1996 CANADIAN CENSUS DEMOGRAPHIC VARIABLES IMPUTATION

Michael Bankier, Anne-Marie Houle, Manchi Luc and Patricia Newcombe
Anne-Marie Houle, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6, houlann@statcan.ca

KEYS WORDS: Minimum change hot deck imputation, inconsistent response, couple edit rules.

1. INTRODUCTION

Among the basic questions asked of every Canadian on Census Day are the five questions related to the demographic variables age, sex, marital status, common-law status and relationship to Person 1. The responses given to these questions are examined simultaneously for all persons in a household to identify missing and inconsistent responses and to make the appropriate corrections.

A New Imputation Methodology (NIM) was used in the 1996 Canadian Census to carry out Edit and Imputation (E&I) for these variables. This methodology allows, for the first time, minimum change imputation of numeric and qualitative variables simultaneously for large E&I problems.

There exist a variety of imputation methods. Two important types are deterministic imputation and hot deck imputation. Some imputation methodologies use only deterministic imputation or only hot deck imputation, while others incorporate the two types of imputation. The NIM is a hot deck imputation methodology that uses some deterministic imputation in a prederive module.

In Section 2, the objectives of an imputation methodology are presented as well as the basic concepts of the NIM. In Section 3, some common response errors are described and illustrated by examples. Section 4 presents the major innovation in the editing of couples compared to previous censuses. Also, examples of imputation actions are provided to illustrate how the NIM works. Finally, Section 5 provides some concluding remarks.

More information on the NIM is available in Bankier, Luc, Nadeau and Newcombe (1996).

2. OBJECTIVES AND OVERVIEW OF THE NIM

The objectives of an automated hot deck imputation methodology should be as follows:

(a) The imputed household should closely resemble the failed edit household.

(b) The imputed data for a household should come from a single donor, if possible, rather than two or more donors. In addition, the imputed household should closely resemble that single donor.

(c) Equally good imputation actions, based on the available donors, should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population.

Besides respecting these objectives, the NIM attempts to deal more effectively with the frequent response errors that were not well resolved by the E&I system used in the previous censuses. To achieve this objective, the new methodology was developed in parallel with the modification of the edit rules such that they now reflect more accurately the changes in the Canadian family structures over the last few decades.

The objectives of an imputation methodology are achieved under the NIM by first identifying the passed edit households which are as similar as possible to the failed edit household. This means that the two households should match on as many of the qualitative variables as possible while having small differences between the numeric variables. Households with these characteristics are called nearest neighbours. For each nearest neighbour, the smallest subsets of the non-matching variables which, if imputed, allow the household to pass the edits, are identified. One of these imputation actions which passes the edits and resembles both the failed edit household and the passed edit household is then randomly selected.

The E&I system, called CANEDIT, used in the 1976 to 1991 Censuses, is based on the imputation methodology proposed by Fellegi and Holt (1976). CANEDIT, unlike the NIM, first determined the minimum number of variables to impute and then searched for a nearest neighbour. Reversing the order of the operations allows the NIM to solve larger imputation problems.

In 1996, 1% of the private households of the Atlantic Provinces were households with total non-response while 10% of the households failed because of partial non-response only. Only 2% of the private households of these provinces had one or more variables imputed because of inconsistencies.

3. FREQUENT RESPONSE ERRORS

In the households that failed because of inconsistencies, some common response errors are observed. First, Person 1's spouse is sometimes reported as a son/daughter. In the households with such an error, the difference between the age of Person 1 and the age of the "erroneous" son/daughter, is smaller than the accepted difference between the age of a parent and the age of a child. Another frequent situation, which is not an error but which needs to be dealt with carefully, is when Person 1's spouse is not in position 2 in the household. If it is not possible to identify this person as Person 1's spouse and then make sure that the marital status and common-law status of this person and of Person 1 are appropriate, there could be a loss of legitimate couples. The household displayed in Table 1 illustrates the two problems described. In this household, Person 1's spouse is reported as a son/daughter and, moreover, this person is in position 5.

