCTION

This paper is a discussion of the application of a Systematic Method of Automatic Edit and Imputation to the 1976 Canadian Census.

The following presentation consists of a statement of the rationale for the edit and imputation of the Census and a brief non-technical description of the methodology. A mathematical description and a computer systems description are given in articles by Fellegi and Holt[(1)]and Graves[(2)]. An evaluation is then given of the method with a final section suggesting directions for further work on the development of edit and imputation methodologies arising from the experience of the application to the 1976 Canadian Census.

2. THE RATIONALE FOR THE EDIT AND IMPUTATION OF THE CENSUS DATA

The terms 'edit and imputation' (E&I) as used here in reference to the Census are twin aspects of a single operation. 'Edit' refers to the detection of an error; 'imputation' to the correction of an error. Imputation as the correction of an error is taken to mean any modification of the data that produces a record that will pass the edits, other than by reference back to the source of the data to elicit a 'true' response. This operation of edit and imputation is undertaken with the intention of minimizing the errors in the data at the micro-level.

The reason for imputing, rather than making a correction attempting to obtain a 'true' value, is that after a certain stage in the operation it becomes costly, if not impossible, to retrace one's steps. The choice at this stage is either to edit and impute the data or to publish data that include unspecified or erroneous information. There are three reasons why edit and imputation is undertaken:

(1) To obtain the required estimates, adjustments must be made for errors at either the macro or the micro level. Correction (by edit and imputation) at the micro-level can make maximum use of the available information and in principle achieve the best estimate.

(2) Subsequent operations in the Census, for example, the formation of families, would be much more complicated, if not impossible, with incomplete and inconsistent data.

(3) Consistent official estimates are essential as a service to the users both outside and within Statistics Canada. Few users will wish to take responsibility for adjusting the estimates, and difficulties may arise as a result of differing unofficial estimates.

3. THE METHODOLOGY AND ITS IMPLEMENTATION

3.1 The Methodological Objectives

Fellegi and Holt state three objectives for the methodology underlying the edit and imputation system:

(1) As much as possible of the original data should be retained by changing the minimum number of fields in a given "dirty" record in order to produce a "clean" record.

(2) The data after imputation should retain as far as possible the distributional properties of the clean records.

(3) The imputation action should arise directly out of the edit rules.

These objectives are clearly aimed at ensuring data quality; their validity will be discussed below in the section on evaluation. The third objective is a practical consideration as it serves to greatly simplify the operation of defining imputation.

3.2 Implementation of These Objectives

The initial attempt at the implementation of the methodology was by a system that consisted of a method to analyze the edit rules and the edit and imputation system that operates on the data. The first stage in the edit and imputation operation is the analysis of the edit rules. This stage consists of two steps. The edits are written in a conflict form. They may be either within-person edits or between-person edits. The edit rules are then analyzed and any inconsistencies, conflicts or redundancies are identified. The final output of this stage is a minimum set of edit rules (explicit rules) and a set of implied edit rules, that are generated from the minimum set. These two sets combined comprise the complete set.

The edit and imputation of the data can now be undertaken. First, the edit that defines which rules have failed for each record must be identified. Second, the fields to impute must be selected. This involves two stages:

- the identification of which field(s) represent(s) the minimum number of field(s) that need to be changed to ensure a clean record; and
- the selection at random from among alternatives if there is more that one minimal set. The information that existed in the fields selected for imputation is now ignored and will in no way influence the imputation action.

There are two stages of imputation, known as primary and secondary imputation. Primary imputation is a method by which one donor record gives a 'dirty record' all the values necessary to complete the imputation. To do this the donor must match the 'dirty record' for those fields that will not be changed, and are linked by an edit rule to the fields to be imputed. These conditions ensure that a new record is clean. A donor record is found by selecting at random an acceptable record from a 'hot-deck' of about 2,000 records. If no acceptable record is found, the search continues by the method of secondary imputation.

