IMPUTATION OF DEMOGRAPHIC VARIABLES FROM THE 2001 CANADIAN CENSUS OF POPULATION

Patrick Mason, Michael Bankier and Paul Poirier
Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

I. INTRODUCTION

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

The main difference between the NIM and the Fellegi/Holt imputation methodology is that the NIM first finds donors and then determines the minimum number of variables to impute based on these donors. The Fellegi/Holt methodology determines the minimum number of variables to impute first, and then finds 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. The NIM, however, can only be used to carry out imputation using donors while the Fellegi/Holt can be used with any imputation methodology.

For the 2001 Census, a more generic implementation of the NIM has been developed. It is called the CANadian Census Edit and Imputation System (CANCEIS). It is written in the ANSI C programming language and uses ASCII files. As a result, with only minor modifications, it can be used on many platforms such as the PC or mainframe, and under different operating systems. Besides the demographic variables, it will be used in the 2001 Canadian Census to perform E&I for the labour, mobility, place of work, and mode of transport variables. This corresponds to about half of all variables on the 2001 Census questionnaire. For the 2006 Canadian Census, CANCEIS will be used to process all census variables.

Section II describes how the demographic E&I was performed in the 2001 Canadian Census using CANCEIS. Section III discusses how CANCEIS parameters can induce more plausible results for the demographic data. Finally, some concluding remarks are given in Section IV. For more details regarding the NIM and CANCEIS, see Bankier, Lachance and Poirier (2000, 2001).

II. OVERVIEW OF THE E&I PROCESS FOR THE DEMOGRAPHIC VARIABLES

For the Canadian Census of Population, five demographic questions are asked of each person. There are questions related to age, sex, marital status, common-law status and relationship to the household representative (also known as Person1). The responses given for each of these variables are edited and imputed simultaneously for all persons within a household. Furthermore, while respondents are not asked explicitly who are the couples and families in the household, these characteristics are disseminated. Therefore these variables need to be derived and edited.

There are four steps in the E&I of the demographic variables. The first step is the correction of known systematic reporting errors found in the demographic data. The second step is the derivation of the Couple, ChildOf and GrandchildOf variables. These variables are used during the editing to identify which persons are couples, parent/child pairs or grandparent/grandchild pairs. The third step is the editing, where edit rules are used to define inconsistencies in the data, such as a married 3-year old. The final step is the imputation of any missing/invalid or inconsistent responses. These four steps are described in parts A to D of this section.

A. Treatment of Systematic Errors by Deterministic Imputation

Sometimes, imputing the minimum number of variables is not the best way to resolve edit failures. This is particularly evident in the case of systematic errors that can be found in demographic data. Obviously, the ability to correct these systematic errors is dependent on the ability to detect these errors. This may not always be an easy task given that the response patterns are different for each census due to modifications to the questionnaire and changing social trends.

For 2001, there were several systematic errors that were corrected during production. One example of a systematic reporting error is where everyone in a family reports being in a common-law union. This is illustrated in Table 1 below. For this household, minimum change imputation would likely impute only the two variables, Relationship for person 4 and Common-law Status for person 5. A more plausible imputation action can be achieved by correcting the systematic error by imputing the Common-law Statuses for persons 3, 4, and 5.

Table 1: Systematic Reporting Error of Common-law Status Variable

Position on questionnaire	Relationship to Person 1	Age	Sex	Marital Status	Common-law Status
Unimputed Data					
1	Person 1	43	Male	Single	Yes
2	Opposite-sex Partner	39	Female	Divorced	Yes
3	Son/Daughter	19	Male	Single	Yes
4	Son/Daughter	16	Female	Single	Yes
5	Son/Daughter	14	Female	Single	Yes
Resulting household under correction of reporting error					
1	Person 1	43	Male	Single	Yes
2	Opposite-sex Partner	39	Female	Divorced	Yes
3	Son/Daughter	19	Male	Single	**»No**
4	Son/Daughter	16	Female	Single	**»No**
5	Son/Daughter	14	Female	Single	**»No**
Possible resulting household under minimum change imputation					
1	Person 1	43	Male	Single	Yes
2	Opposite-sex Partner	39	Female	Divorced	Yes
3	Son/Daughter	19	Male	Single	Yes
4	**»Son/Daughter-in-law**	16	Female	Single	Yes
5	Son/Daughter	14	Female	Single	**»No**

B. Derivation of Couple, ChildOf and GrandchildOf Variables for Editing

The Couple variable is used to identify potential couples prior to editing. In order to derive the Couple variable, a score is assigned to each possible pair of persons in the household based on the unimputed responses to all of the demographic variables and the proximity of the persons to each other on the questionnaire. The given score reflects the likelihood of the pair being an actual couple. The pairs with the highest scores are retained with a person being allowed to belong to only one potential couple. The Couple variable is set to the same value for the two persons of a specific couple. This variable is then used in editing, imputation and to determine the final Census and Economic Families.

