STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 25
English Only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(Helsinki, Finland, 27-29 May 2002)

IMPUTATION OF DEMOGRAPHIC VARIABLES FROM THE 2001 CANADIAN CENSUS OF POPULATION

Invited paper

Submitted by Statistics Canada[1]

Abstract: CANCEIS (CANadian Census Edit and Imputation System) is a generalised Edit and Imputation (E&I) system that was used in February 2002 to process the demographic variables from the 2001 Census. Later in 2002, it will perform E&I on the labour, mobility, place of work and mode of transport variables as well. It performs minimum change imputation on a mixture of qualitative and quantitative variables for a given nearest neighbour donor. In this paper, an overview of the E&I process for the demographic variables is given. In addition, it is emphasized that it is not always appropriate to impute the minimum number of variables. Also, the impact of modifying certain parameters in the distance measures that determine what are the optimal imputation actions is reviewed. The advantages of processing on the Personal Computer (PC) versus the mainframe are examined. Finally, the operational feasibility and the organisational issues associated with the implementation of CANCEIS are considered.

I. INTRODUCTION

1. 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 nonresponse 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.

2. 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

[1] Prepared by Michael Bankier, Patrick Mason and Paul Poirier, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. Mike.Bankier@statcan.ca. Revised May 14, 2002.

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.

3. 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. In addition, CANCEIS has been used by the Canadian Census of Agriculture Coverage Evaluation Survey and will be used this summer by the Canadian Survey of Household Spending.

4. CANCEIS (or an earlier version of the software) will be used to process some of the variables in the 2001 Ukrainian Census, the 2000 Brazilian Census and the 2001 Swiss Census. In addition, the 2001 Italian Census, having studied CANCEIS, will use a similar approach in their imputation methodology.

5. Section II describes how the demographic E&I was performed in the 2001 Canadian Census using CANCEIS. Section III explains how CANCEIS parameters were selected for the demographic data. Section IV highlights the performance of CANCEIS. Section V provides an overview of the organisational issues associated with implementing an E&I system such as CANCEIS. Finally, some concluding remarks are given in Section VI. For more details regarding the NIM and CANCEIS , see Bankier, Lachance and Poirier (2000a , 2000b).

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

6. 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.

7. 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

8. 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.

9. For 2001, there were several systematic errors that were corrected during production. One systematic error occurs when a child completes the questionnaire for their parents and put themselves in position 3. The child reports all the Relationship to Person 1 responses using themselves as the reference person. An example is provided in Table 1 below. This often occurs when an older child fills in the questionnaire for their parents because the parents do not have as good a grasp of the English or French language as their child.

10. To correct for this, the household in Table 1 is reordered by moving the child into position 1 and the original Person 1 into position 2. In addition, their relationships are changed to Person 1 and Father. The household can now pass the edits without further imputation. If the systematic error was not detected and corrected prior to minimum change imputation, the resulting household would look dramatically different as can be seen in Table 1. With minimum change imputation, the family in this household, which initially resembled a 2 parent, 3 children family, would become a 1 parent, 4 children family. The correction of the systematic reporting error, in contrast, preserves the initial family structure.

Table 1: Reference Person for Relationship in Position 3 instead of Position 1

Position on questionnaire	Relationship to Person 1	Age	Sex	Marital Status	Common-Law Status
1	Person 1	55	Male	Married	No
2	Mother	51	Female	Married	No
3	Blank	27	Female	Single	No
4	Sister	24	Female	Single	No
5	Nephew	2	Male	Single	No
Data after correction of systematic reporting error					
»3	**»Person 1**	27	Female	Single	No
»1	**»Father**	55	Male	Married	No
»2	Mother	51	Female	Married	No
4	Sister	24	Female	Single	No
5	Nephew	2	Male	Single	No
Data after minimum change correction of systematic reporting error					
1	Person 1	**»25**	Male	Married	No
2	Mother	51	Female	Married	No
3	**»Sister**	27	Female	Single	No
4	Sister	24	Female	Single	No
5	Nephew	2	Male	Single	No

11. Another example of a systematic reporting error is where everyone in a family reports a common-law status of yes. An example is given in Table 2 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 reporting error by imputing the Common-Law Statuses for persons 3, 4, and 5.