Secondary imputation is a method of field-by-field hot-deck imputation. The crucial condition for accepting a donor is not a perfect match which has already proved impossible, but rather that the imputed record will pass the edit rules involving fields left unchanged or previously imputed. Once a field is imputed it is incorporated into the record for the search to continue so as to impute the next field.

3.3 Modifications and Enhancements Consistent with the Original Methodology

As a result of experience in attempting to apply the system, various modifications and enhancements were introduced. Some that are consistent with the methodology are described here.

Auxiliary constraints, first suggested in [1], are fields used in matching during the search for a donor record irrespective of whether or not they are required as a matching condition to ensure a clean record. They are used in both primary and secondary imputation. Fields used as auxiliary constraints will normally be those highly correlated with the fields to be imputed.

During early testing excessive matching conditions forced a large number of records to have to go to secondary imputation. In the original version of the system a match was made with every field linked to the fields to be imputed by edit rules. However, because two fields are linked by edit rules, it does not necessarily mean that the value in the field to remain unchanged restricts the acceptable values in the field to be imputed. If, therefore, the datum in a linked field does not restrict the possible values in a field to be imputed, the linked field is not used in matching. This process was termed data dependent decoupling.

A stratification system was developed to partition the data into subsets that shared a common set of edit rules and that manifested a degree of homogeneity beyond that of sharing edit rules. Edit and imputation is then undertaken independently within each stratum.

In a sense the Census represented three if not four surveys rolled into one and part of the complexity of attempting to edit it lies in this multiple nature. The difficulty lies in the interrelationship between person, family and household data. At the start of the operations the number of persons in households has been frozen. There is, of course, variation in household size. There is now a choice between treating the person or the household as the editable unit. This problem, which was not addressed by Fellegi and Holt, represented a major political issue. The methodology is based on a Cartesian data space which in a specific case, i.e., a household of a certain size, has a fixed number of dimensions. It was not possible to have sets of edit rules that addressed spaces of different dimensions, because each rule spans all dimensions of the space. Therefore, if there are to be edit rules between persons, each size of household requires a unique set of edit rules. Single unit editing is a method in which the person is the editable unit. This means there can be no edit rules between persons. Multiple unit editing is a method in which the household is the editable unit, and allows edit rules between persons.

In 1976, the latter method was used for editing the 100% data in private households principally because of the need to establish clean family data. Single unit editing was used to edit most of the persons in collective dwellings, the 13th person onwards in very large households and all sample data.

3.4 Modifications and Enhancements Inconsistent with the Original Methodology

In developing the Census system two features were included that conflicted with the original objective of changing the minimum number of data fields. These two features were both systems external to the CAN-EDIT system but utilized a specific property of that system to achieve their effect. They were: (1) a"derive" system used prior to edit and imputation and (2) a hierarchical edit and imputation structure. The Fellegi-Holt methodology specified that the amount of change in the observed data should be minimized. By implication all fields are equal candidates for change. The system recognized that there were control variables fixed prior to editing and that the system should include the possibility of distinguishing between 'imputable' and 'non-imputable' fields.

The"derive"system creates an environment within which additional variables may be derived for the edit and imputation operation. One use of this function was the deriving of a variable to force imputation in conflict with the original objectives.

The derived variable was frozen as a nonimputable variable. This meant that where an edit involved this field and other fields, some of the other fields were forced to change. This was used to force a specific imputation outcome. In general, this meant changing more than the minimum number of fields.

Hierarchical editing is a system of editing in which one set of fields is edited, imputed and frozen before another set of fields is edited.

There exists at least one edit rule linking the two sets. If there are no rules linking the two sets the order is irrelevant. If, however, there are linking rules, freezing some fields in an earlier hierarchy may force more than the minimum change in the record as a whole. The principle of minimum change only applies to a single hierarchy.

In 1976, there were two main hierarchies, one for the 100% data and one for the sample data. This structure clearly had implications for the sample questionnaire only, primarily in relating to the age question. Age was frozen in the first hierarchy and may have been inconsistent with the data on education, labour force status and mobility status. In practice, such inconsistencies were rare and the effect on the data was negligible. An additional minor hierarchy was used for questions within filters in the sample data.

