

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 5 of the provisional agenda

1996 CANADIAN CENSUS DEMOGRAPHIC VARIABLES IMPUTATION

Submitted by Statistics Canada¹

¹ Prepared by Michael Bankier, Anne-Marie Houle and Manchi Luc.

KEYWORDS: Minimum Change Hot Deck Imputation, Inconsistent Responses, Couple Edit Rules

I. INTRODUCTION

1. Among the basic questions asked to 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 the household to identify missing and inconsistent responses.

2. 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. 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 a major innovation in 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).

II. OBJECTIVES AND OVERVIEW OF THE NIM

3. 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.

4. Besides respecting these objectives, the NIM attempts to deal more effectively with certain 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 to the modification of the edit rules that now reflect more accurately the changes in the Canadian family structures over the last decades.

5. 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 close to each other or nearest neighbours.

Then for each nearest neighbour, the smallest subsets of the non-matching variables (both numeric and qualitative) 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.

6. 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.

7. Table 1 presents some 1996 Census results for the private households of the Atlantic Provinces. 87% of the private households passed the edit rules. Conversely, 13% of the households failed at least one edit rule. 10% of the households failed because of partial non-response only. Only 2% of the households had one or more variables imputed because of inconsistencies.

Table 1: Private Households of the Atlantic Provinces

Private Households	Proportion
Passed Edit Households	87.1%
Failed Edit Households	
Total Non-Response	0.6%
Partial Non-Response only	10.2%
Inconsistencies only	1.7%
Non-Resp. and inconsistencies	0.5%

III. FREQUENT ERRORS MADE BY RESPONDENTS

8. In the households that failed because of inconsistencies, some common response errors were observed. First, step-children are sometimes erroneously reported as sons or daughters-in-law. Consequently, the presence of single son/daughters-in-law not living in a common-law relationship is frequent. The household displayed in Table 2 illustrates this situation. Tables 3 and 4 present this household imputed by CANEDIT and by the NIM.

Table 2: Step-Child Reported as Son/daughter-in-law

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

9. The household in Table 2 failed because the person in position 3 is reported as a single daughter-in-law not living in a common-law relationship. The minimum change imputation action is to change either the marital status, the common-law status or the relationship of this person. If there is more than one minimum set of variables to impute, CANEDIT selected one of them at random. Therefore, in this situation, the marital status or the common-law status would have been imputed 2/3 of the time, independently of the plausibility of the resulting responses. For this household, CANEDIT imputed the marital status of the daughter-in-law to married. This imputation action is illustrated in Table 3. The imputed household thus presents a rare combination of responses and CANEDIT had, in this way inflated a small group in the population.

Table 3: CANEDIT Imputation Action

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

10. On the other hand, the imputation action selected by the NIM is based on the frequency with which each possible imputation action appears among the donors because the NIM first identifies the donors similar to the failed edit households. For this household, the NIM imputation action was to change the daughter-in-law to a daughter (see Table 4). This is a more plausible imputation action than the one selected by CANEDIT.

Table 4: NIM Imputation Action

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

11. Another common response error is Person 1's spouse 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.

12. 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.

13. The household displayed in Table 5 illustrates these last two situations described. In this household, Person 1's spouse is reported as a son/daughter and, moreover, this person is not in position 2.

Table 5: Person 1's spouse Reported as Son/daughter and not in Position 2

Relationship	Marital Status	Age
Person 1	Divorced	35
Son	Single	8
Daughter	Single	12
Son	Single	15
Daughter	Single	36

14. For this household, the minimum change imputation action is to change one variable: either the age of Person 1, the age of the oldest son/daughter or the relationship of the last person. With CANEDIT the age of Person 1 was increased. The household imputed by CANEDIT is presented in Table 6. There is only a difference of 9 years between the imputed age of Person 1 and the age of the oldest son/daughter. This is because the decade of birth was used in the edit rules in 1991 since CANEDIT did not allow the use of numeric variables. The edit rule that failed was "The decade of birth for a son or daughter is the same or precedes the decade of birth reported for Person 1". In 1996, with the NIM, numeric variables can be used in the edit rules and the household now fails the rule "The difference between the age of a parent and the age of a child is less than 15 years". However, in the household of Table 5, 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. This imputation action is presented in Table 7. More than the minimum number of variables was imputed by the NIM while CANEDIT imputed only one variable. This household illustrates a situation for which imputing the minimum number of variables is not the right decision.