12. Deterministic imputation is also utilised during the Couple formation process (see part B of this section). In this process, sometimes "blanking out" either the Relationship, the Common-Law Status or the Marital Status variables is advantageous in order to induce imputation by CANCEIS. For example, if there is sufficient evidence that this pair of persons is likely a couple and one of the relationships is misreported, then the misreported relationship is blanked out. This is only done when one person is related to Person 1 but the other person is not (such as a lodger or roommate).

Table 2: Systematic Reporting Error of Common-Law Status Variable

Position on questionnaire	Relationship to Person 1	Age	Sex	Marital Status	Common-Law Status
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**
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**

13. The household presented in Table 3 illustrates how the process works. In this household, the persons in positions 3 and 4 are opposite in sex, have appropriate ages and marital statuses for a couple, but one is the daughter of Person1 while the other one is reported as a Lodger. These two persons will be identified as a couple. Since the daughter is related to Person 1 and the lodger is not, the lodger relationship is set to Blank to allow CANCEIS the choice, through imputation, of keeping this pair as a couple or not.

Table 3: Blanking Out a Relationship to Force a Plausible Imputation

Position on questionnaire	Relationship to Person 1	Age	Sex	Marital Status	Common-Law Status
1	Person 1	45	Male	Married	No
2	Wife	44	Female	Married	No
3	Son/Daughter	21	Female	Married	No
4	**»Lodger → Blank**	23	Male	Married	No
5	Grandchild	3	Male	Single	No

14. Table 4 gives two more examples where it is beneficial to deterministically blank out either common-law status or marital status. With these two examples, the persons' relationships, ages and sexes are consistent with them being couples but no indication of this is given in terms of marital status or common law status. Imputing one variable (the relationship for person 2) rather than two (both marital statuses or both common-law statuses) is the minimum change imputation action but eliminates the couple. By blanking out one response for marital status or one response for common-law status, retaining or discarding the couple is equally attractive and will be based on the frequency with which such couples appear among the donors.

Table 4: Blanking Out Common-Law or Marital Status to Force Plausible Imputations

Position on questionnaire	Relationship to Person 1	Age	Sex	Marital Status	Common-Law Status
1	Person 1	38	Female	Single	No
2	Common-Law Partner	39	Male	Single	**»No → Blank**
1	Person 1	38	Female	Single	No
2	Husband	39	Male	**»Single → Blank**	No

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

15. 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 retained couples are identified by the Couple variable which 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.

16. 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 the unimputed responses of the relationships, the ages and proximity on the questionnaire.

17. The household presented in Table 5 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.

Table 5: 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

18. Similarly, the person in position 1 (i.e. person 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.

C. Editing with CANCEIS

19. 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 6 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). 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.

20. Decision Logic Tables such as the one illustrated in Table 6 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.

21. The Canadian Census requires that a couple be 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 6 ensures that couples respect these conditions. If a potential couple matches 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 appropriate for a couple.

Table 6: Between Person Edit Rules for {Son/daughter, Son/daughter-in-law} Couples

	Rules								
Propositions	**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

22. 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. Because the plausibility is considered during imputation, the NIM often deals effectively with the many reporting errors made by respondents. A brief review of the NIM follows. Additional details are provided in the papers given in the references.

23. The NIM was created with the following objectives:
(a) The imputed household should closely resemble the failed edit household.
(b) The imputed data for a household should come from a single donor household, if possible, rather than two or more donors. In addition, the imputed household should closely resemble that single donor household.
(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.

