THE APPLICATION OF A SYSTEMATIC METHOD OF AUTOMATIC EDIT AND IMPUTATION TO THE 1976 CANADIAN CENSUS OF POPULATION AND HOUSING[1]

C.J. Hill[2]

I.P. Fellegi and D. Holt proposed a systematic approach to automatic edit and imputation. An implementation of this proposal was a Generalized Edit and Imputation System by the Hot-Deck Approach, that was utilized in the edit and imputation of the 1976 Canadian Census of Population and Housing. This paper discusses that application, evaluating the strengths and weaknesses of the methodology with some empirical evidence. The system will be considered in relation to the general issues of the edit and imputation of survey data. Some directions for future developments will also be considered.

1. INTRODUCTION

This paper is a discussion of the application of a Systematic Method of Automatic Edit and Imputation originally developed by I.P. Fellegi and D. Holt [1] to the 1976 Canadian Census of Population and Housing. The implementation of this methodology as a computer system within Statistics Canada is the system known as 'CAN-EDIT'. This was described by Graves [2]. The Can-Edit system, in turn, became a component of the "Census Edit and Imputation Processing System" which included several other custom-built modules [3]. Some of these modules handled certain special edit and imputation problems. Others operated in conjunction with the CAN-EDIT system and addressed methodological issues not covered by Fellegi and Holt. Some discussion of the methodology of these modules is included here in that they were essential to the application of the Fellegi-Holt method.

---

[1] Adapted from a paper presented at the Annual Meeting of the American Statistical Association, August 14-17, 1978, San Diego, California, U.S.A.

[2] C.J. Hill, Census Survey Methods Division, Statistics Canada.

The census is a multi-purpose survey consisting of both population and housing questions. The housing questions in 1976 were primarily concerned with identifying the type and tenure of the dwelling. The population questions were divided into two parts, a basic set of questions asked of all persons, and a set of sample questions asked of persons 15 years of age and over in 1/3 of all private households, and all collective dwellings. The basic questions were demographic questions on age, sex and marital status, a question on relationship to head, and one on mother tongue. The sample questions were on education, labour force status and mobility status. The 'CAN-EDIT' system was used in the edit and imputation of most of the variables. The only variables not handled by this system were mother-tongue and mobility status.

This paper presents the rationale for the edit and imputation of the Census and a brief non-technical description of the methodology in sections 2 and 3. An evaluation of the method is then given in section 4,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. Edit can be considered separately from imputation in that it may be used to initiate a corrective action involving a return to an earlier state in the processing. Editing may also be undertaken merely to flag erroneous records. 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.

Among others, the following three important reasons influenced the undertaking of edit and imputation in the 1976 Census.

(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. In certain cases, the number of invalid records would increase considerably.

(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 Methodology 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 The Implementation of These Objectives