In the same manner, the ChildOf and GrandchildOf variables are used to identify potential parent/child and grandparent/grandchild pairs prior to editing. Much like the Couple variable, a score is assigned to each possible pair of persons in the household based on unimputed responses and proximity on the questionnaire.

The household in Table 2 illustrates how this algorithm works. In this household, the persons in positions 1 and 2 are likely a couple as they have appropriate relationships, ages and proximity, and the other variables do not indicate that these persons are not a couple. They are identified as a potential couple by setting the Couple variable to the same value (11) for both of them. The persons in positions 4 and 5 are also identified as a potential couple since their proximity, ages, sexes and relationships all indicate a potential couple, even though the common-law status for Person 5 is No. Similarly, the person in position 1 is identified as the potential child of person 7. The ChildOf variable for the potential child is set to the value of the Couple variable for the potential parent(s). For this pair, the relationships are appropriate and there is no evidence that contradicts a parent/child relationship. Person 6 is identified as the child of person 3 since the relationships and ages are appropriate. Note that person 4 was not identified as the potential parent of person 6 since there was not a large enough age difference (15 years or more), and thus this pair did not receive as high a score as the {person 3, person 6} pair. During this process, persons 3 and 4 are also identified as potential children of the couple in positions 1 and 2. In calculating the GrandchildOf variable, persons 3 and 4 are identified as potential grandchildren of person 7. Person 6 is identified as a potential grandchild of persons 1 and 2.

Table 2: Example of Output of Couple, ChildOf and GrandchildOf Algorithm

Position on questionnaire	Relationship to Person 1	Age	Sex	Marital Status	Common-law Status	Couple	ChildOf	Grand childOf
1	Person1	57	Male	Married	No	*11*	*7*	0
2	Husband/Wife	53	----	----	No	*11*	0	0
3	Son/Daughter	35	----	Separated	No	3	*11*	*7*
4	Son/Daughter	23	Female	Single	Yes	*12*	*11*	*7*
5	Son/Daughter-in-law	22	Male	Single	No	*12*	0	0
6	Grandchild	12	Male	Single	No	6	*3*	*11*
7	Father/Mother	----	Female	Widowed	No	7	0	0

C. Editing with CANCEIS

After the identification of the potential family structure, edit rules are applied to identify which households require imputation. There are two primary types of edit rules used: within person edits and between person edits. Within person edit rules (for example, a person cannot be both married and less than 15 years of age) are used to edit individual persons in a household. Between person edit rules (for example, the age difference between a grandparent and grandchild is less than 30 years) are used to edit two or more persons simultaneously within a household. The edit rules related to couples are between person edits that are applied only to the potential couples identified. An example of these couple edit rules is illustrated in Table 3 for the {Son/Daughter, Son/Daughter-in-law} couples. Edit rules similar to those presented in this table exist for pairs of persons with other relationships that could form couples (for example a {Brother/Sister, Brother/Sister-in-law}). The ChildOf and GrandchildOf variables are used in a similar fashion as the Couple variable to specify the between person edit rules relating to those pairs.

Decision Logic Tables such as the one illustrated in Table 3 are used to generate rules for all possible combinations of two persons in the household. The quantities "#1" and "#2" are used to represent any combination of positions for the two persons. The first proposition ensures that the rules are applied only to the potential couples identified in the second step.

The Canadian Census requires that a couple is married or living in a common-law relationship and if they are married they must be of opposite sex. In addition, the partner of someone in a common-law relationship must be present in the household. The set of rules in Table 3 ensures that couples respect these conditions. If a potential couple match one of these edit rules then there are two possible outcomes. Either the variables that caused the household to fail are changed so as to be appropriate for a couple, or the relationship of one person is changed such that the relationships are no longer appropriating for a couple.