24. These objectives are achieved with the NIM 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.

25. The notion of similarity is based on a distance function. It will be assumed that F households fail the edits rules while P households pass the edits 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

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

26. 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.

27. The distance function $0 \leq D_i(V_{fi}, V_{pi}) \leq 1$ can be different for each variable i. See Section III for a discussion of the distance function used with the quantitative variable age. 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 7 where it is assumed that the distance is always 0 when $V_{fi} = V_{pi}$.

28. The Common_Law_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 Common_Law_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 7: Example of a Distance Matrix

$V_f \backslash V_p$	Son-in-law	Son	Wife	Common_Law_ Partner_Person_1	All Other Relationships
Common_Law_Partner_of_Daughter	0.25	1	1	1	1
Stepson	1	0.25	1	1	1
Wife_or_Common_Law_Partner_Person_1	1	1	0	0	1
All Other Relationships	1	1	1	1	1

29. The value Wife_or_Common_Law_Partner_of_Person1 indicates that both Wife and Common_Law_Partner 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 Common_Law_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.

30. 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 nonmatching variables (those with $V_{fi} \neq V_{pi}$) are, of course, considered for imputation. Various subsets of these nonmatching 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.

31. 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 the 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.

32. An example of where plausibility is preferred over minimum change is given in Table 8. 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.

33. Of all the imputation actions considered which pass the edit rules, only those which minimise or nearly minimise the weighted distance are retained. The imputation actions retained must satisfy the equation $D_{fpa} \leq \gamma \min D_{fpa}$ where γ is a parameter which is greater than or equal to 1. The γ parameter was set equal to 1.1 for the 2001 Census. These imputation actions are called "near minimum change imputation actions", while the imputation actions with $D_{fpa} = \min D_{fpa}$ are called "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.

34. 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.

Table 8: 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

E. Use of Auxiliary Variables

35. 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. 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.

36. Without the auxiliary variables, it is unclear whether the Son in Table 9 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 will result in the majority of donors having an age greater than 20 for the Son.

Table 9: 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*

37. For the first time, the 2001 Canadian Census used auxiliary variables from long form questionnaires during the imputation of the demographic variables. After testing, it was determined that the variables Highest Grade, Hours Worked and Total Income would be used. 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. With a sample of households, where the addition of these auxiliary variables had an impact on the imputation actions, its use was found to be worthwhile, though the proportion of households where the imputed data was better (57% better vs 43% worse) was not striking.

III . OPTIMISATION OF CANCEIS PARAMETERS

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

39. A simplified version of the Age distance function used in demographic E&I is given by:

$$D_i\left(V_{fi}, V_{pi}\right) = \begin{cases} 1 & if \quad \left|V_{fi} - V_{pi}\right| \geq 6 \\ 1 - (1 - \left|V_{fi} - V_{pi}\right| / 6)^r & otherwise \end{cases}$$

where $r \geq 0$. In Chart 1, the age distance function is given for different values of r. With the demographic data, it is sometimes better to change two ages through imputation by a "small amount" (as represented by a small value in the age distance measure) rather than to change a single qualitative variable. Increasing the value of r will decrease what is considered a "small amount". For example, with r = ¼ (the value used in 1996), an age difference of 3 is considered a change of a "small amount" since the distance measure is only 0.15 while with r = 2 (the value used in 2001), an age difference of 3 is no longer considered a change of a "small amount" since the distance measure is now 0.75. The impacts of changing *r* is illustrated in Table 10.

40. When r = ¼, Imputation Action #1 in Table 10, where two ages are changed, will be preferred to Imputation Action #2 where one qualitative variable is imputed. This is because Imputation Action #1 has the smaller distance D_{fpa}. When r = 2, Imputation Action #1 has the larger distance and hence imputing the one qualitative variable of Imputation Action #2 is preferred.

41. The α parameter in the D_{fpa} function was also examined. As the value of α is lowered, it places more emphasis on plausibility (i.e. minimising $D(V_a, V_p)$) rather than the number of variables changed (i.e. minimising $D(V_f, V_a)$).