4. AN EVALUATION OF THE EDIT AND IMPUTATION METHODOLOGY

4.1 Introduction

The method may be evaluated as an instrument in allowing the successful edit and imputation of the data and objectively by an external evaluation against a source of true data. A project is underway to achieve the latter. The discussion here, however, is a consideration of the system as an instrument for producing a clean data base.

4.2 The Evaluation of the Method as an Instrument for the Edit and Imputation of the Data

4.2.1 The Scope of the Method

In developing a generalized edit and imputation system it was necessary to limit the scope of the types of data that it could handle. As indicated by Fellegi and Holt, the methodology addressed itself primarily to coded or qualitative data. Quantitative fields can, of course, be treated as if they were qualitative variables and, therefore, be handled in the same system. There are, however, two important objections to doing this:

(1) the loss of information in throwing away the metric; and
(2) the potentially vast number of edit rules that may be generated in attempting to treat arithmetic rules as logical rules between categories.

Despite these objections, the system was applied in the Census to records that contained a mixture of quantitative and qualitative data. This was justified insofar as the variables were predominantly qualitative and the edits applied to the quantitative variables were of a limited nature. However, as the Census was attempting to edit variables outside the scope for which the editing system was designed, the results were not totally satisfactory.

The only quantitative variable in the 100% data was date of birth or, by implication, age. Date of birth was defined by 3 variables: decade, year, and month of birth, this last being more correctly the two periods January to May, June to December. Each of these taken separately could be used as a qualitative variable and indeed was so treated. There were two main problems:

(1) A crucial age barrier occurs at age 15. The sample questions were only to be answered by persons at or over this age. Also certain conditions were only allowable at or above this age, e.g., Head of household or Married. The problem was that after edit and imputation there were more than the expected numbers of certain groups of persons close to the 15 year age boundary, in particular widowed or divorced persons. The only consolation was that the problem was greatly reduced when compared with the 1971 data.
(2) It was impossible to write edits to ensure reasonable age spacing between parents and children. The number of edits required to ensure a 15 year minimum difference was astronomical. The decision was therefore:

(i) to limit such edits to age differences between the Head and Spouse and their children, (the main group of edits this excluded was edits between the Head and his parents);
(ii) to use only decade of birth in the edits;
(iii) to ensure that at least one parent was born in an earlier decade than all the children.

The net result of this was only partially successful in removing strange data. A successful solution to this problem awaits the development of a methodology that can be implemented as a system that will not only edit and impute quantitative data but also quantitative data in combination with complex qualitative data.

4.2.2 Finiteness

The population of Canada is 23 million. The number of households is 7 million. The complete data space representing households has very many more cells than the total number of households. For households of size 'n', this space contains approximately $2,000^n$ cells.

The number of edit rules required to partition this space is also potentially very large. A particular between-person edit condition that could apply between most persons in the household, in almost all positions, would have generated 100 million edit rules. A tabulation of the data indicated that in fact there were only 1700 persons in Canada who could potentially fail these rules.

The total number of edit rules is a function of household size and the set of edit conditions to be applied. A realistic utilization of computer resources set a limit of 2048 upon the total number of edit rules. This limit was implemented by restricting multiple unit editing to households of 12 or less, or the first 12 persons in large households; and by excluding certain types of conditions from the set of edit rules. A special 'clean-up' programme was used to edit and impute these residual problems.

There are also data limitations in trying to push the method too far. The imputation was by a hot-deck method. In attempting to edit and impute large households, the system came up against the data limit that the number of available records for the hot-deck had become very small. With very large households, a point is reached at which the operation is very costly, the number of records is very small and the quality of the imputation is much reduced by the small hot-deck size. The finite limitations of the system are probably a minor constraint upon the effectiveness of the method given the finite nature of the data.

4.2.3 The Methodological Basis

The three criteria set out by Fellegi and Holt were outlined above in the description of the methodology and will now be assessed.