Table 3: Between Person Edit Rules for {Son/daughter, Son/daughter-in-law} Couples

Propositions	**Rules**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Couple#1 = Couple#2	Y	Y	Y	Y	Y	Y	Y	Y	Y
Relationship#1 = Son/Daughter	Y	Y	Y	Y	Y	Y	Y	Y	N
Relationship#2 = Son/Daught-in-law	Y	Y	Y	Y	Y	Y	Y	N	Y
Sex#1 = Sex#2	Y								
Marital status#1 = Married	Y	Y	N			N			
Marital status#2 = Married	Y	N	Y				N		
Common-law status#1 = Yes				Y	N	N		Y	
Common-law status#2 = Yes				N	Y		N		Y

D. Imputation with CANCEIS

CANCEIS imputes using the Nearest-neighbour Imputation Methodology (NIM). This method is based on the principle of minimum change while taking into consideration the plausibility of the imputation actions.

NIM performs E&I by first identifying the passed edit households which are as similar as possible to the failed edit household. These households are called nearest neighbours donors. For each nearest neighbour donor, the NIM attempts to impute each combination of variables that do not match the responses for the failed edit household. One of these minimum change imputation actions that passes the edits and most resembles both the failed edit household and the passed edit household is then randomly selected.

The notion of similarity is based on a distance function. It will be assumed that F households fail the edit rules, while P households pass the edit rules. The responses for the households that failed and passed the rules are labelled respectively by $V_f = [V_{fi}]$, $f = 1$ to F and $V_p = [V_{pi}]$, $p = 1$ to P, $i = 1,..,I$. These are Ix1 vectors containing the responses for all the persons in a household, where I will vary according to the household size. The distance between each failed edit household V_f, and each passed edit household V_p is defined as:

$$D_{fp} = D(V_f, V_p) = \sum_{i=1}^{I} w_i D_i(V_{fi}, V_{pi})$$

The weights, w_i, of the variables (which are non-negative) can be given smaller values for variables where it is considered less important that they match, for example, variables considered more likely to be in error. For the demographic data in the 2001 Canadian Census, all weights were set to one except for auxiliary variables, which are described in Section II.E.

The distance function $0 \leq D_i(V_{fi}, V_{pi}) \leq 1$ can be different for each variable i. For qualitative variables, the distance function often simply takes on the value 0 (if $V_{fi} = V_{pi}$) or 1 (otherwise). Another frequently used distance function is the distance matrix, which is used when some responses to qualitative variables are somehow similar.

For example, a distance matrix was implemented in 2001 for the Relationship variable to indicate similar responses and to resolve some multiple responses. An example of a distance matrix is given in Table 4 where it is assumed that the distance is always 0 when $V_{fi} = V_{pi}$. The Comlaw_Partner_of_Daughter relationship can also be

reported as a Son-in-law and thus a smaller distance of 0.25 was chosen instead of 1. However, the relationship Comlaw_Partner_of_Daughter is still preferred so it still receives a distance of 0. Similarly, stepsons are often reported as sons and thus a smaller distance of 0.25 was chosen instead of 1.

Table 4: Example of a Distance Matrix

$V_f \backslash V_p$	Son-in-law	Son	Wife	Comlaw_Partner_of_Person_1	All Other Relationships
Comlaw_Partner_of_Daughter	0.25	1	1	1	1
Stepson	1	0.25	1	1	1
Wife_or_Comlaw_Partner_of_Person_1	1	1	0	0	1
All Other Relationships	1	1	1	1	1

The value Wife_or_Comlaw_Partner_of_Person1 indicates that both Wife and Comlaw_Partner_of_Person1 were reported on the questionnaire. Since multiple responses are not allowed for this variable, only one of these two responses can be retained. With a distance of 0 in the distance matrix, CANCEIS can impute, without penalty, either Wife or Comlaw_Partner_of_Person1 while imputing any other value will have a distance of 1. Thus the distance matrix allows multiple responses to be resolved based on the frequency of the two responses among the nearest neighbour donors.

For each failed edit household, the N passed edit households (N might equal 40) with the smallest distances are considered as potential donors for the failed edit household. Only non-matching variables (those with $V_{fi} \neq V_{pi}$) are, of course, considered for imputation. Various subsets of these non-matching variables are imputed to determine which are the optimum imputations for a given {failed edit household, passed edit household} pair. Each of these subsets will be called an imputation action. The different possible imputation actions based on these N potential donors are generated and one of the optimal ones (as defined below), which passes the edit rules, is randomly selected to be the actual imputation action used for the failed edit household. For each possible imputation action that passes the edit rules, the following weighted distance is calculated:

$$D_{fpa} = \alpha D(V_f, V_a) + (1-\alpha) D(V_a, V_p)$$

where $D(V_f, V_a)$ is the distance between the failed edit household and the imputed household (this measures the amount of change to the data), and $D(V_a, V_p)$ is the distance between the imputed household and the passed edit household (this measures he plausibility of the imputation action). The parameter α can take on a value between 0.5 and 1. As α approaches 0.5, more emphasis is placed on minimising $D(V_a, V_p)$ rather than minimising $D(V_f, V_a)$. This weighted distance is calculated for each potential imputation action and is used to determine the probability of selection.

