STUDY SERIES
(Computing #2007-1)

CANCEIS Experiments of Edit and Imputation with 2006 Census Test Data

Bor-Chung Chen

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

Report Issued: December 10, 2007

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

CANCEIS Experiments of Edit and Imputation with 2006 Census Test Data [*]

Bor-Chung Chen
Statistical Research Division
U.S. Bureau of the Census

Abstract

In this report, we demonstrate the CANCEIS (CANadian Census Edit and Imputation System) experiments of edit and imputation with the 2006 test data. The major effort is to translate the if-then-else rules of current edit and imputation system of the decennial census into the decision logic tables (DLT) of CANCEIS. We also formulate the input files that are needed to run the CANCEIS software. The advantages of using DLT are that it is easy to understand the edit rules; and DLTs are input, not part of the software, making it easier to change when edit rules are changed. We also compare the imputation results between the CANCEIS experiments and the 2006 Census Edited File. The comparison is for our curiosity beause the constructed DLTs are not identical to the edit rules specified in the 2006 edit specs. Although the edit rules are not identical, the comparison still shows some similarities between the CEF and CANCEIS results.

1 CANCEIS: An Introduction

The CANadian Census Edit and Imputation System (CANCEIS) was designed based on the Nearest-Neighbor Imputation Methodology (NIM) developed by Mike Bankier of Statistics Canada in 1992. CANCEIS works with three sources of information provided by the user: (1) unedited input data files (2006 Census Test 100% Census Unedtied Files (HCUF) in this experiment), (2) data dictionary files, and (3) edit rules (edit specifications) defined in decision logic tables (DLTs).

There are three major components of the CANCEIS software (see Figure 1): (1) DLT Analyzer: The DLT Analyzer uses the decision logic tables and the data dictionary information to check the edit rules specified by the user for any syntax error or inconsistency and then creates one unified DLT that is to be used by the Derive Engine or the Imputation Engine. (2) Derive Engine: The unified derive DLT is generated by the DLT Analyzer and processed by the Derive Engine. It allows the user not only to specify the edits but also to specify deterministic imputation actions that should be performed to fix a failed record without reference to any donor. (3) Imputation Engine: The Imputation Engine performs the hot deck imputation. It applies the rules of the unified hot deck DLT to the actual data and determines which units pass and fail the edit rules. Using the NIM methodology, it searches for passed units that resemble each failed unit (these passed units are called nearest-neighbor donors) and uses data from a nearest-neighbor donor to perform minimum change imputation. This donor search and selection is based on distance measures applied to each variable.

The DLT Editor is a VBA (Visual Basic for Applications) program in Microsoft EXCEL which accesses the flat input files for a CANCEIS module to facilitate the creation of propositions for DLTs.

There are several advantages of using CANCEIS DLTs for edit and imputation of census or survey data (listed with all possible applications of DLTs):

1. it is easy to understand DLTs after learning their structure, an example is shown in Appendix A;
2. DLTs are input, not part of the software;
3. when edit rules are changed, we just need to change DLTs, not the software;
4. it is easy to use for research on the impact of imputed results when edit rules are modified;

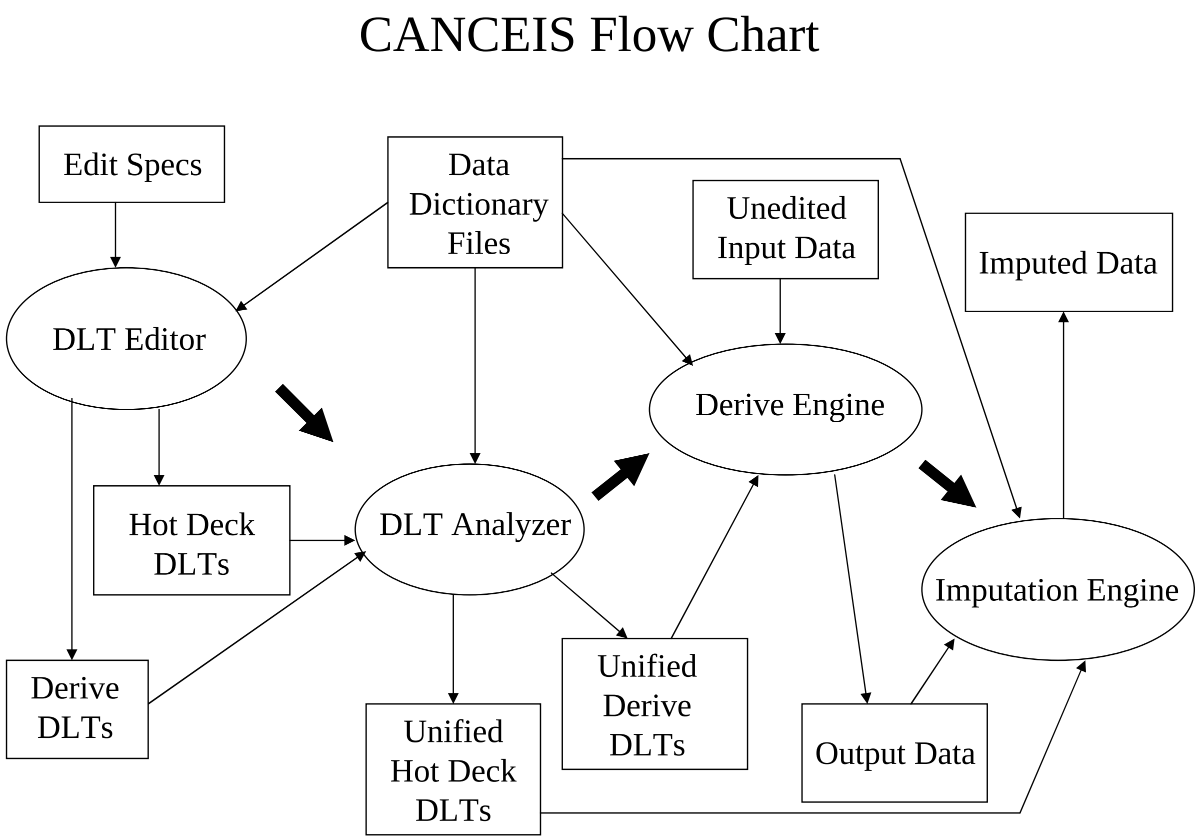
[*] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

5. the set of DLTs is a special purpose programming language for edit and imputation that we really need;
6. with the current edit specs we have, it is easy to plant bugs into the program when edit rules need to be changed, and CANCEIS does not have this problem.

If CANCEIS applications are attractive to the Bureau, we would recommend to form a DLT review team to start formatting the edit rules into a database of Decision Logic Tables. We strongly recommend that a complete set of thoroughly reviwed DLTs be available for the 2008 census dress rehearsal and the 2010 Decennial Census.

CANCEIS also has potential to be used in other surveys, such as American Community Survey (ACS). Therefore, we would recommend that the translation of the edit specifications to DLTs should be done automatically for decennial censuses, ACS, and other surveys.

Figure 1: CANCEIS Data and Processing Flow Chart.



2 CANCEIS Experiment with the 2006 Census Test Data

2.1 Introduction

As part of research effort for Census 2010, CANCEIS (Canadian Census Edit and Imputation System)[2] was applied for the household relationship, sex, and age questions. Results from CANCEIS and from two other methods used in the research - AR (Administrative Records) direct assignment for sex and age and traditional hot deck - were compared using a “truth deck” created from non-imputed values from Census 2000 records by PRED (Planning, Research, and Evaluation Division). The preliminary results from the three methods indicated that for the four states (DE, GA, NY, FL) chosen for this analysis CANCEIS performed slightly better on imputing spouse and nonrelative for the question on household relationship and imputing at the lower and upper end of the age spectrum [5].