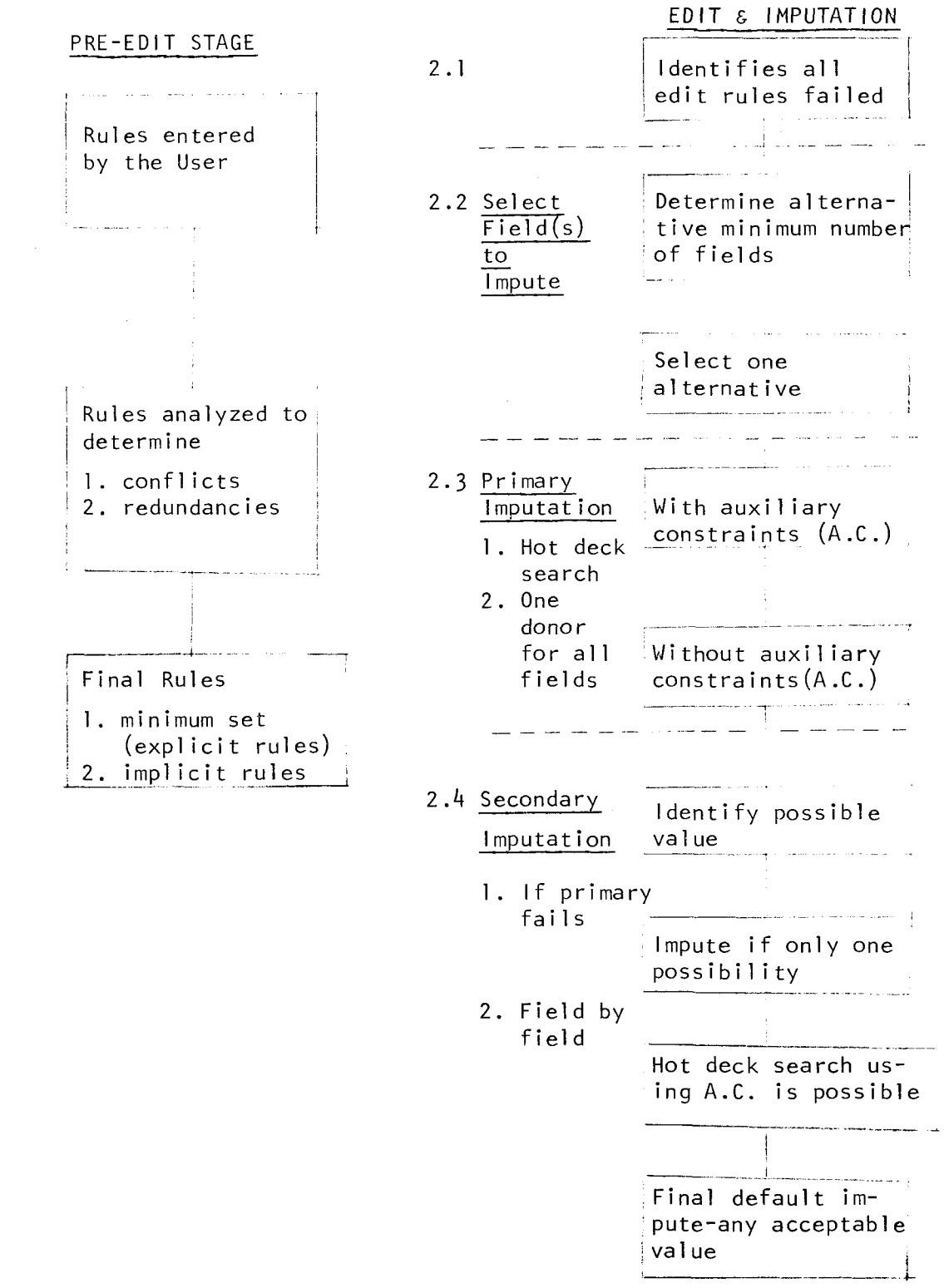
The initial attempt at the implementation of the methodology was by a system that consisted of two basic sub-systems:

(1) A system to analyze the edit rules.
(2) The edit and imputation system that operates on the data.

These operations are shown in Diagram 1.

Diagram 1

A Flowchart of 'CAN-EDIT' Processes to be Undertaken for Each Stratum*



* Stratification and Auxiliary Constraints are explained in Section 3.3.

(1) The System to Analyze the Edit Rules

The first stage in the edit and imputation operation is the analysis of the edit rules. This stage consists of the following steps:

The edits are written in a conflict form. They may be either within-person edits or between-person edits.

> An example of a within-person edit is:
> 'It is a conflict if the third person in the household is married and is less than 15 years of age'.

> An example of a between-person edit is:
> 'It is a conflict if the sixth person in the household is the parent of the head of the household and male and the ninth person in the household is the parent of the head of the household and male'.

It is important to note that one concept requires many edit rules. If, for example, an edit is required to exclude the possibility that the head of household has two parents of the same sex, edit rules have to be written between all possible pairs of persons. This essential feature creates some limitations to the system that will be discussed in a later section.

The edit rules are then analyzed and the output defines:

i) Any inconsistencies or conflicts between the rules.
ii) Any redundancies in the rules.

Once inconsistencies are removed, the final output is:

i) A minimum set of edit rules (explicit rules).
ii) A set of implied edit rules, that are generated from the minimum set.

