Enhancements to the 2006 Canadian Census Edit and Imputation System

Wesley Benjamin
Statistics Canada, Ottawa, ON, K1A 0T6

Abstract

The CANadian Census Edit and Imputation System (CANCEIS) will do deterministic imputation plus perform minimum change donor imputation for all variables (both numeric and categorical) in the 2006 Canadian Census. Significant enhancements have been made to CANCEIS for the 2006 Census, including the ability to perform deterministic imputation, process alphanumeric variables, do outlier detection of numeric variables and use failed records as donors. In addition, it is now easier to write compact decision logic tables and improvements have been made in the handling of numeric variables. These changes were implemented by methodologists, subject matter experts and systems developers working in a collaborative fashion. CANCEIS, or an earlier version of the software, has been used in the 1996 and 2001 Canadian Censuses, as well as the 2000 Swiss, 2000 Brazilian and 2005 Peruvian Censuses.

Keywords: edit, imputation, census, Canada, donor, deterministic

1. Introduction

For the 1996 Canadian Census of Population, a new method of imputing for non-response and inconsistencies was introduced for the demographic variables (age, sex, marital status, common-law status and relationship). It allowed, for the first time, the simultaneous minimum change donor imputation of both qualitative and quantitative variables for large edit and imputation (E&I) problems. This new method, the Nearest-neighbour Imputation Methodology (NIM), replaced a computer system based on the methodology proposed by Fellegi and Holt (1976). Many minimum change imputation systems are based on the Fellegi/Holt approach, including CANEDIT and GEIS at Statistics Canada and DISCRETE and SPEER at United States Bureau of the Census (USBC). The main difference between these two imputation methodologies is described in Bankier et al.(2001). Where Fellegi/Holt first determines a theoretical minimum number of variables to impute and then searches for a suitable donor, NIM first finds potential donors and then determines the minimum number of variables to impute given a donor. Reversing the order of these two operations creates significant computational advantages with NIM while maintaining the well-accepted Fellegi/Holt objects of minimum change and the preservation of sub-population distributions.

A more generic implementation of the NIM was developed for the 2001 Canadian Census, called the CANadian Census Edit and Imputation System (CANCEIS). This system was used to perform donor edit and imputation for approximately 40% of the census variables while the remaining 60% of donor imputation and all deterministic imputation was performed by the existing mainframe E&I system.

CANCEIS, or some form of the NIM methodology, has been used by a number of clients outside of the Canadian Census. The Survey of Household Spending, one of the larger surveys at Statistics Canada, has used CANCEIS for several years now. Switzerland and Brazil both used a prototype software based on the NIM methodology for their 2000 censuses, Peru used CANCEIS to process their entire 2005 Census and Brazil and the United Kingdom plan to use CANCEIS for their 2007 and 2011 censuses (respectively).

CANCEIS will be used to perform 100% of the E&I for the 2006 Census, including both donor and deterministic imputation. Numerous enhancements to the system have been required in order to accommodate the increase in the number of variables and types of data to be processed. This paper will review the existing CANCEIS functionality and identify these new features that were introduced.

2. CANCEIS in 2001

For the 2001 Census, CANCEIS performed the donor E&I for variables from five subject matter topics: Demography, Labour, Mobility, Place of Work and Mode of Transport. For these five topics, it applied up to 43,000 edit rules simultaneously for the entire Canadian population of approximately 30,000,000 persons. CANCEIS was processed on personal computers with all input present in ASCII text input files.

The donor E&I for the remaining subject matter topics, along with the deterministic imputation for all topics, was processed on the mainframe by SPIDER (System for Processing Instructions from Directly Entered Requirements), which was introduced for the 1981 Census. Deterministic imputation is done in *pre-derive* and *post-derive* modules which are typically run before and after donor imputation modules respectively. These *derive* modules are generally used to define universes and create stratification variables for the donor modules, derive new variables, perform deterministic imputation and perform default imputation.