Chart 1: Age distance function

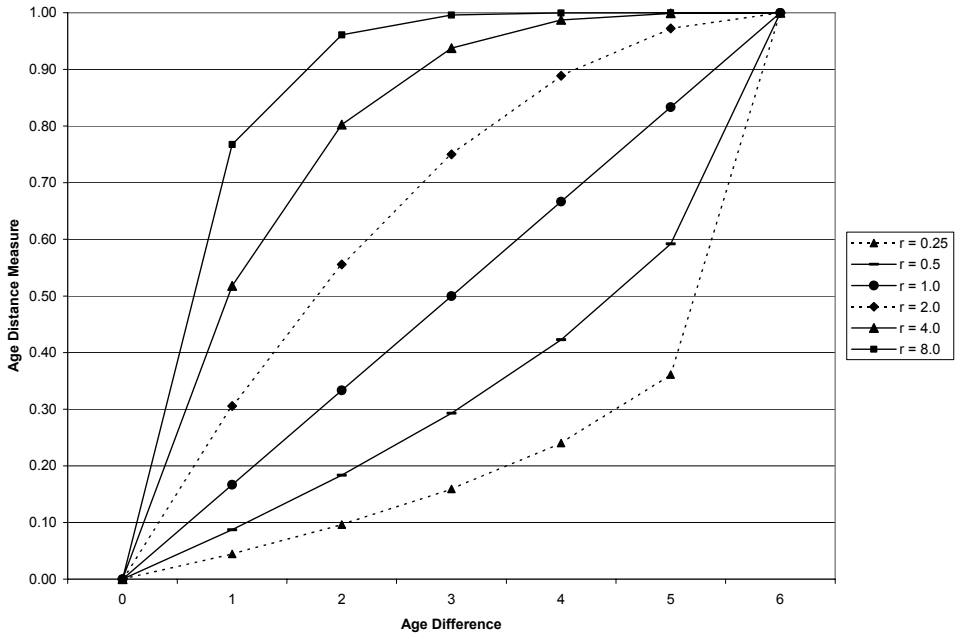


Table 10: Age Distance Function for two *r*-Values

	Unimputed			Imputed	
	Relationship to Person 1	**Age**		**Relationship To Person 1**	**Age**
Imputation Action #1	Person 1	50		Person 1	**»45**
	Wife	43		Wife	43
	Father	56		Father	**»60**
Imputation Action #2	Person1	50		Person 1	50
	Wife	43		Wife	43
	Father	56		**»Brother**	56

42. In the 1996 Census, $\alpha = 0.9$ was used. For the 2001 Census it was decided to continue the use of $\alpha = 0.9$ for the smaller household sizes, but use $\alpha = 0.75$ 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 11 (the failed edit household), Table 12 (using potential donor #1) and Table 13 (using potential donor #2).

Table 11: 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

43. In Table 12, 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.

Table 12: Potential Donor #1

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
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**

44. In Table 13, 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 13: Potential Donor #2

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**
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. PERFORMANCE OF CANCEIS

A. Mainframe vs PC

45. The Census of Population E&I has used a mainframe for its computing environment since the beginnings of automated Census E&I in 1971. For the 2001 Census of Population, it was planned to continue to use the mainframe based on the assumption that this would result in processing being completed in the shortest time possible. It was also decided to develop CANCEIS on the PC and later port the software to the mainframe. To the surprise of most people involved, early versions of CANCEIS ran between 2 and 5 times faster on the PC than on our mainframe. The time difference was found to depend on whether the process was more I/O intensive (where the PC does a lot better than the mainframe) or more CPU intensive (where the mainframe processing time is similar to the PC). Recent tests with a more powerful PC (a Pentium IV 1.7GHz was used) have shown that the difference is now of the order of 3 to 8 times faster on the PC. Many attempts to improve the mainframe version to reduce the time difference were unsuccessful.