These two sets combined comprise the complete set.

(2) The Edit and Imputation System

The analysis of the rules having been completed, the edit and imputation of the data can be undertaken. This operation divides into four stages:

(2.1) The edit that defines which rules have failed for each record.

(2.2) The selection of fields to impute. This has two parts:

i) the identification of which field(s) represent(s) the minimum number of field(s) that need to be changed to ensure a clean record,

ii) the selection at random from among alternatives if there is more than 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.

(2.3) 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. (Refinements on this principle will be discussed below). A donor record is found by selecting at random an acceptable record from a file of about 2,000 records. This is a form of the method of imputation known as 'hot deck' imputation. If no acceptable record is found, the search continues by the method of secondary imputation.

(2.4) Secondary imputation is a method of field-by-field hot deck imputation. In this method certain matching conditions may be applied during the search for a donor. However, the crucial condition for accepting a donor is not a perfect match which has already proved impossible, but rather that the new 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.

One important discovery that was made during the development testing of 'CAN-EDIT' is that for primary imputation only the minimum set of rules is required, whereas secondary imputation needed the complete set of rules. Failure to use the complete set could result in creating a situation in which a partially imputed record could become impossible to complete.

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 of these were consistent with the methodology, four of which are considered here. Section 3.4 will consider two modifications that conflicted with the original objectives. Two are important refinements to the principles of imputation within the 'CAN-EDIT' system. These are (1) 'Auxiliary Constraints' and (2) 'Data Dependent Decoupling'. The other two are elements of the 'Census System' that address methodological problems not covered by Fellegi and Holt. These are (3) the Stratification Sub-System and (4) the choice between single or multiple unit editing.

(1) Auxiliary Constraints

Auxiliary constraints 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. This enhancement was suggested in the paper by Fellegi and Holt.

In primary imputation, they have to be used as a complete set or not at all. The system was designed this way because there is no very obvious algorithm for relaxing constraints when the entire record is imputed simultaneously. In effect, therefore, primary imputation has two levels of matching, the optimum matching conditions that include auxiliary constraints and a degraded option matching on the necessary fields only.

In secondary imputation, with field by field imputation, one can attempt to match on as many fields as specified and take the best match.

(2) Data Dependent Decoupling

During a test of an early version of 'CAN-EDIT' excessive matching conditions forced a large number of records to have to go to secondary imputation. An analysis of the problem indicated that the matching conditions in the search for a donor were too restrictive.

In the original version, 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. An example of this is in the field "relationship to head", with reference to the previously mentioned rule preventing two parents of the head with the same sex. Clearly, if there is a person in the household coded head's parent and male, this places a restriction

upon imputing the code parent to another male. If on the other hand there is no such person, there need no longer be this restriction.

(3) The Stratification System

The function of the stratification system was to partition the data into subsets that (1) shared a common set of edit rules and (2) manifested a degree of homogeneity beyond that of sharing edit rules. Edit and Imputation is then undertaken independently within each stratum.

The control variables[1], document type and collective dwelling type were used for this purpose, for the 100% data, together with a variable defined in terms of the mix of persons in the household. Age, sex and collective dwelling type were used to stratify the sample data.

A full appreciation of the nature of stratification needs to be considered in conjunction with the question of single and multiple unit editing, since one of the dimensions of stratification for multiple unit editing was the number of persons in the household.

(4) Single or Multiple Unit Editing

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 dwelling data stands alone and presented only minor problems. The difficulty lies in the interrelationship between person, family and household data. At the start of the editing operations the number of persons (the low level unit) in households (the high level unit) has been frozen. There is, of course, variation in household size.

---

[1] The operation prior to edit and imputation determined whether a household was a private or a collective dwelling, occupied or unoccupied and whether or not it was in the sample. It also ensured that all collective dwellings had an identified type, e.g. hospital, orphanage, hotel. This information was frozen as the control variables document type and collective dwelling type.

The family at this stage has yet to be defined. 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 practical issue when integrating 'CAN-EDIT' into the 'Census System'. 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 the method of editing in which the person is the editable unit. This means there can be no edit rules between persons.