POP (Population Division) and HHES (Housing and Household Economic Statistics Division) wanted to extend the application of CANCEIS to race, Hispanic origin, and housing tenure variables to see how well CANCEIS would perform. The Imputation Workgroup also wanted to include type of vacancy as part of the research effort. However, due to the constraints of the available resources and timing, we did not include type of vacancy in this experiment.

2.2 Scope of the Experiment

The person level variables of interest in this experiment are relationship (QREL), sex (QSEX), age (QAGE), race (QRACEX), and Hispanic Origin (QSPANX). We also include the household level variable, tenure, in the experiment. The main effort of this research is the decision logic table (DLT) construction, that is based on the edit specifications (called the edit specs, hereafter) of the *2006 Post Processing Legacy Edit Design – 2006SD* [6]. We ran CANCEIS using data from the 2006 Census Test for the central portion of Travis County, Texas and the Cheyenne River American Indian Reservation and Tribal Trust Lands in South Dakota. We used CANCEIS Version 4.5.4 of February 2007 for this experiment.

2.3 Pre-edits or Derive Imputations

Here, the pre-edits (also called derive imputations that are performed by the CANCEIS derive engine) are defined, for the purposes of this analysis, as imputations that can be derived from other variables of the same person or from a different person in the same household. For example, a missing age can be derived from the reported date of birth (other variable from the same person). Another example is that a missing race of the householder can be derived from the corresponding response of the mother in the same household (same variable of different person from the same household). In contrast, the hot deck imputation is the imputation that any missing, invalid, or inconsistency items are imputed from the valid and consistent items of one of the other households, called donors.

2.4 The Setups for the Experiment of CANCEIS

We ran each test site separately and the stratification variable is the household size. A total of 12 strata is used; strata 1 to 12 are for household size 1 to 12, respectively, and stratum 13 is for household size 13 and above. In stratum 13, everyone in the 13+ person household is treated as an individual of a single person household. This special treatment of the households in stratum 13 is due to the limitation of the effective application of CANCEIS that we might not be able to obtain enough donor pool when the household size is larger than 12. We would look to provide a more adequate application in the future. We ran CANCEIS for the person level variables and for the household level variables separately.

2.4.1 Person Level Variable Imputation

The variables included are FINAL_MAFID (household ID, not imputable), PP_PNC3 (person ID in the household, not imputable), QREL (relationship), QSEX (sex), QAGE(age, 0–115), QRACEX (race), and QSPANX (Hispanic origin). The pre-edits are

1. §D.1 Relationship Pre-edit (page 21 of the edit specs [6]);
2. §D.2 Sex Pre-edit (page 22 of the edit specs [6]);
3. §D.3 Age and Date of Birth Pre-edit (page 22 of the edit specs [6]);
4. §D.4 Race Pre-edit (page 30 of the edit specs [6]);
5. §D.5 Hispanic Pre-edit (page 34 of the edit specs [6]); and
6. §F.3 Create QSPANX and QRACEX Variables (page 79 of the edit specs [6]).

However, we don't have the following from the edit specs [6] due to lack of data required, such as the administrative records, ancestry data, and Hispanic surnames:

1. pre-edits from administrative records,
2. ancestry pre-edits, and
3. Hispanic pre-edits from the surnames.

2.4.2 Household Level Variable Imputation

The household level and person level variables included are FINAL_MAFID (not imputable), PP_PNC3 (not imputable), QAGE (the householder's age, imported from the results of CANCEIS runs of the person level variables, not imputable), QRACEX (the householder's race, imported from the results of CANCEIS runs of the person level variables, not imputable), TEN (tenure), SEQSTATUS (type of unit record, not imputable), SEQVACANT (vacancy status, not imputable), and QRELs (all household members' relationship, imported from the results of CANCEIS runs of the person level variables, not imputable). Each person in the household is a record. Each record has its unique value for PP_PNC3 and QREL. All records of a housing unit have the same values for all other variables.

Since there is a consistency check between tenure (vacant, i.e., TEN = 0) and vacancy status (occupied, i.e., SEQVACANT = 0), both variables should be included in the same CANCEIS runs. The variables of the householder's QAGE and QRACEX, and all QRELs are included because they are involved in the allocation matrices (pages 117–121 of the edit specs [6]) of tenure edits, and household size can also be identified with the number of QRELs. No pre-edits are needed in the CANCEIS runs of housing unit level variables. The CANCEIS runs with the household level variables are also stratified by the household size.

The `tenure` variable and other household level variables if included could have been used in the same CANCEIS runs of the person level variable imputaions. However, we decided to have separate runs from the person level variables (other than those described above). The reason for this was to avoid the unnecessary comparisons when searching for a potential donor that would have resulted from combining tenure and all of the person level variables into a single run. In addition, a separate run is more computationally efficient and increases the donor pool because there are fewer households that fail the tenure edits than those that fail at least one of the person level variables.

3 Person Level Variable Derive and Hot Deck Imputation Input Files

3.1 CANCEIS Input Files

In this experiment for the derive and hot deck imputations, we have fourteen input files each to run the DLT Analyzer, the Derive Engine, and the Imputation Engine. Some of the files are conditionally optional (i.e. if a user wishes to use coded variables, then the user must create a coded variable file; a coded variable is a variable that a response is given a numerical code, such as 2 is assigned to spouse of QREL). Each of the fourteen files are required per stratum. Thirteen of the files listed in Table 1 are for the 4-person household stratum, where the `04` in the file names is the household size 4. The other one file is the input data file (the UNIT file, not shown in Table 1). In Table 1, a value set is the set of all possible values, valid or invalid, of a variable and the validity set only consists if the valid values of a variable. Therefore, the validity set is a subset of the value set. A more detailed description of each of the fourteen files can be found in the the CANCEIS User's Guide [1] and a detailed version [3] of this paper. All the input files are given by the users to the derive and hot deck modules described in Section 3.2.

Table 1: CANCEIS Input Files

	File Name	Description
Input Data Layout	dc04var.txt	This file describes the different variables that are part of the derive or hot deck module, whether each variable is repeated or not, the validity set associated with each variable.
Value Sets and Validity Sets	dc04set.txt	The first two columns are the validity set ID and the variable type. The third column represents the value set ID to which the validity set is associated. The last column represents the number of digits of precision after the decimal point for continuous variables.
Labels for Coded Variables	dc04vscode.txt	These two files contain the list of valid responses associated with each coded variable value set.
	dc04code.txt	
Intervals for Numeric Variables	dc04interv.txt	We use the interval ID, the minimum, maximum, and step values for discrete variables (D) only.
	dc04num.txt	The validity set ID and inverval ID are placed into the file.

Table 1: CANCEIS Input Files (Continued)

	File Name	Description
Classes of Coded and Numeric Variables	dc04class.txt	This file contains all the class names as well as the validity set IDs. A class name is only written once. It defines class names and indicates what value set they are associated with.
	dc04clcode.txt	This file contains the class ID and the label (response name) for coded variables (qualitative validity sets). It lists the responses associated with each coded class name. These responses must come from the value set linked to the class in the `dc04class.txt` file.
	dc04clnum.txt	This file contains all the interval IDs associated with a class. It lists the intervals associated with each numeric class. Classes can be comprised of multiple intervals. The intervals appearing in this file must be declared in the `dc04interv.txt` file.
Imputation	dc04impparam04.txt	This file is used to identify several imputation characteristics of the variables defined in the input data file (the UNIT file).
Parameters	dc04imp04.txt	This file lists the cases where specific sub-units (a subunit is a person within a household in our experiment) have a different imputability value than the default defined for the given variable in the IMPPARAM file.
Permutability of Sub-Units	dc04permu04.txt	Since it is assumed that all sub-units (persons) within a record (household) are **not** permutable, this file is only used to list the exceptions, that is, those sub-units which are permutable. A permutable sub-unit is one that may be shuffled among all permutable sub-units to find the best donor match. A fixed sub-unit cannot be shuffled, and will always stay in the same position.
System Parameters	dc04SYSP04.txt	This file is a list of all user-defined system parameters. System parameters are values that are used by CANCEIS to perform editing, donor searches, and imputation. See [1] for detailed descriptions of the system parameters and their possible values.