Changing the Fewest Possible Items of Data

The principle of changing the fewest possible data items (fields) is considered by Fellegi and Holt to be of overwhelming importance. This position is more than justified as a reaction against the enthusiastic over-correction of data that has been known to occur. Their formulation, however, is a specific case of a general principle that data modification should be kept to a minimum. The problem is that the number of fields is somewhat arbitrary. The number of fields covering the same information may be modified by changes in the questionnaire or in its data capture. A simple, easily defined concept may be reliably captured by one question, whereas a number of questions may be used to define a single, potentially ambiguous concept. On the other hand, one cannot pretend to start counting concepts as if they had the same concrete existence as a question.

Alternative formulations of the principle of minimum change may be changing a weighted minimum number of data items, or moving the minimum distance in some conceptual space.

The justification for using the first alternative may relate to the conceptual intentions of the questionnaire or to the reliability of each field. This may be illustrated with reference to the questions on education.

One education question asks for the respondent's highest school grade; three other questions ask for the respondent's post-secondary education and qualifications. By 'post-secondary', the Census had intended to refer to education of an advanced nature requiring a certain minimum schooling as an entrance requirement. Unfortunately, a surprisingly high proportion of respondents interpreted this as any education obtained after leaving school. Typically, the respondents making this error were giving two wrong answers consistent with each other but in conflict with the highest grade to be incorrectly up-graded. It was finally decided that the best strategy was to modify certain rules to avoid the risk of serious distortion of the highest grade response by imputation.

Imputation Rules Derived from Corresponding Edit Rules

Among the subject-matter oriented benefits of the system listed by Fellegi and Holt are:

'(1) Given the availability of a generalized edit and imputation system, subject-matter experts can readily implement a variety of experimental edit specifications whose impact can therefore be evaluated without extra effort involving systems development. This is particularly important given the generally heuristic nature of edit specifications.
(2) Only the edits have to be specified in advance, since the imputations are derived from the edits themselves for each current record. This represents a major simplification for subject-matter experts of the workload of specifying a complete edit and imputation system'.

The first of these two benefits, 'a parametric' approach to editing, was clearly an advantage. The second of these two benefits, however, is not necessarily an unqualified advantage. The fact that the imputation actions arise directly out of the edit rules, precludes the possibility of any error-specific data correction. The methodology facilitates experimentation with the edit rules but removes any control the user may otherwise have over the imputation. A means of utilizing a specific feature in the system was however identified that returned some control over imputation. This was the use of an unimputable derived variable.

One type of error that justified the use of a derived variable was erroneous responses associated with common-law relationships. The intention of the Census was that consensual unions should be treated the same way as legal unions, hence allowing the identification of families. However, the frequent response pattern in these cases was to give the legal marital status, i.e., 'not-married', together with the de facto relationship to head, either spouse or common-law spouse. A typical pattern of responses was:

Person 1	Head of Household	Divorced
Person 2	Spouse of Head	Single

In such a case, the minimum change of data fields was to change the relationship of person 2 to 'head' rather than the marital status of both individuals. It was decided that the best strategy was to force the data using an uneditable derived variable. This was given a value 'Spouse Confirmed' whenever cases such as the above occurred. Then the responses were forced into the pattern:

Person 1	Head of Household	Married
Person 2	Spouse of Head	Married

During the application of the E&I System in the 1976 Census, it became evident that there were other situations in which control over the imputation could have achieved more appropriate outcomes. Certain types of response errors caused edit failures for which a clearly identified correction procedure could be specified. The principal examples of these were:

- The incorrect coding of relationship to head of household by reversing the relationship, i.e., son or daughter of the head was coded 'Father' or 'Mother' of the head.
- Incorrect coding of relationship to head where there are children in the household, the head and spouse being coded as 'Father of head' and 'Mother of head'.

The main problem with the data in both these two examples is that they are cases of infrequent errors on common conditions being mis-allocated to infrequent conditions. In both examples, deterministic editing would have been appropriate.