2.1 A Review of CANCEIS Functionality

CANCEIS allows the user to define a custom *data dictionary* for their module which defines the necessary information for all of the variables. In 2001, users were able to define both categorical and numeric variables for simultaneous processing. Defining the data dictionary includes specifying the names of the variables, the set of valid responses for each variable, text labels associated with numeric responses for categorical variables, weights and distance measures associated with each variable, classes which group similar responses and allow for more compact edit rules and defining a set of system parameters that allow you to control the E&I process. The input files which defined this information were in text format.

Edit rules are defined within CANCEIS through the use of *Decision Logic Tables (DLTs)*. Any records which match one or more of the edit rules defined in a DLT are determined to have inconsistent data and are flagged as 'failed' in the editing process. These records will be resolved with donor imputation.

Figure 1: Example of a CANCEIS DLT in 2001.

```
@ AGE < 15                ;Y;Y;
@ MARITAL_ST = MARRIED    ;Y; ;
@ INCOME = 0              ; ;N;
```

Figure 1 presents an example of a DLT with 2 columns of edit rules and 3 rows of *conditions* (the series of logical statements beginning with an '@' symbol). The first rule states that the record should fail if someone has an age of less than 15 and is married. The second rule states that the record should fail if someone has an age of less than 15 and reported an income other than 0. In the imputation process, CANCEIS will replace data in the failing record with that from a donor record so that none of the edit rules fail.

DLTs are used to identify records which fail for having inconsistent responses. Records can also fail if they contain invalid responses. The data dictionary defines a validity set for each variable. For example, the validity set for the variable SEX would contain the two responses FEMALE = 2 and MALE = 4. Categorical variables such as SEX are stored on the data files as numbers, and the data dictionary associates labels with these numeric values. If a numeric response other than the two defined in the validity is present (often indicating a BLANK or INVALID response), it will be flagged as invalid. For categorical variables, the validity set contains the list of valid numeric codes. For numeric variables, the validity set defines the valid range of responses. For example, the validity set for the variable AGE would be defined as all discrete values in the range from 0 to 120. If a response outside this range is detected, it will be flagged as invalid. Records with invalid data or inconsistencies are flagged as failing. Records which contain no invalid data and no inconsistencies are flagged as 'passed' and will be used as donors. Failing records will have their invalid and inconsistent responses replaced in donor imputation with valid responses from the donor record which eliminate all inconsistencies.

The best potential donors for a failed record are determined by comparing the response for each variable between the failed record and the potential donor record. If there are I variables present in the module, then CANCEIS calculates the distance measure $D_{fp} = \sum D_i W_i$, where f and p indicate that this distance is between the failed and passed (donor) record. D_i is the distance measure within the range [0,1] which indicates how similar the response between the failed and passed record is for the i^{th} variable. CANCEIS offers a selection of distance functions from which the user can choose to suit each variable individually. W_i is the non-negative weight assigned by the user to the i^{th} variable. A higher weight would be assigned to variables for which it is most important to match between the failed and donor record. This weighted distance measure is summed across all *I* variables. The donor records with the smallest distance, which are called *nearest neighbours*, are retained for further analysis to determine the best *imputation actions*, which represent the failed record modified with the donor record data to pass all edit rules. Only imputation actions involving the minimum number (or near the minimum number) of changes will be considered. Two new distance measures are calculated: D_{fa}, where f and a indicate that the distance is between the failed record and the imputation action, and D_{ap}, where a and p indicate that the distance is between the imputation action and the passed (donor) record. D_{fa} represents a measure of minimum change while D_{ap} represents a measure of plausibility. These two distance measures are combined to form a new distance measure:

$$D_{fpa} = \alpha\, D_{fa} + (1- \alpha)\, D_{ap}$$

where $0 \leq \alpha \leq 1$ can be used to place more importance on imputing the minimum number of variables or creating a plausible record. One of the best imputation actions (those with the lowest D_{fpa}) will be randomly chosen as the final solution to resolve the failed record. A detailed explanation of CANCEIS' donor imputation methodology is found in Bankier et al.(2001) and Bankier (2006).