3.2 Decision Logic Tables (DLTs)

CANCEIS 4.5.4 has Hot-Deck modules and Derive modules:

1. **Hot-Deck** modules are used with the Imputation Engine to perform minimum change donor imputation, where only Hot-Deck DLTs can be used;

2. **Derive** modules are used with the Derive Engine to perform deterministic imputation, where only Derive DLTs can be used.

Each DLT consists of header information and prospositions. Two examples, one for a Derive DLT and the other for a Hot Deck DLT, are shown in the appendix. The header information is a set of parameters that the user specifies in order to customize the CANCEIS application. Header lines are denoted by a `%` at the beginning of the line. A detailed description of the header information can be found in the the CANCEIS User's Guide [1]. There are three different types of propositions that can be used for Derive modules. Each proposition has to be preceded by a special symbol and given in a specific order. They are:

1. **Common Actions:** Common actions are actions to be performed on all records unconditionally. They must be specified below all header information, but above any condition propositions. They are denoted using a `$` symbol at the beginning of the line.

2. **Conditions:** Conditions must be specified below all header information and common actions, but above any conditional actions. They are denoted using a `@` symbol at the begining of the line.

3. **Conditional Actions:** Conditional actions are denoted using a `&` symbol at the beginning of the line. Conditional actions must be specified below all condition propositions so that if some condition propositions hold, certain conditional actions will be executed.

In Hot-Deck DLTs, only conditions can be used. They must be specified below all header information and denoted using a `@` symbol at the begining of the line. A Derive DLT may consist of common actions only, conditions and conditional actions, or all three types.

3.2.1 Census Test 2006 Person Level Variable Derive and Hot Deck DLTs

In translating the edit rules from the edit specs to DLTs, no attempts are made to optimize the DLT constructions. There are also no reviews from the subject matter experts regarding the edit rules specified in the DLTs. The set of edit rules specified in the DLTs of this experiment is not identical to that specified in the 2006 Post Processing Legacy Edit Design, but we tried to minimize the differences. We have two sets of DLTs: Derive DLTs (for Derive Engine) and Hot Deck DLTs (for Imputation Engine). Both sets of DLTs need to go thru DLT Analyzer to check syntax error and inconsistency.

Table 2 shows the numbers of DLTs constructed for the experiment, in which the names of the DLTs indicate what variables are involved in the edit rules specified in the DLTs. For example, the `Relationship--Age` DLT means that the variables of `Relationship` and `Age` are involved in the edit rules.

Table 2: Numbers of Person Level Derive and Hot Deck DLTs

DLT(s)	Derive	DLT(s)	Hot Deck
Relationship	2	Relationship	2
Sex	1	Spouse--Age	2
Age	21	Same Sex Marriage	1
Relationship--Age	1	Relationship--Age	4
Race	4	More than one spouse	1
Hispanic Origin	1		--
Race--Hispanic Origin	48		--
Total	78	Total	10

4 Household Level Tenure Hot Deck Imputation Input Files

The input file names are same as those given in Section 3 with different set of imputation variables, which were described in Section 2.4.2. There is only one DLT for the household level hot deck imputation and it is shown in Appendix B.

5 Data Processing Flow of the Experiment

The processing flow for the experiment is given in Figure 2. There are three imputation processing modules: person level derive imputation, person level hot deck imputation, and household level hot deck imputation. The input data to the *person level derive imputation* (the first imputation module) are unedited QREL, QSEX, QAGE, QRACEX, QSPANX, allocation flags, and intermediate variables, such as PYOB (Year of Birth), PMOB (Month of Birth), PDOB (Day of Birth), etc. The output data are the derive imputed QREL, QSEX, QAGE, QRACEX, and QSPANX based on the person level derive DLTs.

The input data to the *person level hot deck imputation* (the second imputation module) are the output data from executing the imputation module of *person level derive imputation* and the ouput data are the final imputed QREL, QSEX, QAGE, QRACEX, and QSPANX based on the person level hot deck DLTs.

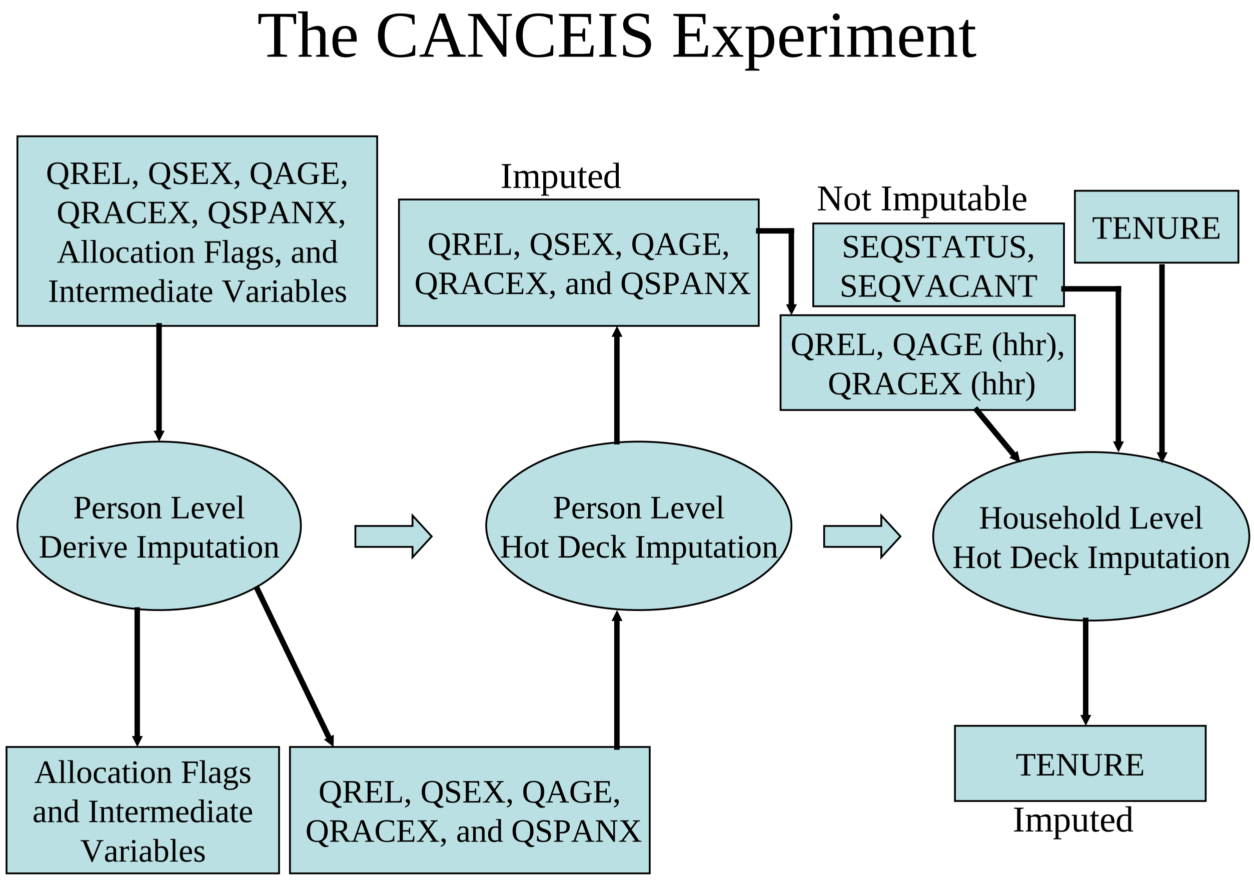
Finally, the input data to the *household level hot deck imputation* (the third imputation module) are TENURE, SEQSTATUS, SEQVACANT, and the imported QREL, the householder's (hhr) QAGE, and the householder's QRACEX from the output of the *person level hot deck imputation.* The data of QREL, QAGE, and QRACEX in the third imputation module are not imputable. The output data are the imputed TENURE. The values of SEQSTATUS and SEQVACANT were not imputed for the purposes of this experiment.