An example of where plausibility is preferred over minimum change is given in Table 5. In this example, the minimum change imputation action would be to impute 3 variables. However, this would introduce a lodger to the household and would create a family where the wife and son have the same age. Although this is conceptually permissible, it should rarely occur. The more plausible imputation action will likely impute 4 variables since it is more prevalent among the donors.

Table 5: Example of Plausibility Vs. Minimum Change

Position on questionnaire	**Relationship to Person 1**	**Age**	**Sex**	**Marital Status**	**Common-law Status**
Unimputed data					
1	Person1	46	Male	Married	No
2	----	45	Female	Separated	No
3	Son	21	Male	Single	No
4	Common-law Partner of Son	21	Female	Married	Yes
Imputed data with minimum change imputation action					
1	Person1	46	Male	Married	No
2	**»Lodger**	45	Female	Separated	No
3	Son	21	Male	Single	No
4	**»Wife**	21	Female	Married	**»No**
Imputed data with plausible imputation action					
1	Person1	46	Male	Married	No
2	**»Wife**	45	Female	**»Married**	No
3	Son	21	Male	Single	**»Yes**
4	Common-law Partner of Son	21	Female	**»Single**	Yes

Of all the imputation actions considered which pass the edit rules, only those which minimise or nearly minimise the weighted distance are retained. These imputation actions are called "minimum change imputation actions" or "near minimum change imputation actions". Near minimum change imputation actions are retained because, for practical purposes, they are (particularly with quantitative variables) nearly as good as minimum change imputation actions.

A size measure defined as R_{fpa}= (min D_{fpa} / D_{fpa})^t is calculated for each near minimum change imputation action. The parameter t is used to give more or less weight to the minimum change imputation actions as opposed to the near minimum change imputation actions. For the 2001 Census, the parameter t had a value of 1. One of the potential near minimum change imputation actions is randomly selected, with probability proportional to R_{fpa}, to be the actual imputation action for V_f.

E. Use of Auxiliary Variables

For the Canadian Census long form (given to 20% of the households), the E&I of the demographic variables is done before the E&I of the other variables on the questionnaire. This is done because of operational and computational considerations. Some of the other long form questions are only to be completed by adults (i.e. at least 15 years old). Thus, if a "true" adult has an age of less than 15 imputed, then any responses received for these long form questions will be dropped. In the same manner, if a "true" child has an age of 15 or more imputed then all these long form questions would not have been answered and will require imputation. In order to minimise this problem, long form variables can be used to give an indication of a person's "true" demographic characteristics. Such long form variables, which do not enter the demographic edits, will be called auxiliary variables.

Without the auxiliary variables, it is unclear whether the Son in Table 6 should be an adult or not. Although the auxiliary information is unedited and thus somewhat less reliable, it provides substantial evidence that the Son is, in fact, an adult. Having the auxiliary information in the distance measure for this example will result in the majority of donors having age greater than 20 for the Son.

For the first time, the 2001 Canadian Census used auxiliary variables from long form questionnaires during the imputation of the demographic variables. The total weight for the auxiliary variables equalled 1, which is the same as each individual demographic variable. The auxiliary variables were given smaller weights than the demographic variables to reflect the fact that their unimputed responses were considered somewhat less reliable.

Table 6: Illustration of Use of Auxiliary Variables

Position on questionnaire	**Relationship To Person 1**	**Age**	**Sex**	**Marital Status**	**Common-law Status**	***Highest Grade***	***Hours Worked***	***Total Income***
1	Person 1	48	Male	Married	No	*13*	*40*	*$ 49,000*
2	Husband/Wife	46	Female	Married	No	----	----	*$ 32,000*
3	Son/Daughter	----	Male	Single	No	*17*	*45*	*$ 62,000*

III. OPTIMISATION OF PARAMETERS

There are many parameters within CANCEIS that can have an effect on processing time and the quality of the imputation. For the 2001 Census, a few parameters were altered from the values assigned for the 1996 Census including the α parameter in D_{fpa} calculation, the age distance function, and donor search stage sizes.