3. Enhancements to CANCEIS for 2006

The 2001 Census represented the 5th time that SPIDER was used to process the Canadian Census and the first time for CANCEIS. It was decided for the 2006 Census that processing would move from a custom Statistics Canada data base (RAPID) to SYBASE, a commercially available data base. Rather than rewrite SPIDER for this new environment, and based on the successful use of CANCEIS for the 2001 Census, it was decided for the 2006 Census that CANCEIS would be used to perform E&I on 100% of the Census variables. In order to accommodate this large increase in the types of variables to be processed and in order to fully replace SPIDER, numerous enhancements to CANCEIS were required prior to processing the 2006 Census data.

3.1 Derive Methodology

The major advantage of SPIDER over CANCEIS in 2001 was the fact that SPIDER could perform donor imputation, deterministic imputation and derive variables while CANCEIS could only perform donor imputation. Thus, one of the main challenges for 2006 was to introduce into CANCEIS the ability to perform deterministic imputation and derive variables. It was decided to expand upon the existing CANCEIS syntax for donor imputation to incorporate all necessary functionality rather than try to mimic the SPIDER functionality exactly within CANCEIS. Over 5000 SPIDER DLTs had to be rewritten for CANCEIS for 2006. The syntax had to be as simple and efficient as possible in order to be easily adopted by the subject matter representatives who performed this translation.

The main difference between donor and derive (deterministic) DLTs is that fact that with donor imputation the user defines only conditions and allows the system to determine the imputation actions while with deterministic imputation the user must define both the conditions and the actions to take when a rule is matched.

Figure 2: Example of a CANCEIS Derive DLT.

```
$ DO DERIVE_AGE
$ FLAG = 0

@ AGE < 15               ;Y;Y;
@ MARITAL_ST = MARRIED   ;Y; ;
@ INCOME = 0             ; ;N;

& MARITAL_ST = SINGLE    ;X; ;
& INCOME = 0             ; ;X;
```

Figure 2 presents an example of a CANCEIS derive DLT that performs deterministic imputation. Lines beginning with an '@' symbol represent conditions, as they do in donor DLTs. Lines beginning with an '&' symbol represent conditional actions. When an edit rule is matched by a record, all conditional actions associated with that edit rule (indicated by an X in that edit rule's column) will be performed. For example, when the first edit rule is matched by somebody who is under the age of 15 and married then their value for the marital status variable 'MARITAL_ST' will be changed from 'MARRIED' to 'SINGLE'. When the second edit rule is matched by somebody who is under the age of 15 and reported an income other than 0, their value for the variable 'INCOME' will be changed to 0. Lines beginning with an '$' symbol represent common actions. These actions are performed unconditionally on all records which enter the DLT.

In donor imputation, every DLT and edit rule is independent of the others, and the order in which they are presented is usually organized based on clarity for the user. In deterministic imputation it is necessary to create a flow of logic for each record to follow as different actions will be taken depending on the data present. The first line in Figure 2 demonstrates the use of a 'DO' statement which is used to call another DLT. When actions such as this are encountered, the DLT which is called is entered by the current record and all appropriate actions are performed before returning to the original DLT. 'DO' statements can be used to call DLTs in both common and conditional actions.

A major challenge when designing CANCEIS's derive functionality was mimicking SPIDER's looping functionality. The household, in CANCEIS notation, is called a *unit* and the persons are called *subunits*. SPIDER applied a DLT to all members of a household using a standard looping system of letting a variable (typically I or J) represent the person number, setting it to 1 initially and incrementing the number by 1 after each iteration of the DLT. CANCEIS is designed to allow a set of edits to be applied to everyone in a household simultaneously. *Variable Position Subunits*, an established method of writing a single condition which applies to all members of the household, provided the necessary functionality to reproduce SPIDER's method of looping.