6 The Results from the CANCEIS Experiments

6.1 The Person Level Variable Imputation

In this section, we present the imputation results of the CANCEIS experiments and the values from the 2006 CEF (Census Edited File) using the traditional edit and imputation methodology. We only show the households with sizes 1 to 12. The households with size 13 and above are processed as group quarters, in

Figure 2: CANCEIS Experiment.


The CANCEIS Experiment
QREL, QSEX, QAGE, QRACEX, QSPANX, Allocation Flags, and Intermediate Variables
Imputed
QREL, QSEX, QAGE, QRACEX, and QSPANX
Not Imputable
SEQSTATUS, SEQVACANT
TENURE
QREL, QAGE (hhr), QRACEX (hhr)
Person Level Derive Imputation
Person Level Hot Deck Imputation
Household Level Hot Deck Imputation
Allocation Flags and Intermediate Variables
QREL, QSEX, QAGE, QRACEX, and QSPANX
TENURE
Imputed

which there is no between-persons edit. In other words, they are treated as one-person households for each individual in the households.

6.1.1 Pass Rates

One of the criteria to evaluate a hot deck imputation methodology is to see if there are enough donors for the failed records. There are two types of edit rules: `conflict` rules or `validity` rules. A household record satisfying any of the `conflict` rules is not a "good or clean" record. If a household record satisfies all of the `validity` rules, it is a "good or clean" record. Currently, CANCEIS only works with conflict rules, so a passed household record must not satisfy any of the edit rules specified in the hot deck DLTs and a failed record satisfies at least one of the edit rules. When the household size is large, it may not get enough donors, Table 3 shows that the pass rates are very low for household size 7 and above with Texas data and for household sizes 10 and 12 with South Dakota data. One way to get more donors is to combine the adjacent geographical areas into an imputation group.

6.1.2 Statistical Comparisons

One of the important criteria raised by Fellegi and Holt[4] was to maintain the frequency distribution of variables before and after imputation when imputation is necessary. In this section, we compare the frequency distribution of the imputed values[1] for the CANCEIS results and the 2006 Census Test CEF results with the reported values[2] from the 2006 Census Test, despite the fact, mentioned above, that the DLTs and the 2006 edit specs are not identical.

[1] Here, an imputed value is defined as a value of a person level variable of a failed household after imputation. A failed household is a household that fails at least one of the edit rules. Therefore, an imputed value may be *different from* or *same as* the reported value depending on the imputation procedure of an imputation system.

[2] A reported value is defined as a value reported by the respondents. For example, a reported value of age with age allocation

Table 3: Percentage of Passed Households with Person Level Variables Used as Donors for the CANCEIS Imputation Engine with 2006 Census Test Data

	Cheyenne River Reservation, South Dakota				Portion of Travis County, Texas			
size	Total Households	Passed Households	Failed Households	Pass Rate(%)	Total Households	Passed Households	Failed Households	Pass Rate(%)
1	556	497	59	89.4	62929	51604	11325	82.0
2	655	569	86	86.9	55334	46594	8740	84.2
3	401	330	71	82.3	27955	23228	4727	83.1
4	346	291	55	84.1	20554	16931	3623	82.4
5	253	213	40	84.2	10238	8292	1946	81.0
6	146	112	34	76.7	4757	3709	1048	78.0
7	77	70	7	90.9	2024	1144	880	56.5
8	47	42	5	89.4	899	502	397	55.8
9	20	14	6	70.0	389	211	178	54.2
10	7	4	3	57.1	192	106	86	55.2
11	2	2	0	100.0	86	42	44	48.8
12	5	3	2	60.0	38	16	22	42.1
total	2515	2147	368	85.4	185395	152379	33016	82.2

We intend to look at the imputation system that has a "closer" frequency distribution to that of the reported values of the person level variables, such as age. We define the "closeness" measurement between the sets of the imputed values of the person level variables and the reported values of the corresponding variables as the sum of the absolute deviations between their frequency distributions:

$$C = \sum_{i=1}^{n} g_i = \sum_{i=1}^{n} \frac{|x_i - r_i|}{r_i},$$

where n is the number of categories or the number of all possible valid values of a variable; x_i is the number of individuals of category i, and r_i is the number of individuals of category i with reported data, i.e., allocation flag of 0. A variable value with allocation flag of 0 is usually *a reported value* (except the age value, please see footnote 2). A small value of C would represent a "look alike" frequency distribution of the reported values of the person level variables. For example, Table 4 shows the frquency distributions of the imputed values of age for household sizes of 1 to 12 for the portion of Travis County, Texas in the 2006 Census Test with the original age allocation flags of 0 and 1 (the columns labeled with (2) and (4)). These allocation flags are assigned by the current imputation system (not CANCEIS) and defined in the edit specs [6]. An age value with age allocation flag of 0 is *consistent as reported* and an age value with allocation flag of 1 is *age only.* In Tabel 4, the other possible allocation age flags are *date of birth only* (2), *inconsistent age and date of birth* (3), *allocated from hot deck* (4), *substituted* (7), and *age of householder or spouse adjusted to be consistent with age of children* (8). The valid age is between 0 and 115 that is divided into 23 categories with 5 years in each category except the last one with 6 years. An extra category for age 0 is also shown in Table 4 because some of age 0 are reported as "months-old" and could be mistakenly recorded as "years-old". From the table, the values of C (all the values of C do not include the g_i values of category age 0) are 2.157 and 2.327, respectively, for the CEF and CANCEIS results indicating that the imputation results from the systems are "look alike" to each other. Table 4 also shows the frequency distributions of imputed values of age for all allocation flags (the columns labeled with (3) and (5)). The CANCEIS imputation results have smaller value of C ($= 4.449$) indicating that their distribution of overall imputed values is more "look alike" the distribution of the reported values than the CEF results ($C = 5.754$). For the category of age 0, the CANCEIS imputed results are "closer" to the reported values than the CEF imputed results (the g_i values of 0.042 vs. 0.066 for flags of 0 and 1, and 0.095 vs. 0.127 for all flags). The values of C and g_i are only shown for the age variables because we are more interested in the distributions of age values. Figure 3 shows the bar charts of the frequency distributions of age with allocation flag 0 or 1 from CEF and CANCEIS imputation results of

flag of 0 as specified in [6] is consistent with the reported value of date of birth. A reported value of age with age allocation flag of 1 is the value reported with missing value of date of birth also as specified in [6]. Therfore, a reported value of age with flag of 0 is more reliable than that of age with flag of 1.

household sizes 1 to 12 from the portion of Travis County, Texas in the 2006 Census Test. The bar charts of the frequency distributions of age with all flags are shown in Figure 4.

Table 5 and Figure 5 show the frequency distributions of the relationship of household sizes 1 to 12 from the portion of Travis County, Texas in the 2006 Census Test. The frequencies of the relationship are very similar between the CEF and CANCEIS results except `unmarried partner`, `parent-in-law`, `parent`, and `son/daughter-in-law` categories. The distinctions for these four categories are based on the largest percentage differences. Here, the percentage difference is defined as

$$\frac{|m_i - n_i|}{n_i}\%,$$

where m_i is the CANCEIS result and n_i is the CEF result. The largest percentage differences are the categories of `unmarried partner` (15.45%) and `parent-in-law` (14.49%). The next level of the percentage differences are the categories of `parent` (9.77%) and `son/daughter-in-law` (8.34%). For the values of `unmarried partner`, the difference came from the fact that the edits of the current system assign the `unmarried partner` to the relationship when certain conditions hold (see Section E.4.b.1 of the 2006 edit specs, page 48). The hot deck imputation of CANCES is more likely than the current system to assign `parent-in-law`, `parent`, and `son/daughter-in-law` to the relationship when an imputed value of relationship is needed.

