MINIMUM CHANGE EDIT AND IMPUTATION FOR THE 2006 CANADIAN CENSUS

Darryl Janes and Michael Bankier
Statistics Canada, Ottawa, ON, K1A 0T6

I. INTRODUCTION

Many minimum change imputation systems are based on the approach proposed by Fellegi and Holt (1976). For example, DISCRETE and SPEER at United States Bureau of the Census, and CANEDIT and GEIS/Banff at Statistics Canada all use, or had as their starting point, the Fellegi/Holt imputation methodology. A somewhat different approach was successfully used in the 1996 Canadian Census of Population to impute for non-response and inconsistencies for the demographic variables of all persons in a household simultaneously. This method, called the Nearest-neighbour Imputation Methodology (NIM), permitted for the first time the simultaneous minimum change imputation of qualitative and quantitative variables for large E&I problems. An overview of the NIM algorithm is provided in Bankier (1999).

Bankier et al. (2001) describes the main difference between the NIM and the Fellegi/Holt imputation methodologies. NIM first finds donors and then determines the minimum number of variables to impute based on these donors, while the Fellegi/Holt methodology determines the minimum number of variables to impute first, and then attempts to find donors. Reversing the order of these operations confers significant computational advantages to implementations of the NIM while still meeting the well-accepted Fellegi/Holt objectives of minimum change and preserving sub-population distributions.

NIM was limited to hot deck edit and imputation (E&I) in the 1996 and 2001 censuses. In 2006, NIM must be able to handle variable derivations and deterministic imputation. This paper will identify the challenges of creating such an E&I system for 2006.

II. E&I IN 2001 CANADIAN CENSUS

For the 2001 Census, a more generic implementation of the NIM was developed, called the CANadian Census Edit and Imputation System (CANCEIS). It was used to perform hot-deck donor E&I on personal computers for about 40% of the variables including Demography, Labour, Mobility, Place of Work, and Mode of Transport. It was able to apply up to 43,000 edits for 30,000,000 people.

The remaining 60% of the hot deck imputation was performed by the mainframe E&I system called SPIDER (System for Processing Instructions from Directly Entered Requirements), which had been used since the 1981 Census. SPIDER was also used to handle the *pre-derives* and *post-derives* of all census variables. A *derive* is any module that is performed before or after a hot-deck donor imputation module and typically generates stratification identifiers, derives variables, and performs deterministic imputation.

A. A Review of CANCEIS in 2001

Besides developing and testing CANCEIS, it was necessary to work with subject matter specialists to develop *Decision Logic Tables (DLTs)* and *data dictionary files*. DLTs define edit rules that specify conditions that are not permitted for a dataset. A data dictionary is a set of files used to define information such as the names of the variables, their sets of valid responses, text labels associated with numeric responses, variable weights and associated distance measures, and parameters to be applied to the E&I process. All input files and output files are in text format.

Figure 1: Example of a CANCEIS DLT in 2001.

```
@ AGE < 15                      ;Y; ; ; ;
@ MAR_STAT = NEVER_MARRIED      ;N; ;Y; ;
@ AGE - YEARS_MARRIED < 15      ; ;Y; ; ;
@ YEARS_MARRIED > 0             ; ;Y;Y; ;
@ AGE < 10                      ; ; ; ;Y;
@ INCOME > 0                    ; ; ; ;Y;
```

During the E&I process, CANCEIS uses the edit rules to flag as "failed" any records with inconsistent data. Any failed records with invalid or inconsistent data will be sent to hot-deck donor imputation. In Figure 1, we see an example of a DLT where there are 4 columns of edit rules, and 6 conditions or *propositions* (beginning with an '@' symbol). The first rule, for example, states that the record should fail if someone has an age below 15, and has also been married. CANCEIS imputes new data for a failed

record from a donor that resembles the failed record so that none of the edit rules fail.

Edit rules are generally used to define inconsistent responses while the data dictionary is used to define which responses are valid. For example, if we define AGE to have a valid range from 0 to 120, then a value of 200 will be flagged as invalid and will be replaced in donor imputation.

Donors are typically found in a *ripple search*, alternating the search up and down on the data file, progressively moving away from the failed record. The dataset is usually sorted so that adjacent units in the dataset would be geographically close (within neighbourhoods of similar demographics, income, etc.).

CANCEIS stores a list of the best potential donors by comparing the value of each variable of the failed record with the value in the potential donor. If there are I variables, then CANCEIS retains the smallest values of the distance score $D_{fp} = \sum D_i W_i$, *where* f and p indicate a score between the failed and passed (donor) record. D_i is a distance measure within [0,1] between the i^{th} variable responses for the failed and donor records, and W_i is the non-negative weight of the i^{th} variable for $i=1,\ldots,I$. The best potential donors, or *nearest neighbours*, are then further examined to find the best *imputation actions that pass the edits*. Only imputation actions involving the minimum number (or near the minimum number) of changes will be examined, and one of the best imputation actions will be chosen randomly.