46. Since the assumption that the mainframe was faster than the PC was not true, a decision was made to evaluate the possibility of processing the data with CANCEIS on PCs, but keep the Census data on the mainframe for more secure storage. The advantages of running on the mainframe were found to be few. Among them is the extra security, the fewer number of file transfers between platforms, the storage capacity, the backup facility and the stability. The advantages of running on the PC were numerous. Among them is the speed, the cost ("free" PCs in our offices could be used), the increased flexibility, the data quality (multiple runs could be done to fine tune the process and more donors could be considered given the increased speed), the possibility of running many PCs in parallel (e.g. could run twice as fast with two machines), the programming/debugging/analysis tools are superior, the independence (our mainframe is shared with many surveys), the user-friendliness and the interest of the staff. The interest of the staff is more of a long-term impact. The mainframe is viewed as a dinosaur and working on it is not perceived as a career opportunity. Because of this, it has been hard to find people interested in working in this environment. Staff recruitment and retention for methodologists, computer programmers and technical production officers has been problematic for mainframe related projects. Given these facts, the decision was made to use the PCs to process all variables using CANCEIS for the 2001 Census.

B. Highlights of Demography E&I Production

47. E&I of the 2001 Census Demographic variables (for both the short and long forms) was carried out successfully with CANCEIS in February 2002. With short forms, CANCEIS processed 30 million persons with up to 14,000 edit rules. With long forms, CANCEIS processed some 6 million persons with up to 43,000 edit rules. All processing (short and long forms) was done on typical PCs and CANCEIS ran flawlessly over a 4 day period. Traditionally, demographic data processing was done on the mainframe over a 10 to14 day period. Tests done after production showed that long form production could have been done in 22 hours on a single 1.7 GHz Pentium IV PC while short form production could have been done in 45 hours on the same machine.

48. CANCEIS was built to handle the processing of millions of records and thousands of edit rules. Table 14 shows the number of edit rules per household size for the 2001 Census Demographic data processing of the long Census form.

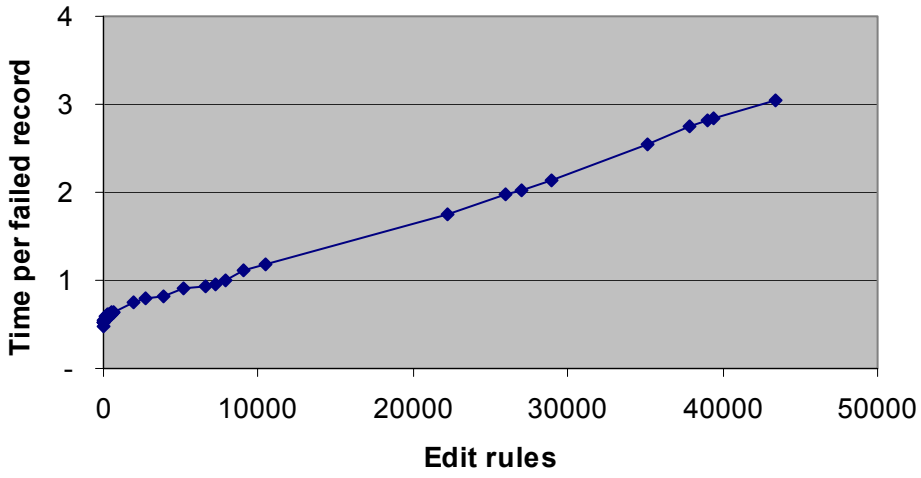
Table 14: 2001 Census demographic data, long form