The frequency distributions of sex, race, and Hispanic origin for the portion of Travis County, Texas in the 2006 Census Test are shown in Tables 6, 7, and 8 and their corresponding bar charts are shown in Figures 6, 7, and 8, respectively. They don't show any significant differencies except that CEF assigns more `American Indians or Alaska Natives (AIAN)` and `Native Hawaiian and Other Pacific Islander (NHOPI)` to the race.

6.2 The Household Level Tenure Imputation

Table 9 shows the CANCEIS pass rates of the tenure hot deck imputation. The pass rates in the table for both the Cheyenne River Reservation in South Dakota and the portion of Travis County, Texas in the 2006 Census Test are very high, therefore, the frequency distributions of the tenure variable of the CEF and CANCEIS results, shown in Table 10 and Figure 9, are very similar. It indicates that CANCEIS results are consistent with the CEF results and CANCEIS has the advantages over the current system as described in Section 1.

7 Conclusion and Discussion

In conclusion, we have demonstrated the use of CANCEIS DLTs and the CANCEIS methodology for edit and imputation with the 2006 census test data. It shows several advantages of using CANCEIS software over the current If-Then-Else system as described in Section 1. Although the sets of edit rules between the current system and CANCEIS DLTs are not identical, the imputed results between the two systems are exceptionally consistent. The major differencies are in the `unmarried partner` category of relationship and the `American Indian or Alaska Native` and `Native Hawaiian and Other Pacific Islander` categories of race as discussed in Section 6.1.2. We don't have good explanation for some of the differences due to the fact that the sets of edit rules between the current system and CANCEIS DLTs were not identical and some of the data were not available to us, such as the administrative records, which were used as a source for imputing census records as provided in the current imputation system.

In these CANCEIS experiments, the imputation results are consistent with the CEF results. A detailed statistical comparison between the CANCEIS and the CEF results would have made sense if the CANCEIS DLTs had completely been constructed from the 2006 edit specs and had been reviewed by the demographic subject matter experts of the edit rules. We would recommend to form a DLT review team to start formatting the edit rules into a database of CANCEIS Decision Logic Tables. We also strongly recommend that a complete set of thoroughly reviwed DLTs be available for the 2008 census dress rehearsal and the 2010 Decennial Census for research purpose.

Table 4: The Frequency Distributions of Age for Portion of Travis County, Texas in 2006 Census Test

			CEF				CANCEIS			
Label	Age Group	(1)	(2)	g_i(2)	(3)	g_i(3)	(4)	g_i(4)	(5)	g_i(5)
-	Age 0	6689	7132	0.066	7538	0.127	6971	0.042	7323	0.095
0	000-004	30930	34113	0.103	35667	0.153	34017	0.100	35728	0.155
5	005-009	26007	29458	0.133	30758	0.183	29405	0.131	30935	0.189
10	010-014	22356	24958	0.116	26013	0.164	24939	0.116	26238	0.174
15	015-019	24897	28420	0.142	30088	0.208	28355	0.139	29994	0.205
20	020-024	45260	55037	0.216	57781	0.277	55070	0.217	58101	0.284
25	025-029	41841	50206	0.200	52657	0.259	50212	0.200	52681	0.259
30	030-034	35292	41752	0.183	43795	0.241	41827	0.185	43831	0.242
35	035-039	29237	34188	0.169	35881	0.227	34254	0.172	36047	0.233
40	040-044	25116	28757	0.145	30338	0.208	28669	0.141	30049	0.196
45	045-049	23249	25742	0.107	27189	0.169	25827	0.111	27110	0.166
50	050-054	20043	22009	0.098	23166	0.156	21986	0.097	23047	0.150
55	055-059	15815	16894	0.068	17799	0.125	16972	0.073	17772	0.124
60	060-064	9980	10572	0.059	11116	0.114	10576	0.060	11007	0.103
65	065-069	6898	7262	0.053	7647	0.109	7243	0.050	7570	0.097
70	070-074	5643	5876	0.041	6222	0.103	5885	0.043	6144	0.089
75	075-079	4600	4760	0.035	5042	0.096	4764	0.036	4964	0.079
80	080-084	3382	3490	0.032	3678	0.088	3493	0.033	3644	0.077
85	085-089	1691	1747	0.033	1857	0.098	1764	0.043	1855	0.097
90	090-094	622	644	0.035	687	0.105	633	0.018	666	0.071
95	095-099	145	150	0.034	169	0.166	152	0.048	177	0.221
100	100-104	26	30	0.154	34	0.308	23	0.115	25	0.038
105	105-109	5	5	0.000	6	0.200	4	0.200	6	0.200
110	110-115	1	1	0.000	3	2.000	1	0.000	2	1.000
	C			2.157		5.754		2.327		4.449

Notes: (1) Frequency Distribution (r_i) of Reported Values of Age with Flag of 0
(2) Frequency Distribution (x_i) of Imputed Values of Age with Original Flags of 0 and 1
(3) Frequency Distribution (x_i) of Imputed Values of Age with All Original Flags
(4) Frequency Distribution (x_i) of Imputed Values of Age with Original Flags of 0 and 1
(5) Frequency Distribution (x_i) of Imputed Values of Age with All Original Flags

Figure 3: The Frequency Distribution Bar Charts of Age with Flag = 0 or 1 for Portion of Travis County, Texas in 2006 Census Test

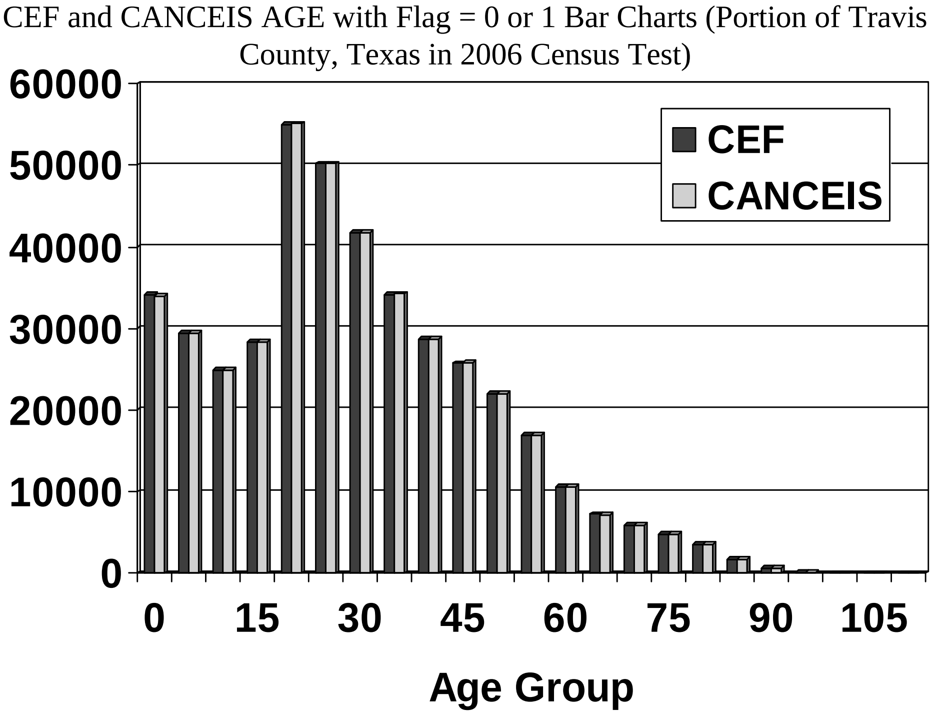


Figure 4: The Frequency Distribution Bar Charts of Age with All Flags for Portion of Travis County, Texas in 2006 Census Test