The generation of imputation actions for a specific failed edit/donor pair is done in an efficient fashion. First, blanks and invalids are imputed. If an imputed record still fails the edits because of inconsistent responses, failing edits are evaluated to determine if certain variables must be imputed. For example, if an edit rule fails because of someone who claims to be 5 years old and married, but the donor is 5 years old and single, then the only imputation possible to resolve the edit failure is to change the marital status from married to single. These *essential-to-impute* variables must be imputed. Edit rules that do not link to responses not found in the failed or donor records cannot fail, and are eliminated. For example, any edits involving grandparents are dropped if there are no grandparents in either the failed edit or donor household. Then the only variables that are candidates for imputation are those that enter these "simplified" edits and for which the failed edit record and the donor has a different response. All possible imputations are generated using a binary tree and assessed against the simplified edits. Frequently branches of these binary tree can be "pruned" without generating all the imputation actions if it is determined that none of them will pass the simplified edits.

B. SPIDER vs. CANCEIS

The 2001 Census represented the 5th time that SPIDER was used, but only the first time for CANCEIS. In the 20 years after the introduction of SPIDER, there have been major advances in information technology. CANCEIS was able to take advantage of more recent technology, whereas SPIDER was a system built from technology around 1980 and was limited to a mainframe environment using a custom data base called RAPID. For the 2006 Census, it was decided to use the commercial data base, SYBASE. This precluded SPIDER continuing to be used. We compared, as noted below, the attributes of CANCEIS and SPIDER to determine what extensions were required to CANCEIS to allow it to replace SPIDER for the 2006 Census.

1. SPIDER had high usage costs since it ran on a mainframe computer. By running CANCEIS on personal computers, we avoided hundreds of thousands of dollars in mainframe costs. Furthermore, the software could be run on multiple computers simultaneously. We also found that running CANCEIS on a 1.7GHz PC was between 1 and 3 times as fast as the mainframe, depending on the module. CANCEIS was particularly faster during file I/O. Modules that took many days with SPIDER could often be done in hours with CANCEIS.

2. SPIDER was linked to the RAPID database at Statistics Canada, and therefore was unable to be expanded for use in other domains. CANCEIS was not linked to a database and used text files as input. This allowed CANCEIS to be used in non-census surveys both within and outside of Statistics Canada.

3. SPIDER used a crude approximation of the Fellegi and Holt (1976) methodology, while CANCEIS was built using the NIM methodology.

4. The SPIDER system did not allow much flexibility or user control during the imputation process. CANCEIS, however, was built to be more versatile; a system parameter file allows for

imputation parameters to be easily changed, and a Windows Graphical User Interface will allow for more control and functionality.

5. The major advantage of SPIDER over CANCEIS in 2001 was that it did both hot-deck and deterministic imputation, plus it had the ability to generate new variables. CANCEIS was limited to hot deck imputation in 2001. However, CANCEIS will be undergoing significant enhancements in order to be able to do variable derivation and deterministic imputation for the 2006 Census.

III. CHALLENGES FOR 2006

Rather than build CANCEIS to mimic SPIDER's syntax, it was decided to expand CANCEIS syntax to be more in line with the style used in hot-deck imputation. The retirement of SPIDER also means that all modules have to be modified to the new CANCEIS format. Another challenge is to improve CANCEIS' level of user friendliness. This involves the creation of a user interface, a new error-message system and other helpful tools.

A. NEW DERIVE FUNCTIONALITY IN 2006

The expansion of CANCEIS to include deterministic imputation and variable derivation is a complex process that must be done quickly in order to leave time for other upgrades, the preparation of modules, and testing before the 2006 Census. Nearly 5000 SPIDER DLTs need to be rewritten into a new CANCEIS format, and almost all of these perform deterministic imputation and variable derivation. Since most census modules have to be prepared and tested by August 2005, the new CANCEIS syntax must be simple and efficient enough to be designed and programmed, and then adopted by subject matter in autumn 2004.

When designing new syntax within CANCEIS decision logic tables, it was necessary to add functionality to handle issues such as looping, deterministic actions, assigning values to variables and DLT calls. It was also desirable to allow the creation of compact DLTs using *Variable Position Subunits* and *Text Substitution*.