Table 1: Person 1's spouse Reported as Son/daughter and not in Position 2

Relationship	Marital Status	C-Law Status	Age
Person 1	Divorced	NO	35
Son	Single	NO	8
Daughter	Single	NO	12
Son	Single	NO	15
Daughter	Single	NO	36

For this household, the minimum change imputation action is to change one variable: either the age of Person 1, the age of the last daughter or the relationship of this person. With CANEDIT the age of Person 1 was increased (see Table 2). In this imputed household, there is only a difference of 9 years between the imputed age of Person 1 and the age of the oldest daughter. This is because the decade of birth was used in the edit rules with CANEDIT since it didn't allow the use of numeric variables. In 1996, numeric variables can be used, increasing the precision of the edit rules. For this household, the NIM changed the relationship of the last person to Person 1's husband/wife and also changed the marital status of Person 1 and of the last person to married (see Table 3). More than the minimum number of variables was imputed by the NIM while CANEDIT imputed only one variable. In this situation, imputing the minimum number of variables is not the right decision.

Table 2: CANEDIT Imputation Action for Household of Table 1

Relationship	Marital Status	C-Law Status	Age
Person 1	Divorced	NO	45
Son	Single	NO	8
Daughter	Single	NO	12
Son	Single	NO	15
Daughter	Single	NO	36

Table 3: NIM Imputation Action for Household of Table 1

Relationship	Marital Status	C-Law Status	Age
Person 1	Married	NO	35
Son	Single	NO	8
Daughter	Single	NO	12
Son	Single	NO	15
P1's wife	Married	NO	36

Another frequent situation are single son/daughters-in-law not living in a common-law relationship. These son/daughters-in-law are also sometimes younger than 15 years old. Based on the structure of the households, it is suspected that these son/daughters-in-law are in fact step-son/daughters. The household displayed in Table 4 illustrates this situation.

Table 4: Step-Child Possibly Reported as Son/Daughter-in-law

Relationship	Sex	Marital Status	C-Law Status	Age
Person 1	M	Married	NO	35
P1's wife	F	Married	NO	47
Daughter-in-law	F	Single	NO	24
Son	M	Single	NO	19

In this household, the daughter-in-law in position 3 is single and not living in a common-law relationship. With CANEDIT, couples were not identified (except Person 1's couples and their parents) and the between person edit rules for couples could not be applied because too many rules would have been required. Only within person edit rules were applied, such as "A person reported as a son or daughter-in-law is single and not living in a common-law relationship". The minimum change imputation action was then to change either the marital status, the common-law status or the relationship of this person. In 1991, however, the common-law status variable was deterministically imputed in a prederive module and could not be changed by imputation. Therefore, in 1991, the only two variables that could be imputed for this household to pass this edit rule were the marital status and the relationship to Person 1. If there was more than one minimum set of variables to impute, CANEDIT selected one of them at random. Therefore, in this situation, each of the two variables had one chance in two of being imputed, independently of the plausibility of the resulting responses. For this household, CANEDIT changed the marital status of the daughter-in-law to married. This imputation action is illustrated in Table 5. The imputed household had a rare combination of responses and CANEDIT had, in this way, inflated a small group in the population.

Table 5: CANEDIT Imputation Action for Household of Table 4

Relationship	Sex	Marital Status	C-Law Status	Age
Person 1	M	Married	NO	35
P1's wife	F	Married	NO	47
Daughter-in-law	F	**Married**	NO	24
Son	M	Single	NO	19

On the other hand, with the NIM, couples are identified prior to imputation, as will be explained in the next section, and then couple edit rules can be applied. Therefore, if the son and the daughter-in-law are viewed as a potential couple, the only minimum change imputation action is to change the daughter-in-law to a daughter which is what NIM did (see Table 6). This is more plausible than the imputation action selected by CANEDIT.

Table 6: NIM Imputation Action for Household of Table 4

Relationship	Sex	Marital Status	C-Law Status	Age
Person 1	M	Married	NO	35
P1's wife	F	Married	NO	47
Daughter	F	Single	NO	24
Son	M	Single	NO	19

In a general way, this household illustrates the usefulness of the couple edit rules in the editing of couples who have non-unique relationships to Person 1. The next household is another example of this situation.

Table 7: Household with Couples with Non-unique Relationships to Person 1

Relationship	Sex	Marital Status	C-Law Status	Age
Person 1	M	Married	NO	56
P1's wife	F	Married	NO	55
Son	M	Married	NO	32
Son	M	Married	NO	34
-	F	Married	NO	30
-	F	Married	NO	26
Son	M	Married	NO	30

This household presents a complex situation because there are three sons married and two married women. Therefore there are many possible pairs of persons that could form couples. The persons the most likely to be couples should be identified and they must have appropriate marital statuses and common-law statuses after imputation. It is therefore necessary to have a strategy to deal with this problem. The solution developed, which is a 2-step process, is presented in the next section.