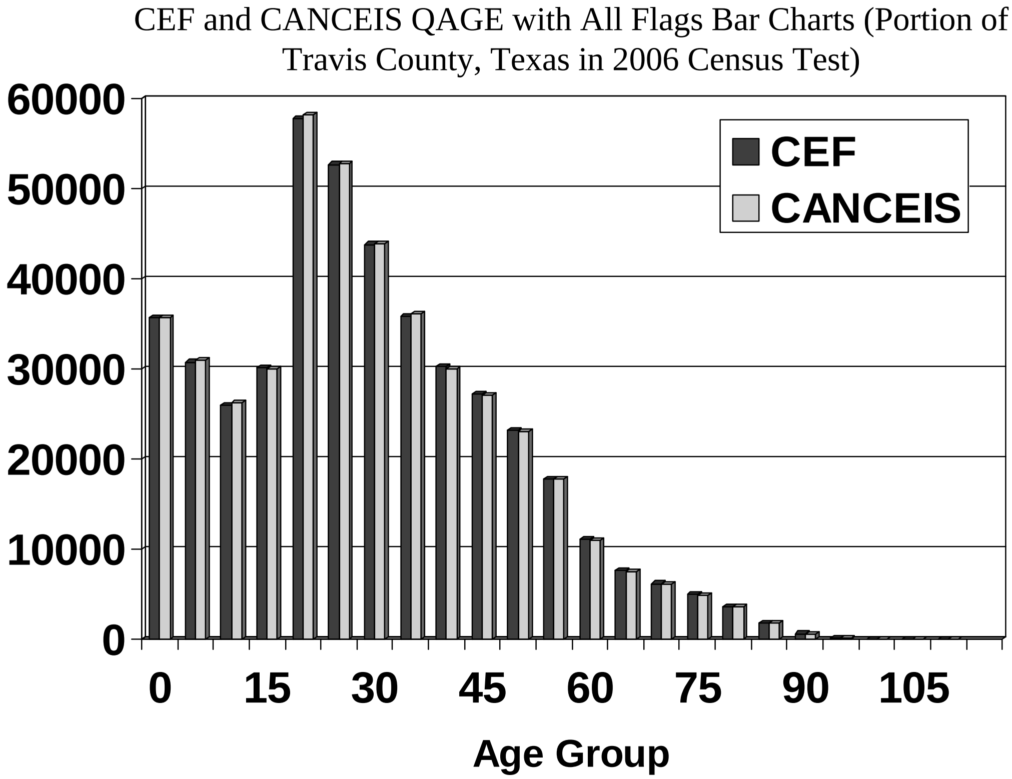


Table 5: The Frequency Distributions of Relationship for Portion of Travis County, Texas in 2006 Census Test

code	relationship	CEF	CANCEIS	code	relationship	CEF	CANCEIS
1	Householder	185395	185395	9	Parent-in-law	809	927
2	Husband or wife	62198	62302	10	Son/Daughter-in-law	1367	1481
3	Biological child	109901	109656	11	Other relative	12113	11783
4	Adopted child	1757	1771	12	Rommer or boarder	3793	3875
5	Stepchild	4020	4049	13	Housemate or roommate	26086	27330
6	Brother or sister	9409	10037	14	Unmarried partner	12075	10209
7	Parent	4146	4551	15	Foster child or adult	360	362
8	Grandchild	8627	8506	16	Other nonrelative	5537	5359
total						447593	447593

Table 6: The Frequency Distribution of Sex for Portion of Travis County, Texas in 2006 Census Test

Sex	CEF	CANCEIS
male	229574	229592
female	218019	218001
total	447593	447593

Figure 5: The Frequency Distribution Bar Charts of Relationship for Portion of Travis County, Texas in 2006 Census Test

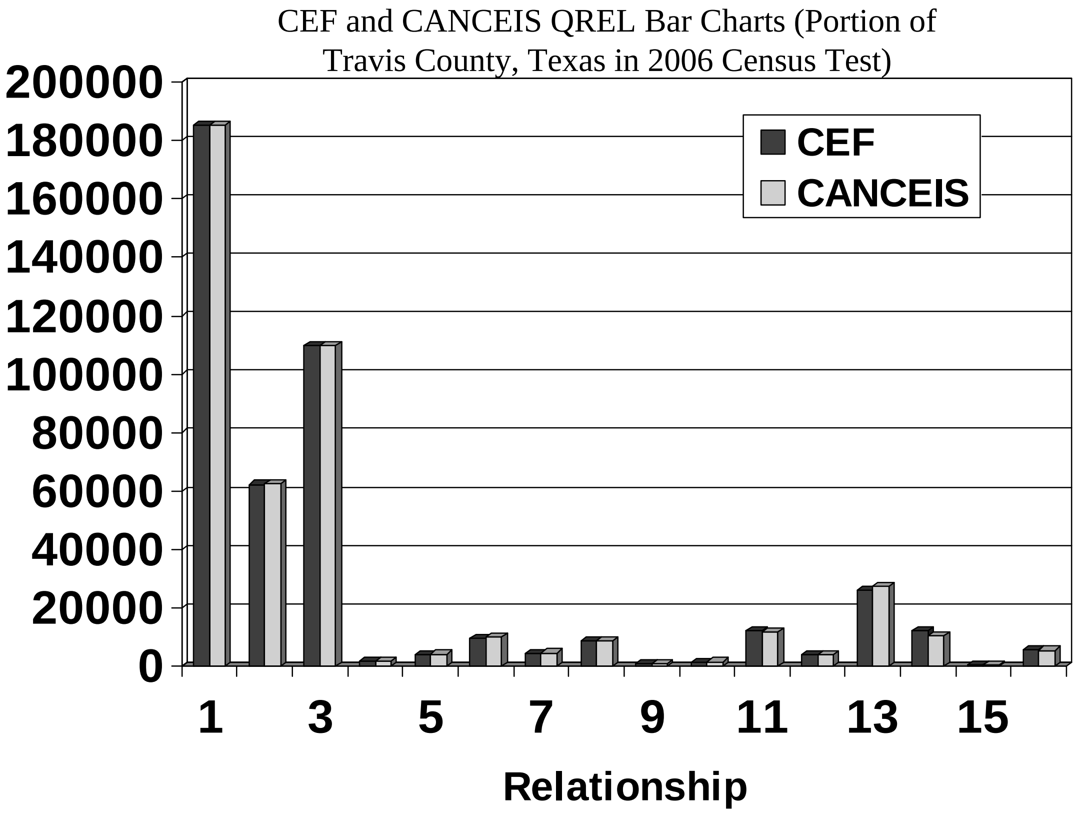


Figure 6: The Frequency Distribution Bar Charts of Sex for Portion of Travis County, Texas in 2006 Census Test

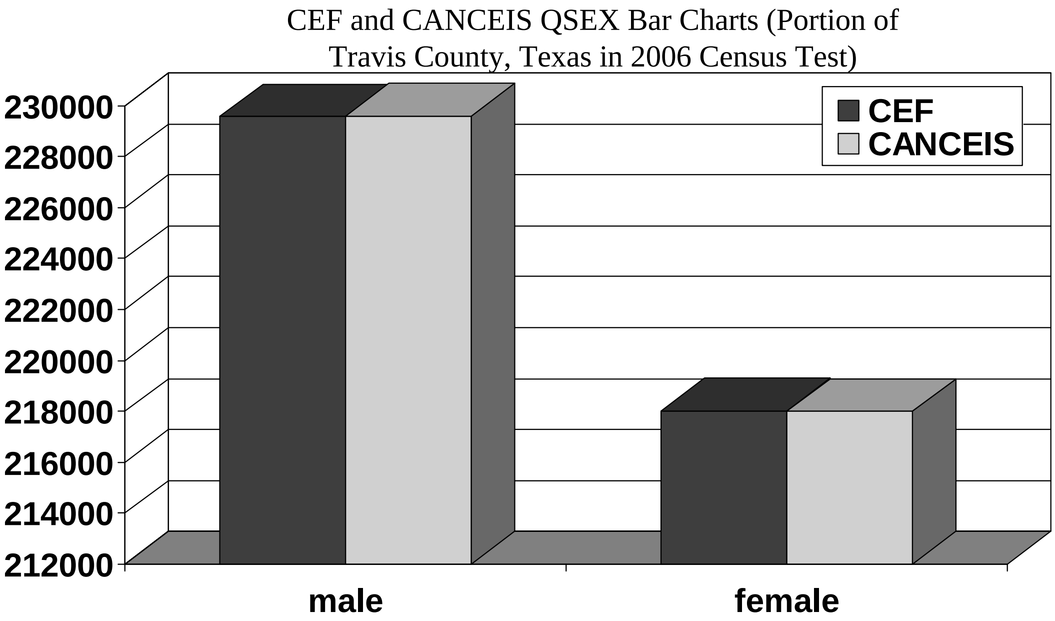


Table 7: The Frequency Distribution of Race for Portion of Travis County, Texas in 2006 Census Test