The particular problems mentioned above, which may be remedied by systematic corrections, must, however, be weighed against the advantages of the method. There are very many rules to which the data should conform; each failed by a small number of records. Separate imputation rules for each of these would have required a much more complicated system.

The system created a framework within which alternative edit specifications could be reviewed, evaluated and modified very easily. It required a certain amount of work on the part of subject matter personnel to familiarize themselves with the system and its language. Once this had been achieved however, considerable progress could be made in understanding the problems in the data and refining the edits.

One incident illustrated the flexibility of the system. A tabulation during processing indicated that a rule had been omitted from one particular set of rules. This omission was corrected within 48 hours. The system naturally cannot ensure that the user has included a complete set of edits, but it can ensure that the existing set is clean and consistent. It took much longer to make corrections to tailor-made programmes with the risk always that a correction introduced a new error.

Retaining the Distributional Properties of the Clean Data

In the absence of any additional information, retaining the distributional properties of the clean data is the most appropriate strategy to take during imputation. The effectiveness of the system to achieve this was increased by the use of auxiliary constraints, that is, fields used as matching criteria in the hot-deck search by reason of their correlation with the field to be imputed, irrespective of any links by edit rules. There were, however, situations in which the dirty records were clearly drawn from a distribution very different from that of the clean records.

There were two main reasons for this type of problem arising:

(1) Certain sub-groups of the population have difficulty selecting the correct response and are therefore more likely to fail to respond;
(2) Many questions include a 'null', or

'none' category. No device has yet been invented to prevent the relatively high non-response from persons who should have used one of these codes.

An example of this second type of non-response occurred with answers to Labour Force Status. There was a tendency for non-response to increase as the proportion of persons not in the Labour Force increases. **This suggests that there is a tendency for non-re-**spondents to be drawn more heaviıy from **the non-participating population. It is possible to control imputation with re-**spect to tne variables in the Census,but **not for any relationship beyond these.**

An evaluation of this problem is currently being undertaken. Some consideration has also been given to possible enhancements to the methodology to adjust for this differential non-response. However, in order to utilize such enhancements, external information is needed to estimate the differential non-response rates with respect to the target variable.

5. Some Conclusions Concerning Current System

The edit and imputation system developed from the methodology outlined by Fellegi and Holt was designed to be a generalized system. The major motive behind the development, however, was the needs of the Census as manifested in problems experienced during the edit and imputation of the 1971 Census. It was an attempt to bring order to a complex and potentially chaotic operation. The system was very successful in achieving this objective. The data were available relatively earlier than the 1971 data. There has been no need for post edit fixes. The residual problems in the data in general seem less serious than those found in 1971. There is more knowledge about data problems and means of correcting them. This system has in fact allowed a more critical analysis of the data and made it possible to identify problem areas such as'systematic response error and non-response bias. Future work can be concentrated on a better handling of these problems within a controlled structure.

6. Direction for Future Developments in System

The following are some of the areas that need to be considered:

(1) A means for handling systematic errors that can be integrated with the existing system needs to be found.
(2) Alternatives to the principle of changing the minimum number of fields need to be investigated. Such alternatives may prove of limited value compared with the handling of systematic errors.
(3) Strategies for the handling of non-response to adjust for the differences between the responding and non-responding population should be considered.
(4) An experimental system for arithmetic edit and imputation is already being developed. The integration into this system of means of handling both quantitative and qualitative variables is among the possible long-term plans.

Errors cannot be avoided no matter how carefully the survey is designed. The appropriateness of the edit and imputation strategy lies in its ability to recover the 'true' values. To achieve this, there is a need for more empirical evidence concerning the nature of errors in the data.

REFERENCES

[1] I.P. Fellegi and D. Holt. A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, March 1976. Volume 71, Number 353.

[2] R.B. Graves, Can-Edit. A Generalized Edit and Imputation System in a Data Base Environment. A report to the working party on electronic data processing, Conference of European Statisticians. (CES/WP.9/142). Feb. 1976.