4. THE E&I SYSTEM: A 2-STEP PROCESS

The first step is a prederive module, called REORDER7, in which potential couples are identified prior to imputation. The second step is the hot deck imputation where couple edit rules are applied to the potential couples to confirm whether these pairs are, in fact, couples.

4.1 REORDER 7

Initially a score is assigned to each possible pair of persons in the household based on the unimputed responses to all the demographic variables. For a N person household, a score is assigned to each of the N*(N-1)/2 possible pairs. Any pair with a score less than a fixed parameter are dropped because it is felt that there is not sufficient evidence to form this pair into a couple. Among the pairs remaining, the pairs with the highest scores are identified and a maximum of [N/2] pairs are retained, where a person can belong to only one potential couple. These couples retained are identified by a person level variable, COUPLE, that is set to the same value for the two persons of a specific couple so the couple can be recognized by the NIM. Finally, a subsequent review of the couples formed is executed by applying a set of rules to each of the [N/2] couples. It is then decided to retain or not retain each couple. This decision is based on the score of the couple and on the relationships of the two persons of the couple:

(a) If the relationships are appropriate for a couple and if they necessarily imply a couple (for example a father and a mother), then the couple is retained. Otherwise, if the relationships don't necessarily imply a couple (for example two grandparents), then the couple is retained only if the score is high enough, that is if the other demographic variables strongly suggest that the two persons form a couple.

(b) If the relationships are not appropriate for a couple (for example a brother and a roommate) and if the score is not high enough then the couple is not retained. However, if the relationships are not appropriate for a couple but if the score is high enough, then the couple is usually retained. If one person is Person 1 and the other is not Person 1's spouse then this second relationship is set to blank. If one person in the couple is related to Person 1 but the other person is not, the second relationship is set to blank.

Couple edit rules are then applied in the NIM but only to the couples retained after the above set of rules has been applied to the [N/2] couples. REORDER7 reduces the number of NIM edit rules required because the between person edits in the NIM are applied only to the couples identified by REORDER7.

If REORDER 7 blanks out a relationship, this forces the NIM to impute a value. If the imputed value results in the relationship being appropriate for a couple, the NIM will apply the couple edit rules to determine if other demographic variables have to be imputed to be consistent with that pair being a couple. Depending on the number of variables that have to be imputed, the pair may or may not be retained as a couple by the NIM.

The fact that relationships are set to blank is a form of deterministic imputation, where the "deterministic action" is to blank out rather than to impute variables. This allows more plausible imputation actions, through the imputation of more than the minimum number of variables. This combination of deterministic and hot deck imputation is applicable to a wide range of surveys, the common concept being the development of a strategy to determine the optimal variables to blank out so as to guide the hot deck imputation to the most plausible imputation action.

Such a strategy is now illustrated by a modification that could be made to the REORDER7 module for the 2001 Census. Since the NIM performs minimum change imputation, it will tend to eliminate couples who give no indication that they are legally married or in a common-law relationship. For example, for the household of Table 8, the NIM will generally change the relationship of the second person rather than change the common-law status of the two persons, because only one variable is imputed instead of two.

Table 8: Failed Edit Household

Relationship	Marital Status	C-Law Status
Person 1	Separated	NO
C-L Spouse of P1	Separated	NO

In this situation, where it is believed that these two persons possibly form a couple, blanking out variables would increase the chance of the couple being retained by the NIM. When considering how many variables to blank out, the minimum change approach should still be taken. Moreover, it is believed that relationship is more accurately answered than common-law status. Therefore, one possibility would be to blank out the common-law status NO of the second person as shown below:

Table 9: Failed Edit Household

Relationship	Marital Status	C-Law Status
Person 1	Separated	NO
C-L Spouse of P1	Separated	-

From the perspective of minimum change, the two following imputation actions would then be equally attractive because two variables are being imputed in either case.