CANCEIS is designed to allow a set of edits to be applied to everyone in a household simultaneously. The household, in CANCEIS notation, is called a *unit* and the persons are called *subunits*. Frequently, we want to apply an edit to all pairs of persons in a household. Figure 2 shows an example of a set of edits where this is being done. Figure 3 shows the same DLT in compact mode using variable position subunits. We can see that the subunit numbers 1, 2, and 3 in Figure 2 have been replaced by variable position subunit numbers #1 and #2 in Figure 3, which act as looping variables. In these DLTs, lines that begin with '@' are the conditions which make up the edit rules, and the lines beginning with '&' are the actions carried out (as flagged by an 'X') if an edit rule is matched. The headers, which begin with '%', control the looping of #1 and #2 from 1 to 3.

Figure 2: Expanded Form of a CANCEIS DLT

```
@ sexu(1) = sex(2) ;Y;N; ; ; ; ; ;…
@ sexu(2) = sex(1) ; ; ;Y;N; ; ; ;…
@ sexu(1) = sex(3) ; ; ; ; ;Y;N; ;…
@ sexu(3) = sex(1) ; ; ; ; ; ; ;Y;…
@ sexu(2) = sex(3) ; ; ; ; ; ; ; ;…
@ sexu(3) = sex(2) ; ; ; ; ; ; ; ;…

& Var1(1)=A_blank  ; ; ;X; ; ; ;X;…
& Var1(2)=A_blank  ;X; ; ; ; ; ; ;…
& Var1(3)=A_blank  ; ; ; ; ;X; ; ;…
& Do table 2       ; ;X; ;X; ;X; ;…
```

Figure 3: Compact Form Using Variable Position Subunits

```
% symmetry    : no
% sub-unit start position : 1
% sub-unit end position   : 3

@ sexu(#1)=sex(#2)  ;Y;N;

& Var1(#2)=A_blank  ;X; ;
& Do table_2        ; ;X;
```

Figure 4 shows a DLT in expanded form while Figure 5 shows the same DLT in compact mode using text substitution syntax. We can see that numbers 1, 2, and 3 have been replaced by the text '?1', and that the number of conditions and actions has been greatly reduced.

We can also see in Figure 5 an action statement that says, 'DO TABLE2'. This is special syntax that calls another DLT from the current one. This DLT call acts like a subroutine. There are other similar statements and functions that will be useful in the derive process as we provide functionality necessary to replace SPIDER.

Figure 4: 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 5: Compact Form Using Text Substitution

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

$  DECL(VAR?1TEMP,D)
$  DO TABLE1

@ VAR?1TEMP = 25  ;N;

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

B. IMPROVING USER FRIENDLINESS

For 2006, there will be a series of tools available to help with the use of CANCEIS. One of these is the CANCEIS Graphical User Interface. It is a Visual Basic application that can run the CANCEIS programs, but it also greatly assists with the creation of the data dictionary. By using the CANCEIS Interface, spelling errors of variables and parameters are reduced, and files can be created more efficiently.

Another tool being developed is the DLT Editor, which can be accessed through the CANCEIS Interface. It is another Visual Basic application that links to the data dictionary to allow the creation or editing of DLTs. Spelling mistakes and syntax errors are reduced because DLT headers are generated automatically and variable names and responses are selected from lists.

C. OTHER IMPROVEMENTS FOR 2006

There are several other enhancements planned for 2006. For example, a new set of parameters will be available that will give the user more control in terms of how persons are reordered so that the failed edit household more closely resembles a donor.

In 2001, CANCEIS could only process discrete, continuous, and coded variables. In 2006, the introduction of Lp norms will allow more extensive use of continuous variables. Also, CANCEIS will be able to process alphanumeric variables such as names, streets, and Canadian postal codes.

CANCEIS already generates several files of summary statistics or reports. However, these files will be reviewed for the purpose of providing the user with an improved series of useful summary reports. These reports may be generated by either the CANCEIS program or by a new data analysis tool.

IV. CONCLUSION

CANCEIS has evolved from performing donor imputation for only the demographic variables in 1996 to the primary E&I software for the 2006 Canadian Census. CANCEIS continues to expand in order to meet the E&I needs of the Canadian Census and due to its versatility, it is being examined by other national statistical agencies for possible use in their surveys.

References

Bankier, M. (2003), “Current and Future Applications of CANCEIS at Statistics Canada”, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Spain (Madrid). (http://www.unece.org/stats/documents/2003.10.sde.htm)

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. (1999), "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future

Censuses", Proceedings of the UN/ECE Work Session on Statistical Data Editing, Italy. (http://www.unece.org/stats/documents/1999.06.sde.htm)

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.

Janes, D. (2004), "CANCEIS User's Guide", a Complete Reference Guide for the use of CANCEIS, Statistics Canada.

Mason, P., Bankier, M. and Poirier, P. (2002), "Imputation of Demographic Variables from the 2001 Canadian Census of Population", ASA Joint Statistical Meetings, New York.