Household size	Number of edit rules	Non-response/ invalid only	Non-response/ invalid and inconsistencies	Inconsistencies only	Total failure rate
1	4	5.9%	0.02%	0.8%	6.7%
2	93	8.2%	0.6%	1.7%	10.5%
3	985	13.6%	1.4%	3.2%	18.3%
4	3,424	15.5%	1.6%	3.0%	20.1%
5	8,154	16.8%	2.1%	3.6%	22.5%
6	15,919	20.0%	3.5%	4.9%	28.4%
7	27,465	32.5%	7.2%	6.8%	46.5%
8	43,541	32.9%	12.6%	7.8%	53.3%

49. For the demographic data, a household fails if there is non-response, some responses are invalid or if there are inconsistencies among the data in the household. A high failure rate for 7 and 8 person households is partially explained by the fact that there is a higher probability that at least one problem will occur since there is more data (i.e. 32.9% + 12.6% = 45.5% of 8 person households have some non-response or some invalid data in the demographic data). The higher failure rate is also explained by the fact the households with more than 6 persons receive two questionnaires (since only 6 persons can answer on a single questionnaire). This, in turn, can create more problems across the whole Census process.

50. Since CANCEIS is using the NIM methodology, the impact on processing time of adding edit rules is far less taxing than with the Fellegi/Holt approach used by many E&I systems. Chart 1 illustrates the results of a test, with the 8 person long form household stratum, to determine the impact of the number of edit rules on processing time. This chart clearly shows that with CANCEIS, the time to process a failed record (using a Pentium II 350 megahertz computer) increases in a linear fashion with the number of edit rules. Systems using Fellegi/Holt approach tend to have an exponential relationship between the number of edit rules and processing time (see Winkler, 1999).

Chart 2: Time in seconds per failed edit household by the number of edit rules, 8 person households



V. STEPS TAKEN FOR A TROUBLE FREE IMPLEMENTATION OF CANCEIS

51. For the Canadian Census of Population, it is not easy to change the methodology for a major operation like E&I. Many people involved in the decision process have worked through many Censuses and have become accustomed to the tools and processes involved in getting the job done on time. Therefore, any attempt to do a major change to this stable and proven process is met by some understandable resistance. This resistance can come from the subject matter experts that have to certify the imputed data, the programmers who will have to build this new system, the production officers who will need to learn how to use this new tool and by methodologists who are not yet convinced of the benefits of the new methodology.

52. To get this new methodology and new system implemented, many steps were taken to ensure a smooth transition. First, a prototype system was built to show that the overall methodology had potential, that the system could be build without too much effort, that the system could run in a satisfactory amount of time and that the quality of the data was at least as good as with the previous methodology. Once the prototype was successful and with the support from all parties (subject matter experts, computer scientists, production officers and other methodologists), it became easier to convince the upper management of the benefits of implementing a new E&I system. The implementation of the new system was again a group effort to determine the scope of the system, its short, medium and long term functionality and the implementation approach. The process to go from the idea of a better methodology to a fully implemented system used in production took 9 years, over the period of two Censuses.

53. The conversion of E&I processes from the old system to the new system was done in a few steps. The first step was the communication stage, where the features of the new system and the differences from the old system were conveyed and at the same time it was determined what E&I methodology was used with the old system. The goal of the next step was to show that changing the system without changing the overall approach resulted in imputed data of at least as good quality as with the previous system. This was done by using previous census/survey data and comparing the results using both systems. Once convinced that the new system was viable, the move to take advantage of the new features of the new system (more precise distance functions, more variables, less stratification, more donors, more optimising of parameters) was quite natural.

VI. CONCLUSIONS

54. 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.

55. 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. (2000a), "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. (2000b), "2001 Canadian Census Minimum Change Donor Imputation Methodology - Extended Version of Report", Social Survey Methods Division Report, Statistics Canada, Dated August, 2000.

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.

Winkler, William E., (1999), “State of Statistical Data Editing and Current Research Problems”, Working Paper No. 29, Presented at UN/ECE Work Session on Statistical Data Editing, Rome, Italy, June 2-4, 1999. (http://www.unece.org/stats/documents/1999.06.sde.htm)