Table 10: Couple Preserved

Relationship	Marital Status	C-Law Status
Person 1	Separated	YES
C-L Spouse of P1	Separated	YES

Table 11: Couple Eliminated

Relationship	Marital Status	C-Law Status
Person 1	Separated	NO
Brother/Sister	Separated	NO

Thus the frequency with which the common-law couple will be retained will be based on the frequency with which such couples appear among the donors. This way of improving hot deck imputation by deterministically blanking out variables prior to imputation could be applied to a broader range of imputation problems.

4.2 Couple Edit Rules Applied in the NIM

Examples of couple edit rules applied in the NIM to the couples identified by REORDER7 are illustrated in Table 12 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 and a brother/sister-in-law).

Table 12 : 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
relat#1=S/D	Y	Y	Y	Y	Y	Y	Y	Y	N
relat#2=S/D-in-law	Y	Y	Y	Y	Y	Y	Y	N	Y
sex#1 = sex#2	Y								
marital status#1= married		Y	N			N			
marital status#2= married		N	Y				N		
c-law status#1=yes				Y	N	N		Y	
c-law status#2=yes				N	Y		N		Y

The rules illustrated in Table 12 are generated by the NIM Edit Interface for all the combinations of two persons in the household. The first proposition (couple#1=couple#2) ensures that the rules are applied only to the couples identified by REORDER7. The quantities "#1" and "#2" represent any combination of two persons in the household.

These edit rules ensure that, after imputation, the two persons of a couple are opposite in sex and that both are married or both have common-law status YES.

In fact, if two persons with appropriate relationships for a couple, and identified as a couple by REORDER 7, fail one of these edit rules, there are two possible outcomes: either the variables that caused the household to fail the edit rule 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 appropriate for a couple. The pair will then not be considered any longer as a couple.

The first edit rule of Table 12 causes couples formed by two persons of the same sex not to be retained. If, however, it is explicitly mentioned that a person is living in a same-sex relationship with another person, this person is deterministically changed to a roommate. Otherwise, if two persons with appropriate relationships have the same sex, it is possible (and surprisingly frequent) that the sex is misreported. Therefore, depending on the donors available, the NIM either changes the sex or changes the relationship so the two persons don't form a couple after imputation.

The impact of these couple edit rules is illustrated in the next section with a sample of households that represents about 1/5 of all the private households in Canada. The households included in this sample are the households that received a long form questionnaire in the four regions of Canada: East (Atlantic provinces), Quebec, Ontario and West.

4.3 Illustration of the Impact of REORDER7 and of the Couple Edit Rules

To study the effect of the identification of couples prior to imputation and of the application of couple edit rules to the couples identified, the "son/daughter - son/daughter's partner" couples were studied for a sample of private households.

The couples studied have either the two relationships present after REORDER7 or have one relationship missing, the other one being either son/daughter, step-son/daughter, son/daughter-in-law or common-law partner of son/daughter. These couples with a blank relationship are considered because they can be identified by REORDER7 and are potential "son/daughter - son/daughter's partner" couples after imputation. There are in total 22,350 couples in the category "son/daughter - son/daughter's partner" couples identified by REORDER7, of which 86.7% are "son/daughter - son/daughter-in-law" couples. The other types of couples in this category are the "son/daughter - common-law partner of son/daughter" couples, the "step-son/daughter - son/daughter-in-law" couples, the "step-son/daughter - common-law partner of son/daughter" couples, and finally the couples with one relationship missing as mentioned previously.

Of these 22,350 couples identified by REORDER7, 83% were retained by imputation. The fact that a couple is preserved or not by imputation is related to the responses to the other variables. For 98.6% of the 18,522 couples retained, both persons are married or both have common-law status YES before imputation. In addition, for 98% of the couples retained, both persons are older than 15 years old before imputation. The responses to these variables thus indicate that the persons form a couple. On the other hand, for 90% of the couples not retained by the NIM, both persons are not married and both have common-law status NO or missing before imputation. Finally, for 58% of the couples not retained, at least one person is less than 15 years old before imputation. An example of a couple not retained by the NIM is given in the next table.