Race	CEF	CANCEIS
white	268707	271313
black	57796	57782
AIAN	4380	2938
Asian	22949	22484
NHOPI	943	747
SOR	92818	92329
total	447593	447593

Figure 7: The Frequency Distribution Bar Charts of Race for Portion of Travis County, Texas in 2006 Census Test

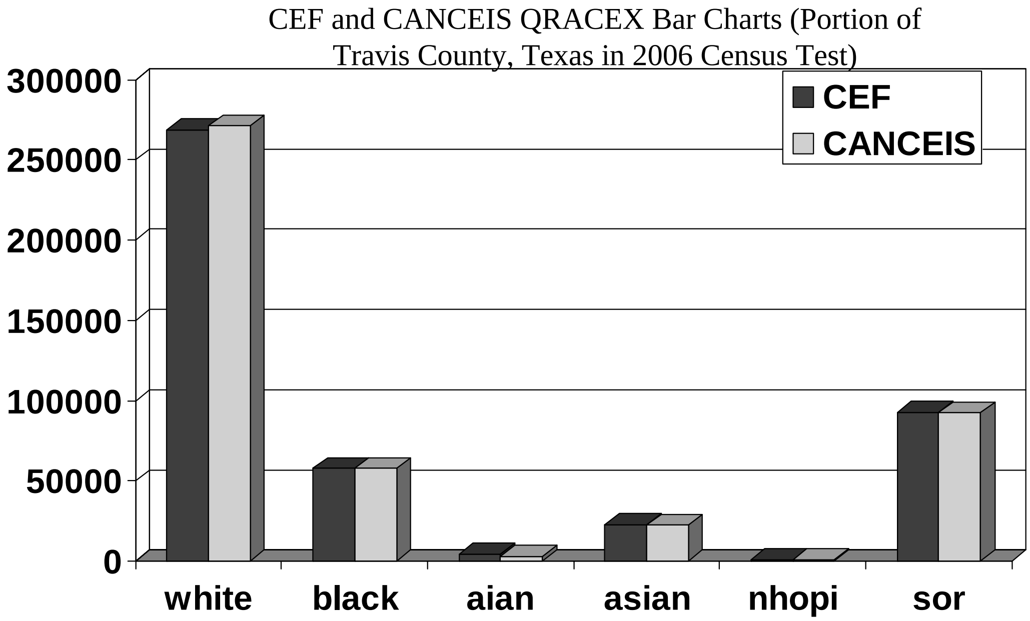


Table 8: The Frequency Distribution of Hispanic Origin for Portion of Travis County, Texas in 2006 Census Test

Hispanic Origin	CEF	CANCEIS
Hispanic	197211	197747
Not Hispanic	250382	249846
total	447593	447593

Figure 8: The Frequency Distribution Bar Charts of Hispanic Origin for Portion of Travis County, Texas in 2006 Census Test

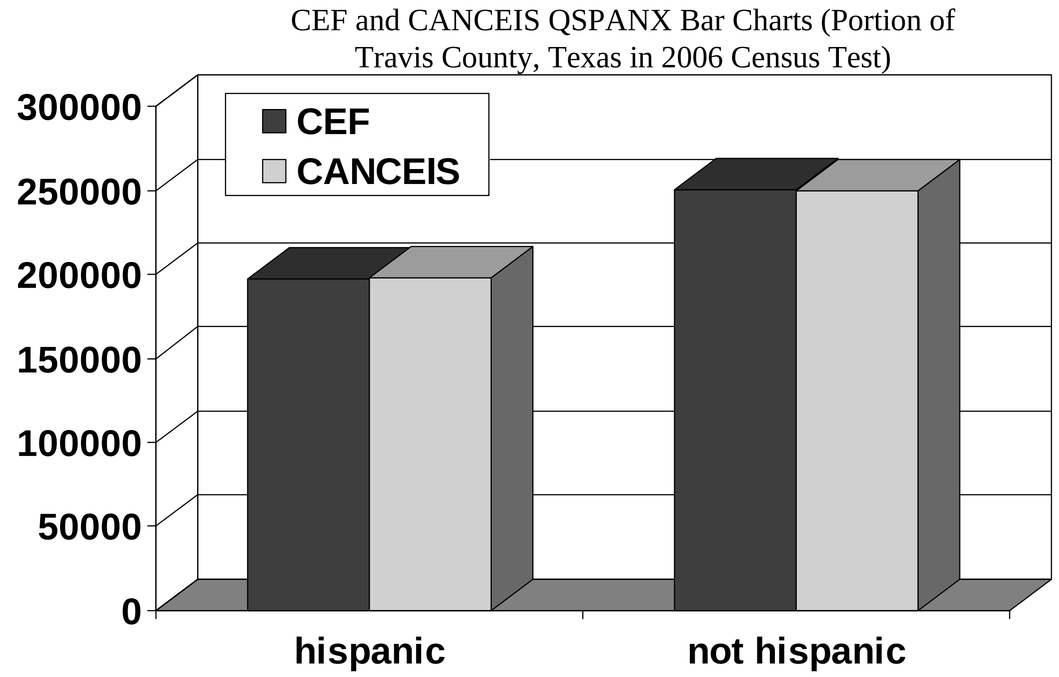


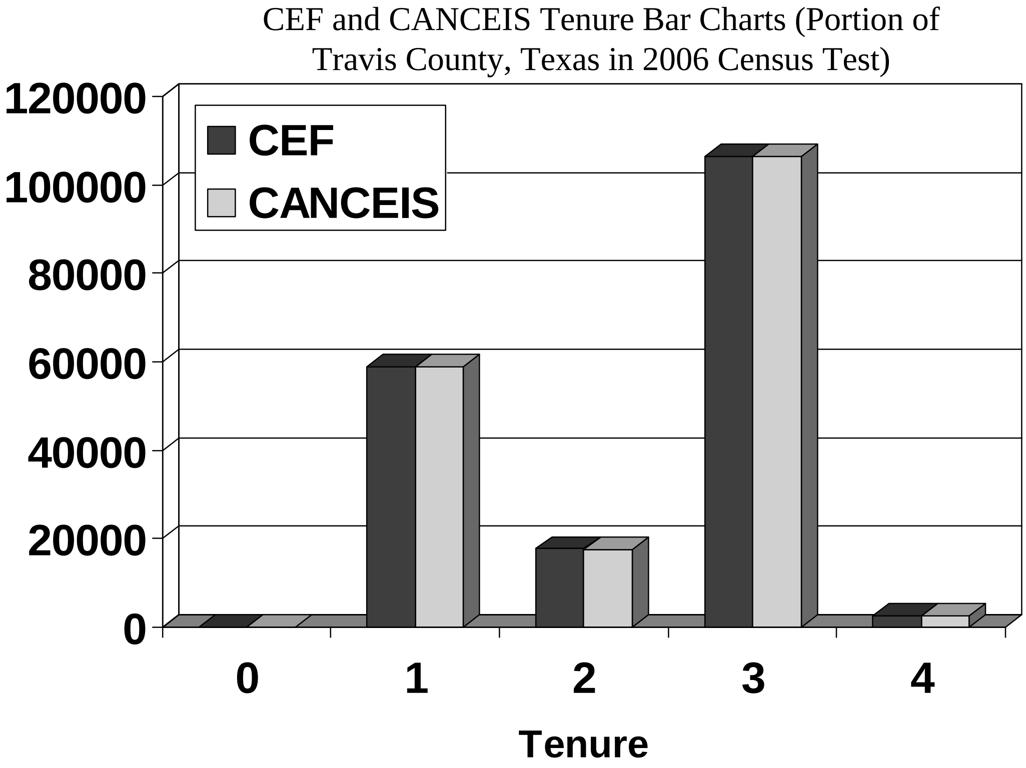
Table 9: CANCEIS Household Level Tenure Imputation Report of 2006 Census Test