Multiple unit editing is a method of editing in which the household is the editable unit. This method allows edit rules between persons. However, this is achieved at certain cost.

i) The data have to be stratified by size of household.
ii) The potential size of the editable unit becomes very large.
iii) There is a cut-off point beyond which it is totally unrealistic to take multiple unit editing which means there must be single unit editing for residual persons in large households.

In 1976, multiple unit editing 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, set out by Fellegi and Holt, 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 'CAN-EDIT' system for very good reasons 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: This piece of software is a semi-generalized system that creates an environment within which additional variables may be derived for the edit and imputation operation.

i) To combine two or more fields into one.
ii) To derive a variable for stratification.
iii) To create class values of a variable.
iv) As a means of forcing an imputation action.

It is this last function that is important to consider here as it conflicts with original objectives. The derived variable was frozen as an non-imputable 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. This is explained in detail in section 4.3.3.

Hierarchical Editing: 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, and in which 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 only had implications for the sample questionnaire, primarily in relation 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 findings of this project will be reported in a census publication [4]. 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

The following points will be considered in evaluating the generalized system as a means of achieving a successful edit and imputation operation.

(1) The methodological scope of the system, i.e. the range of types of variable and edit conditions the system is designed to handle.
(2) Finiteness, i.e. the practical limits to which the system conforms.
(3) The appropriateness of the three objectives outlined by Fellegi and Holt.

4.2.1 The Scope of the Method

In their paper, Fellegi and Holt write "At the beginning, let us restrict ourselves to records containing only qualitative (coded) data, i.e. data which are not subject to a meaningful metric".

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 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.
(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 very large as this would have required an edit rule for each individual age difference. 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. (It is theoretically possible for a step-parent to be younger than an adult child).

The application of these rules removed some, but not all of the erroneous 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 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 $(2000)^n$ 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

Editing is an essentially very straightforward operation and is passive in relation to the final data. The only problem presented by editing is to ensure that the edit rules are clean and consistent. The issue to be discussed here is the methodological basis of the imputation action. The three criteria set out by Fellegi and Holt were outlined above in the description of the methodology and will now be assessed.

4.2.3.1 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 a somewhat arbitrary count. 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.

This problem is implicitly recognized by Fellegi and Holt in the suggestion they made that weights could be attached to fields in relation to their reliability. This suggestion was not implemented for use in the system applied to the 1976 Census. However, careful analysis is required before any alternatives are introduced.

Alternative formulations of the principle of minimum change may be considered.

(1) Changing the fewest possible data items.
(2) Changing a weighted minimum number of data items.
(3) Moving the minimum distance in some conceptual space.

The first of these formulations is given by Fellegi and Holt and the second one is an alternative they suggest. The justification for using the second alternative may, however, relate to the conceptual intentions of the questionnaire rather than 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 that was too low for entry into post-secondary education. In this case the minimum change was causing 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.

4.2.3.2 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 controls the flexibility available to the user. It 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 a non-imputable derive variable referred to in Section 3.4. The variable used to force changes in households with common-law relationships is discussed below.

During the application of this method in the 1976 Census, it became evident that there were 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:

(1) The incorrect coding of relationship to head by reversing the relationship, i.e. son or daughter of the head was coded 'Father' or 'Mother' of the head. A proportion of these errors particularly for younger age groups were correctly imputed.

(2) Incorrect coding of relationship to head where there are children in the household, the head and spouse being coded as 'Father' and 'Mother'. The problem here was that the first person was changed to head but the second person remained as 'Father' or 'Mother'.

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.

(3) One type of error that created special problems 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 patterns of response 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 to head of person 2 rather than the marital status of both persons. Problems of this nature were identified during the test Census. 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

There remained a residual problem as to how to edit children of the common-law partner in these cases. Certain distortions in the data were considered too critical to be left uncorrected. Additional strategies for correction were therefore adopted, either prior to the application of the Fellegi-Holt methodology as with common-law spouses or in certain cases as a clean-up afterwards. Evaluation is currently being undertaken to assess the correctness of the actions taken during the entire edit and imputation.