Table 6: CANEDIT Imputation Action for Failed Edit Household of Table 5

Relationship	Marital Status	Age
Person 1	Divorced	45
Son	Single	8
Daughter	Single	12
Son	Single	15
Daughter	Single	36

Table 7: NIM Imputation Action for Failed Edit Household of Table 5

Relationship	Marital Status	Age
Person 1	Married	35
Son	Single	8
Daughter	Single	12
Son	Single	15
P1's Husband/wife	Married	36

15. A more general problem is the editing of couples (either legally married or living in a common-law relationship) who have non-unique relationship to Person 1. The household displayed in Table 8 illustrates this situation.

Table 8: 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
Son	M	Single	-	30
-	F	Married	NO	26
-	F	Married	NO	30
Son	M	Married	NO	27

16. 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. In fact, if only the last six persons are considered (the persons in positions 1 and 2 already form a couple), and if the sex is disregarded, there are 15 possible pairs of persons. The persons the most likely to form couples should be identified and they must have appropriate marital statuses and common-law statuses after imputation. A very large number of edit rules would be required to handle this type of situation. This solution is feasible for smaller households but, with larger households, the thousands of edits required would be difficult for the NIM to process. It was therefore necessary to have another strategy to deal with this problem. The solution developed is a 2-step process. This solution is presented in the next section.

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

17. 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.

Reorder

18. Initially, in REORDER7, a score is assigned to each possible pair of persons in the household based on the unimputed responses to all the demographic variables. For an N person household, a score is assigned to each of the $N*(N-1)/2$ possible pairs. Secondly, the pairs with the highest scores are identified and a maximum of $[N/2]$ pairs are retained. These couples retained are identified by a person level variable COUPLE## (## is the position of the person in the household) 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. The four following actions are possible:

- (a) If the relationships are appropriate for a couple and these relationships necessarily imply a couple (for example a father and a mother), then the couple is retained;
- (b) If the relationships are appropriate for a couple but 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;
- (c) If the relationships are not appropriate for a couple and the score is not high enough then the couple is not retained;
- (d) If the relationships are not appropriate for a couple but the score is high enough then the couple is usually retained. If one person in the couple 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 person's relationship is set to blank. Blanking out relationships increases the chance that the NIM will impute an appropriate value.

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.

19. Therefore 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. REORDER7 also blanks out some suspect relationships when all the other variables suggest that the two persons form a couple, so that the NIM can possibly impute an appropriate relationship. Consequently REORDER7 allows correct imputation actions to be achieved in cases where more than the minimum number of variables have to be imputed.

Couple Edit Rules Applied in the NIM

20. Some of the couple edit rules applied in the NIM to the couples identified by REORDER7 are illustrated in Table 9 for the "son/daughter - son/daughter-in-law" couples. In this table, the rules are listed in the nine columns while the positions entering these rules

are listed in the left most column.

21. The first proposition (couple#1=couple#2) ensures that the rules are applied only to the couples identified by REORDER7. The two persons of a couple are identified by “#1” and “#2”, where “#1” and “#2” can be any combination of two persons in the household. The purpose of these edit rules is to ensure that the two persons of a couple are opposite in sex and that both are married or both have common-law status YES. This set of edit rules fails the households that don’t respect these conditions. For example, if one person is a son/daughter and the other is a son/daughter-in-law then the household fails if one person is married and the other is not married (rules 2 and 3), or if one person is living in a common-law relationship and the other is not (rules 4 and 5). These rules are generated by the NIM Edit Interface for all the combinations of two persons in the household. Similar rules exist for pairs of persons with other relationships that could form couples.

Table 9 : 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 impact of these new rules is illustrated in the next section with a sample of households that represents about 1/5 of all the private households in Canada.

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

22. To demonstrate the effect of the identification of potential 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.

23. There are four types of couples in this category. There are, of course, the “son/daughter - son/daughter-in-law” couples, but also the “son/daughter - common-law

partner of son/daughter” couples, and finally the “step-son/daughter - son/daughter-in-law” and the “step-son/daughter-common-law partner of son /daughter” couples. In addition, in the study, the couples with one of the four relationships (son/daughter, step-son/daughter, son/daughter-in-law and common-law partner of son/daughter) and a blank relationship are also considered because these couples can be identified by REORDER7 and are potential “son/daughter - son/daughter’s partner” couples after imputation. There are therefore eight types of couples considered. The number of couples identified by REORDER7 are given in Table 10.