Figure 3: Expanded Form of a CANCEIS Derive DLT

```
@ VAR1(1) = INVALID  ;Y;N; ; ; ; ;
@ VAR1(2) = INVALID  ; ; ;Y;N; ; ;
@ VAR1(3) = INVALID  ; ; ; ; ;Y;N;

& VAR1(1) = A_BLANK  ;X; ; ; ; ; ;
& VAR1(2) = A_BLANK  ; ; ;X; ; ; ;
& VAR1(3) = A_BLANK  ; ; ; ; ;X; ;
& DO TABLE2          ; ;X; ;X; ;X;
```

Figure 4: Compact Form Using Variable Position Subunits

```
% sub-unit start position : 1
% sub-unit end position   : 3

@ VAR1(#1)= INVALID  ;Y;N;

& VAR1(#1)= A_BLANK  ;X; ;
& DO TABLE2          ; ;X;
```

Figure 3 demonstrates a derive DLT in which the same edits and actions are being applied to three subunits. Figure 4 presents this same DLT written in compact form using Variable Position Subunits. The three repeated lines have been reduced to one, with the #1 symbol replacing the specific subunit numbers of 1, 2 and 3. The two lines beginning with % belong to the *DLT Header*, which specifies parameters for a particular DLT, and indicate that the edit rules should be repeated three times with the #1 symbol replaced by the values 1 through 3 in the respective iteration. Note that if the 'DO TABLE2' action is performed, then that table will be called and all rules within it will be executed for only the current value of #1 (1, 2 or 3) before returning to this original table and repeating it for the next subunit number. This allows the user to control the flow of logic of their DLTs.

Variable Position Subunits is a very useful tool, but it is limited to looping through subunit numbers. Occasionally the SPIDER DLTs used looping in different situations, such as looping through a list of similarly named variables (i.e. TEMP1 through TEMP10). In order to accommodate this in the CANCEIS DLTs, a new method of writing compact DLTs called *Text Substitution* was created.

Figure 5: Expanded Form of a Derive DLT

```
$ DECL(VAR1TEMP,D)
$ DECL(VAR2TEMP,D)
$ DECL(VAR3TEMP,D)
$ DO TABLE1

@ VAR1TEMP = 25  ;N; ; ;
@ VAR2TEMP = 25  ; ;N; ;
@ VAR3TEMP = 25  ; ; ;N;

& VAR1TEMP = 10  ;X; ; ;
& DO TABLE2      ;X; ; ;
& VAR1TEMP = 25  ;X; ; ;
& VAR2TEMP = 10  ; ;X; ;
& DO TABLE2      ; ;X; ;
& VAR2TEMP = 25  ; ;X; ;
& VAR3TEMP = 10  ; ; ;X;
& DO TABLE2      ; ; ;X;
& VAR3TEMP = 25  ; ; ;X;
```

Figure 6: Compact Form Using Text Substitution

```
% Substitution Start Position: 1
% Substitution End Position: 3

$ DECL(VAR[1]TEMP,D)
$ DO TABLE1

@ VAR[1]TEMP = 25  ;N;

& VAR[1]TEMP = 10  ;X;
& DO TABLE2        ;X;
& VAR[1]TEMP = 25  ;X;
```

Figure 5 presents a CANCEIS derive DLT in which the same pattern of actions and conditions is repeated for the three variables VAR1TEMP, VAR2TEMP and VAR3TEMP. Figure 6 presents the same DLT written in a compact format using Text Substitution. Similar to Variable Position Subunits, the repeated lines have been reduced to a single line each with the numbers 1, 2 and 3 replaced with the text '[1]'. The advantage of Text Substitution over Variable Position Subunits is that it is not limited to replacing subunit numbers in compact DLTs. The disadvantage is that values are not retained between DLTs, so it can not be used to produce a flow of logic through the DLTS in the same way Variable Position Subunits can.