In the 1996 Census, α = 0.9 was used. For the 2001 Census it was decided to continue the use of α = 0.9 for the smaller household sizes, but use α = 0.75 for households with 4 or more persons. The lower α value was preferred for the larger households because it was found that as the complexity and variability of the households increased, it was important to place more weight on the plausibility aspect of the distance function. An example of this is given in Table 7 (the failed edit household) and Table 8 (potential donor #1 and #2).

Table 7: Failed Edit Household

Position on questionnaire	**Relationship to Person 1**	**Age**	**Sex**	**Marital Status**	**Common-law Status**
1	Person 1	38	Female	Married	No
2	Husband/Wife	40	Male	Married	No
3	----	73	Female	Separated	No
4	Nephew/Niece	2	Female	Single	No
5	Brother/Sister	48	Female	Divorced	No
6	Nephew/Niece	22	Male	Single	No

In Table 8, potential donor #1 would impute a single variable (relationship of person 3 to Lodger) to make the failed edit household pass the edits. Potential donor #1, however, does not look much like the failed edit household. Also the resulting imputed record is somewhat implausible with the Sister in position 5 having both a 22-year old and a 2-year old child since person 3 is now a lodger and therefore unrelated to person 4.

Potential donor #2 imputes two variables (relationship of person 3 to Sister and Age to 33) to make the failed edit household pass the edits. Although this is more than the minimum number of variables, potential donor #2 looks very much like the failed edit household and the resulting imputed record is more plausible. With $\alpha = 0.9$, potential donor #1 would be preferred with a distance of 1.87 compared to the distance of 2.25 of potential donor #2. With $\alpha = 0.75$, potential donor #2 is preferred having a distance of 2.62 in contrast to potential donor #1's distance of 3.19.

Table 8: Potential Donors

Position on questionnaire	**Relationship To Person 1**	**Age**	**Sex**	**Marital Status**	**Common-law Status**	**$D(V_f,V_a)$**	**$D(V_a,V_p)$**	**D_{fpa}** α =0.9	**D_{fpa}** α =0.75
Potential Donor #1									
1	Person 1	39	Female	Married	No	0	0.246	0.02	0.06
2	Husband/Wife	42	Male	Married	No	0	0.527	0.05	0.13
3	***Lodger***	47	Female	Separated	No	1	1	1.0	1.0
4	Lodger	46	Female	Divorced	No	0	3	0.3	0.75
5	Lodger	48	Female	Divorced	No	0	1	0.1	0.25
6	Lodger	49	Female	Divorced	No	0	4	0.4	1.0
						Total D_{fpa}		**1.87**	**3.19**
Potential Donor #2									
1	Person 1	35	Female	Married	No	0	0.634	0.06	0.16
2	Husband/Wife	37	Male	Married	No	0	0.609	0.06	0.15
3	***Brother/Sister***	***33***	Female	Divorced	No	2	1	1.90	1.75
4	Nephew/Niece	3	Female	Single	No	0	0.305	0.03	0.08
5	Brother/Sister	46	Male	Divorced	No	0	1.373	0.14	0.34
6	Nephew/Niece	20	Male	Single	No	0	0.556	0.06	0.14
						Total D_{fpa}		**2.25**	**2.62**

IV. CONCLUSIONS

CANCEIS has been shown to be a highly efficient E&I system which can be used by censuses and in various types of surveys to handle minimum change hot-deck imputation. Further enhancements to CANCEIS towards the 2006 Canadian Census of Population, such as adding the ability to perform deterministic imputation and a graphical user interface, will make CANCEIS an attractive choice for an increasing number of surveys within Statistics Canada.

At this point, CANCEIS has been used exclusively with Social and Household surveys (i.e. surveys with a mixture of many qualitative and quantitative variables), using a mixture of hot-deck and deterministic imputation. CANCEIS has the potential to be used with business surveys, but more study of the requirements of these surveys and some extensions to the system may be required.

References

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 (Rome). (http://www.unece.org/stats/documents/1999.06.sde.htm)

Bankier, M., Lachance, M. and Poirier, P. (2000), "2001 Canadian Census Minimum Change Donor Imputation Methodology", Proceedings of the UN/ECE Work Session on Statistical Data Editing, United Kingdom (Cardiff). (http://www.unece.org/stats/documents/2000.10.sde.htm)

Bankier, M., Lachance, M. and Poirier, P. (2001), "2001 Canadian Census Minimum Change Donor Imputation Methodology - Extended Version of Report", Social Survey Methods Division Report, Statistics Canada, Dated February, 2001.

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.