Table 13: Example of a Couple Eliminated

Relationship	Marital Status	C-Law Status	Age	NIM
Person 1	Widowed	NO	48	
Son	Single	NO	27	
Son-in-law	Single	NO	25	--> Son
Daughter	Single	YES	19	
Son-in-law	Single	YES	18	

In this household, the son in position 2 and the son-in-law in position 3 were identified as a potential couple by REORDER 7. To form a couple with these two persons, the NIM had to change three variables: the sex of one person and also either the marital status or the common-law status of the two persons. On the other hand, if the relationship of the person in position 3 is changed, the two persons don't form a couple after imputation but the household passes the edit. Therefore, since the other variables for the persons in positions 2 and 3 do not indicate that these two persons form a couple, the son-in-law in position 3 was changed to a son by the NIM. In the next example a couple was created by the NIM.

Table 14: Example of a Couple Created

Relationship	Marital Status	C-Law Status	Age	NIM
Person 1	Widowed	NO	72	
Son	Single	NO	31	
P1's CLP	Single	YES	33	-> Son-in-law
Daughter	Single	YES	33	
Grandchild	Single	NO	10	

In this household, Person 1 is widowed and is not living in a common-law relationship, but the person in position 3 is reported as Person 1's common-law partner. This person is followed by person living in a common-law relationship. The person in position 4 is reported as the daughter of Person 1 and has the same age as the person reported as Person 1's common-law partner. The NIM then changed the Person 1's common-law partner to a son-in-law, which is a plausible imputation action.

In the examples of Tables 13 and 14, no relationships were missing. In fact, these "son/daughter - son/daughter's partner" couples identified by REORDER7 with no relationship missing represent 97% of the 22,350 "son/daughter - son/daughter's partner" couples identified by REORDER7. Therefore, for this category of couples, only 3% of the couples identified had a missing relationship. However these couples with a blank relationship illustrate an important feature of REORDER7: the possibility of blanking out a non-appropriate relationship for a couple if the other variables indicate that two persons form a couple.

There are 686 couples with a blank relationship identified by REORDER7. 79% of the relationships were present before REORDER7 but were set of blank at this stage. As mentioned in Section 4.1, a relationship is set to blank only if it is not related to Person 1, except when the other person of the couple is Person 1. In this case, if the partner is not Person 1's spouse, but the other variables are appropriate for a couple, then the relationship is set to blank. The relationships lodger and roommate represent 86% of the relationships set to blank. An example of a household where a lodger is set to blank is given in Table 15.

Table 15: Household where Lodger set to Blank by REORDER7

Relationship	Marital Status	C-Law Status	Age	Reorder7	NIM
Person 1	Single	YES	53		
P1's CLP	Single	YES	53		
Lodger	Single	YES	32	-> bk	->son-in-law
Daughter	Single	YES	23		
Grandchild	Single	NO	7		

In this household the persons in positions 3 and 4 are opposite in sex, have appropriate ages for a couple but one is the daughter of Person 1 while the other one is reported as a lodger. These two persons were identified as a couple by REORDER7 because all the variables are appropriate for a couple except the relationships. Since one relationship is related to Person 1 and the other is not, the relationship not related to Person 1 is set to blank by REORDER7 to allow the NIM to impute an appropriate value. The NIM then imputed a son-in-law which is plausible considering the structure of the household.

To evaluate the relative importance of the identification of couples prior to imputation it is also important to examine the couples present in the households after the Edit and Imputation process.

There are 18,756 "son/daughter - son/daughter's partner" couples after imputation. Of these couples, 99% were identified by REORDER7. Most of these couples after imputation (86%) didn't have any variable changed. For 10% of the couples present after imputation, the relationships were appropriate for a couple before imputation but another variable was imputed, either because of non-response or because of inconsistencies. Finally, for 4% of the couples after imputation, at least one relationship was imputed, again either because of non-response or because of inconsistencies. REORDER7 and the NIM, therefore, had some impact on about 14% of the "son/daughter - son/daughter's partner" couples in this sample.

5. CONCLUDING REMARKS

One of the major innovation of the NIM is the identification of couples followed by the minimum change imputation of the demographic variables. This process was computationally feasible and effective and contributed to increase the data quality. This combination of deterministic and hot deck imputation is applicable to a wide range of surveys and censuses.

For the 2001 Census, families, instead of just couples, may be identified prior to imputation. This would allow more detailed edit rules to be applied. In addition, the NIM will be generalized for the 2001 Census so it can process a wider selection of variables.

REFERENCES

Bankier, M., Luc, M., Nadeau, C. and Newcombe, P. (1996), "Imputing Numeric and Qualitative Census Variables Simultaneously", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1996.

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.