3.2 Other Enhancements for 2006

In 2001, CANCEIS could only process discrete, continuous and categorical variables. For 2006, CANCEIS has been expanded to allow for *Alphanumeric Variables* so that 'Postal Code' can be processed. This involved generalizing certain data files, as well as introducing new functionality into the DLTs, such as allowing the creation of substrings and the concatenation of variables.

Figure 7: Example of a CANCEIS Derive DLT.

```
@ -3*DISC_1 + DISC_2 > -DISC_3 + 6.7
@ CONT_1 = CONT_2 + 0.42 – CONT_3
```

CANCEIS in 2001 could already handle numeric propositions such as those seen in Figure 7, where DISC_# and CONT_# represent discrete and continuous numeric variables respectively. The 2006 Census requires extensive processing of continuous variables for the first time with variables such as 'Income'. In order to accommodate these new variables, the donor methodology has been expanded to allow for an *L_p Norm* distance measure. CANCEIS had been programmed assuming a p value of 1, but has now been generalized to accept all values of $p \geq 1$. Additionally, the ability to perform *Outlier Detection* on numeric variables has been introduced into CANCEIS. For variables such as 'Value of Dwelling' and 'Rent Paid', there are often extreme values found in the data which are acceptable as entries for that record but should not be reproduced through imputation. The outlier detection functionality in CANCEIS allows records containing data such as this to remain in the donor pool but restricts those outlier values from being imputed into failing records should those records be chosen as donors.

One of the areas for improvement in the 2001 results was the processing of demographic data for larger households, particularly 8 person households. These large households often experience a higher rate of non-response that limits the size of the donor pool considerably. This can result in some donor records being used many times to resolve multiple failing records. Often these records are failing due to only a few problematic responses for one or two people in the household but contain a large amount of good data for the majority of people in the household. Based on this situation, CANCEIS was enhanced to give the user the option of allowing failing records to be used as donors. If a failing record is chosen as a donor, then any invalid data in the donor record will not be imputed into the other failed record. Data which caused inconsistencies in the donor is still valid for imputation as it may not cause inconsistencies in the failing record. For example, if one record fails due only to an invalid age for the second subunit and another record fails due to only an invalid sex for the eighth subunit, one of these two records could theoretically be used to resolve the problem in the other. This strategy can increase the size of the donor pool considerably and reduce the number of times every given record is used as a donor. The use of this functionality is controlled by a system parameter, allowing the user to decide if it will be beneficial for their current module. The potential use of this feature in 2006 Census production has not yet been evaluated.

In 2001, SPIDER was linked to a custom database (RAPID) and processed in a mainframe environment. CANCEIS was processed on personal computers with text input files. For 2006, CANCEIS has been enhanced to interact directly with a SYBASE database environment. For production, CANCEIS will be processed on this server and access a centralized dictionary rather than processing on a personal computer with text files. It is important to note CANCEIS can still be processed on personal computers using text files for ease of use in the testing environment and for external clients.

4. CONCLUSION

Significant enhancements have been made to CANCEIS in a short amount of time in order to meet the requirements of the 2006 Canadian Census. All enhancements were made in a collaborative fashion between methodology, subject matter and systems representatives in the most efficient manner possible to ensure the system's readiness. CANCEIS has evolved significantly since its initial use in 1996, processing solely the demographic variables, to being the sole E&I software for the 2006 Census. CANCEIS will continue to add functionality to meet the needs of the Canadian Census. For the 2011 Census, it will be assessed whether processing can be made more efficient and data quality improved by processing larger groups of variables simultaneously (e.g. Income with Labour Force Activity).

References

Bankier, M., Poirier, P., and Lachance, M. (2001), "Efficient Methodology Within the Canadian Census Edit and Imputation System (CANCEIS)", ASA Joint Statistical Meetings, Atlanta.

Bankier, M. (2006), "Imputing Numeric and Qualitative Variables Simultaneously", A Technical Report Detailing the Methodology of CANCEIS, Statistics Canada.

Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association", March 1976, Volume 71, No. 353, 17-35.