	Cheyenne River Reservation, South Dakota				Portion of Travis County, Texas			
size	Total Households	Passed Households	Failed Households	Pass Rate(%)	Total Households	Passed Households	Failed Households	Pass Rate(%)
1	556	534	22	96.0	62929	61337	1592	97.5
2	655	635	20	96.9	55334	54224	1110	98.0
3	401	392	9	97.8	27955	27382	573	98.0
4	346	343	3	99.1	20554	20117	437	97.9
5	253	244	9	96.4	10238	9952	286	97.2
6	146	142	4	97.3	4757	4644	113	97.6
7	77	75	2	97.4	2024	1970	54	97.3
8	47	46	1	97.9	899	876	23	97.4
9	20	20	0	100.0	389	376	13	96.7
10	7	7	0	100.0	192	189	3	98.4
11	2	2	0	100.0	86	84	2	97.7
12	5	5	0	100.0	38	36	2	94.7
total	2515	2445	70	83.2	185395	181187	4208	97.7

Table 10: The Frequency Distributions of Tenure for Portion of Travis County, Texas in 2006 Census Test

Label	Tenure	CEF	CANCEIS
0	Vacant	0	0
1	Owned with a Mortgage	58877	58789
2	Owned free and clear	17692	17680
3	Rented for Cash rent	106306	106396
4	Occupied without payment of cash rent	2520	2530
total		185395	185395

Figure 9: The Frequency Distribution Bar Charts of Tenure for Portion of Travis County, Texas in 2006 Census Test



Appendix

A An Example of Edit Specification and DLT

In this appendix, we give an example of FORTRAN pseudo code currently used for sex pre-edits as part of the edit and imputation procedures. We also show the DLT of the sex pre-edits as part of the input files to the CANCEIS software. First, the following is the copy from the 2006 edit specs[6], page 22:

```
D.2 Sex Pre-Edit
PSEX=0
If sum of PSEX(i) is greater than 1, then make PSEX=0; Tally DS(11); go to D.2.a
If PSEX(1) = 1 then PSEX = 1; Tally DS(12)
If PSEX(2) = 1 then PSEX = 2; Tally DS(13)
D.2.a   This section assigns missing sex from the first name file if possible.  For the
        2006 census test, the first name file comes from Census 2000 files.........
        ..........................  Texas should be used for the Austin site.  South
        Dakota population should be used for the Cheyenne site.  Note that if we have a
        first name in any 2006 site that did not occur in these files in 2000, then no
        sex will be assigned from the first name.
FSEX=0
If PSEX=1,2
Then PESEX=PSEX; DS (1);
Elsf PFNAME= blank or PFNAME = 1 character
Then if PFNAME = blank then tally DS(8)or if PFNAME =1 character, then Tally DS(9); fi
               PESEX=0; DS (10) Go to D.3;
Elsf first name one-sex proportion {based on reported sexes with no minimum} is greater
        than or equal to 0.95
then assign that sex from first name.
   Check and then tally when a nonblank PSEX_AR is on the administrative records file
        and is different from sex-from-first-name.  DS(101).
   FSEX=1; {got sex from first name} DS (3)
Elsf we have an administrative records match and sex on the administrative record
Then assign sex from admin records.  FSEX=9.  DS(100)
Else PESEX=ffname(PFNAME); {function to return sex from first name if possible}{Sex is
        assigned based upon the proportional distribution of each name's reported sex
        i.e., to compare random number to the proportion of reported cases - get a new
        random number for each occurrence of the first name} DS(2)
     If PESEX=1,2
     Then FSEX=1; {got sex from first name} DS(3)
     Else PESEX=0; {needs allocation} DS (4)
Fi   Fi
```

In contrast, the Derive DLT is given below:

```
% DLT Name:                  DDLT_SEX
% Strata:                    4
% Purpose:                   Derive
% Type:                      Conflict
% Symmetry:                  No
% Sub-unit Start position:   1
% Sub-unit End position:     4
* common actions
$ FSEX(#1) = As_Reported
* conditions                  1 2 3 4 5 6 7 8
@ PSEX01(#1) = 1             ;Y;N;Y;Y;Y;N;N;N;
@ PSEX02(#1) = 1             ;N;Y;Y;Y;Y;N;N;N;
@ PFNAME(#1) = -1            ; ; ;Y;N;N;Y;N;N;
@ PFNAME(#1) = CLASS(Boys)   ; ; ; ;Y; ; ;Y; ;
@ PFNAME(#1) = CLASS(Girls)  ; ; ; ; ;Y; ; ;Y;
* conditional actions
& QSEX(#1) = Male            ;X; ; ;X; ; ;X; ;
& QSEX(#1) = Female          ; ;X; ; ;X; ; ;X;
& QSEX(#1) = Unknown         ; ; ;X; ; ;X; ; ;
& FSEX(#1) = From_First_Name ; ; ; ;X;X; ;X;X;
```

B The Tenure Hot Deck DLT

```
************************************************************************************
* Description  The following DLT performs the Tenure and Vacancy Status edits
*              in a 4-Person household.
* Edit Specs   2006 Edit Specs
*
* Date         October 31, 2006
* Author       Bor-Chung Chen
************************************************************************************
% DLT Name:    HDDLT_TEN_VAC
% Strata:      4
% Purpose:     Consistency
% Type:        Conflict
% Symmetry:    No
% Sub-unit Start position:          1
% Sub-unit End position:            4
*                                    1 2 3 4
@ TENURE = CLASS(Occupied)          ;Y; ;Y; ;
@ SEQSTATUS = CLASS(Only_Occupied)  ; ;Y;N; ;
@ SEQVACANT = CLASS(Not_Occupied)   ;Y;Y; ; ;
@ TENURE = Not_in_Universe          ; ; ; ;Y;
```

Acknowledgments

The author would like to thank Arthur Cresce of Housing & Household Economic Statistics Division, Inez Chen of Decennial Statistical Studies Division, and Yves Thibaudeau of Statistical Research Division for their reviews and constructive comments on earlier versions of this paper. The author would also like to thank Wesley Benjamin of Statistics Canada for his assistance with questions about the CANCEIS software.

References

[1] CANCEIS Version 4.5.4 USER'S GUIDE. CANCEIS Development Team, Social Survey Methods Division, Statistics Canada, Last Revision: February, 2007.

[2] M. Bankier. Documentation of the New NIM Prototype. Social survey methods division report, Statistics Canada, Ottawa, Dated September 7, 1997.

[3] Bor-Chung Chen. CANCEIS Experiments at U.S. Census Bureau with 2006 Census Test Data (CANCEIS 4.5.4). Technical report, Statistical Research Division, Bureau of the Census, 2007. Manuscript in preparation.

[4] I. P. Fellegi and D. Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71:17–35, 1976.

[5] S. Obenski and J. Farber. Housing Unit Characteristic Imputation Research Results for Short Form Items. Technical report, PRED, U.S. Bureau of the Census, Washington, DC, 2005.

[6] Dan Philipp. 2006 Post Processing Legacy Edit Design–2006SD. Document number: PSS-P-PPLegacyEditDesign-2006SD, U.S. Bureau of the Census, December 13, 2006.