These particular problems 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 first of the two benefits, 'the parametric approach' referred to above must also be weighed against the loss of flexibility in specifying the imputation. However, for these edits even a very imperfect imputation action would have had a negligible impact on the final data.

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. Tabulations were run on the data at stages during the production. A tabulation indicated that a rule had been omitted from one particular set of rules. The erroneous condition detected was a rare condition that had not occurred in the test data, but was a condition that would never the less cause difficulties in the subsequent family formation programme. 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 always the risk that a correction introduced a new error.

4.2.3.3 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. These situations are equally true for any sub-sets of the population defined by other fields in the records. The inadequacy of the imputation as reflected in the final data in this case is a function of the difference between the two distributions and the proportion of dirty 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 fall into this group.

This problem is illustrated by Table 1. This tabulates Labour Force Status defined from the unedited data. Clearly, there is 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-respondents to be drawn more heavily from the non-participating population. It is possible to control imputation with respect to the 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.

Labour Force Status Identified from the Unedited Census Weighted Data

		Status Defined				Not Defined		Total
		In the Labour Force		Not in the Labour Force		Labour Force Status		
		Count	%	Count	%	Count	%	
BOTH SEXES								
	15-19	346,243	42.39	424,653	51.99	45,872	5.72	816,768
	20-24	534,746	71.99	175,551	23.63	32,554	4.38	742,851
	25-54	2,065,994	68.46	851,709	28.22	100,203	3.32	3,017,906
	55-64	340,744	50.57	304,740	45.23	28,304	4.20	673,788
	65+	74,373	9.46	675,876	85.93	36,298	4.61	786,547
	Total	3,362,100	55.68	2,432,529	40.29	243,231	4.03	6,037,860
MALES								
	15-19	194,651	46.50	200,104	47.80	23,892	5.71	418,647
	20-24	300,829	30.33	56,673	15.13	16,979	4.53	374,841
	25-54	1,320,626	86.85	152,289	10.01	47,745	3.14	1,520,660
	55-64	231,330	70.99	82,003	25.16	12,533	3.85	325,866
	65+	52,347	15.73	264,988	79.61	15,541	4.67	332,876
	Total	2,099,783	70.64	756,057	25.43	116,690	3.93	2,972,530
FEMALES								
	15-19	151,592	38.08	224,549	56.50	21,980	5.52	398,121
	20-24	233,917	63.50	118,878	32.27	15,575	4.23	368,370
	25-54	745,368	49.78	699,420	46.71	52,458	3.50	1,497,246
	55-64	109,414	31.45	222,737	64.02	15,771	4.53	347,922
	65+	22,026	4.86	410,888	90.57	20,757	4.58	453,671
	Total	1,262,317	41.18	1,676,472	54.69	126,541	4.13	3,065,330

5. CONCLUSIONS

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 edited 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 a great deal more knowledge about data problems and means of correcting them.

This system has in fact allowed a much 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.

The following four issues are some of the key issues that need to be or are currently being addressed:

(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.

RESUME

À partir de la méthode systématique de vérification et d'imputation proposée par I.P. Fellegi et D. Holt, on a mis au point un système général de vérification et d'imputation par la méthode du hot-deck et on l'a appliqué aux données du recensement de la population et du logement de 1976. Le présent article étudie cette application de la méthode Fellegi-Holt et évalue les points forts et faibles de la méthodologie à partir de certains exemples empiriques. La présentation du système est faite dans un contexte plus vaste, celui des grands problèmes posés par la vérification et l'imputation des données d'enquête. L'auteur énumère aussi quelques avenues de développement possibles.

REFERENCES

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

[2] Graves, R.B., "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.

[3] Turner, M.J., "The Use of Data Base Technology in Large Systems". A report to the working party on electronic data processing, Conference of European Statisticians, April 1977.

[4] Quality of Date, '76 Census, Series 1: Sources of Error - Effect of Edit and Imputation Procedures on the Quality of Data from the 1976 Census of Population and Housing. Catalogue No. 99-843.