Table 10: “Son/daughter - Son/daughter’s partner” Couples Identified by REORDER7

Type of Couple	Number	Proportion
S/D and S/D-in-law	19,384	86.7%
S/D and S/D’s CLP	2,124	9.5%
Step-S/D and S/D-in-law	99	0.4%
Step-S/D and S/D’s CLP	57	0.3%
S/D and blank	564	2.5%
Step-S/D and blank	20	0.1%
S/D-in-law and blank	91	0.4%
S/D’s CLP and blank	11	0.1%
TOTAL	22,350	100.0%

24. These 22,350 couples identified by REORDER7 were either retained or eliminated by imputation. If the data for a household, plus the data present among the donors, supports a pair being a couple then the NIM retains the couple by imputing appropriate values for the different variables if required. On the other hand, if there is not sufficient indication from the data that a pair forms a couple, the NIM does not impute appropriate values for a couple.

25. To study the proportions of couples retained and eliminated by imputation, the eight types of couples are split into three groups. First, the couples where the two relationships are present and one person is son/daughter-in-law (in other words, a son/daughter or a step son/daughter with a son/daughter-in-law) are considered. In this group, 82% of the couples were retained by the NIM. Secondly, the couples where the two relationships are present and one is a common-law partner of son/daughter (in other words, a son/daughter or a step son/daughter with a common-law partner of son/daughter) are considered. It was found that 94% of them were retained by imputation. Finally, 84% of the couples where one relationship is blank were retained.

26. There is an important difference between the proportion of couples retained when the partner is a son/daughter-in-law and when the partner is a common-law partner of the son/daughter. This is explained by the fact that, as mentioned in Section 3, many step children are reported as son/daughters-in-law. If a son/daughter is present in the household, REORDER7 may have identified the son/daughter-in-law and a son/daughter as a couple because these relationships are appropriate for a couple. In this case, the other demographic variables are often not appropriate for a couple and the NIM didn’t retain the couple. On the

other hand, when the partner is a common-law partner of son/daughter, usually the relationship is correctly reported and often the other variables also suggest that the two persons form a couple. Consequently the NIM retained most of them.

27. The fact that a couple is preserved or not by imputation is directly related to the responses to the other demographic 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. 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. And finally, for 58% of the couples not retained, at least one person is less than 15 years old. An example of a couple not retained by the NIM is given in the next table.

Table 11: 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	

28. In the household displayed in Table 11, the son-in-law in position 3 was changed to a son by the NIM because the other variables for the persons in positions 2 and 3 do not indicate that these two persons form a couple. In the next example a couple was created by the NIM.

Table 12: 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	

29. In this household, Person 1 is widowed and is not living in a common-law relationship. Nevertheless, the person in position 3 is reported as Person 1's common-law partner. This person is followed by another 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.

30. In the examples of Tables 11 and 12, no relationships were missing. In fact, these “son/daughter - son/daughter’s partner” couples identified by REORDER7 with a relationship missing represent 97% of all the “son/daughter - son/daughter’s partner” couples identified by REORDER7 (see Table 10). That is to say that, in this category of couples, there is only 3% of the couples identified that had a missing relationship. However some of these couples with a blank relationship illustrate an important feature of REORDER7: the possibility of changing a non-appropriate relationship for a couple to blank if the other variables indicate that two persons form a couple.

31. There are 686 couples with a blank relationship identified by REORDER7 (see Table 10). These blank relationships were either missing or present before REORDER7. Actually, 79% of these relationships (544 relationships) were present before REORDER7 but were set to 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. Most of the relationships set to blank by REORDER7 were either lodger (58.8%) or roommate (27.2%). Therefore, these two relationships represent 86% of the relationships set to blank. An example of a household where a lodger is set to blank is given in Table 13.

Table 13: 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	--> blank	--> son-in-law
Daughter	Single	YES	23		
Grandchild	Single	NO	7		

32. This is a 5 person household where the persons in positions 1 and 2, and the persons in positions 3 and 4, are living in a common-law relationship. 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.

33. So far, the study was focussing on the couples identified by REORDER7. 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.

34. There are 18,756 “son/daughter - son/daughter’s partner” couples after imputation. Of these couples, 99% were identified by REORDER7. However, most of these couples after imputation (86%) didn’t have any variables changed. On the other hand, for 10% of the couples 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, therefore, had some impact on about 14% of the "son/daughter - son/daughter's partner" couples in this sample.

V. CONCLUDING REMARKS

35. The identification of couples followed by the minimum change imputation of the demographic variables was computationally feasible and effective. This New Imputation Methodology is applicable to a wide range of surveys and censuses, and the NIM itself will be generalized for the 2001 Canadian 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.