Simplified Census Edit and Imputation Based on Statistical Principles

Robert D. Sands[1]
U.S. Census Bureau, Decennial Statistical Studies Division, Washington, DC
robert.d.sands@census.gov

Abstract

This study proposes a statistically principled random imputation of the relationship, age and sex characteristics of persons enumerated in the 2010 Census. All three characteristics for all persons in a household are imputed simultaneously. First, a series of probability distributions for all valid, completely-classified households with up to eight persons is produced. Next, the expectation-maximization algorithm distributes the counts of partially-classified households to the completely-classified distribution to produce maximum likelihood estimates. A random draw is made from the maximum likelihood distribution of households that exactly matches the partially-classified household's reported variables. A truth deck of persons is used to calculate measures of agreement for each characteristic. A comparison of the proposed system imputation is made with the imputation results of the 2010 Census production system. The proposed system appears to perform comparably to the legacy system but could use some improvement for age and sex imputation.

Key Words: Demographic Characteristics, Contingency Tables, ML Estimation, EM Algorithm, Inter-rater Agreement, Big Data

0. Introduction

Section 1 of this paper gives a brief synopsis of the legacy imputation of relationship to householder, age, and sex characteristics as used in the 2010 Census production. It attempts to distill the essential attributes of the system as well as describe the historical impetus of its design.

In section 2, the statistical theory and design of the proposed system that is the subject of this paper is given. This proposed system is intended to incorporate the essential attributes of the legacy system while having a compact, straight-forward design that is based upon a simple yet powerful statistical model.

Section 3 documents a simulation of the proposed system using the 2010 Census housing unit person data sets. This simulation follows closely the design described in section 2.

Next, section 4 gives the results of the proposed system simulation and compares these results to the production results. The distributions of the three characteristics for the proposed and legacy systems are examined and compared. In addition, the responses of persons in the independent 2010 Census Coverage Measurement (CCM) survey

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

(Viehdorfer 2011) are employed as a "truth deck" for the two systems. The respective levels of agreement of the two systems, respectively, with the CCM truth deck are given.

Finally, section 5 provides a discussion of the results and directions for further research.

1. Legacy System

The legacy U.S. decennial census edit and imputation system (Brunsman 1960; Treat 1994; Sands 2004; Obenski et al. 2005) has over a half-century history at the U.S. Census Bureau. Most recently, the 2010 Census system requirements document (Cardella 2010) has over 100 pages of edit rules, imputation actions, and hot deck matrices dealing with just the "household formation" person characteristics: relationship to householder, age, and sex. As formally documented in Cardella (2010), the edit rules are interwoven with the various deterministic and random imputation actions. Further, the edits and imputations are iterated in a piecemeal fashion. The documentation of the legacy imputation system follows rather closely the structure of the legacy system computer code that was constructed, apparently, in an ad hoc manner since the advent of electronic data processing in the 1960 Census.

Fifty years of effort by Census Bureau demographers and programming specialists has not gone to waste. This author (Sands 2004) found that, as of last decade, the system works quite well. Comparisons of the 2000 Census results with a "truth deck" provided extensive evidence of the level of performance. Embedded within the legacy system are a number of attributes that the author preserves in his research. These are outlined in the following paragraphs.

1.1 Household Formation

First, the household has a structure (or formation) consisting of the values of three census person characteristics among the n persons in the household (Netting et al. 1984). These three census questionnaire items are: i) Relationship to the householder, ii) Age, and iii) Sex. Henceforth, these three characteristics will be referred to, respectively, as REL, AGE, and SEX; perhaps with a suffix denoting which person, i, of the n persons in the household.

A list of logical household formation edits concerning, exclusively, REL, AGE and SEX have been identified and updated through the decades by census demographers. Although Fellegi and Holt (1976) and others have dealt with the fascinating problem of "error localization" (De Waal et al. 2011) concerning which item(s) to impute when a household fails a particular edit rule(s), this will not be a focus of this paper. The demographic expert's decision on which household person characteristics to "blank" when an edit fails is fixed a priori. Furthermore, this blanking is assumed sufficient to render the resulting household (with the now missing items) able to pass all edits. Any failure of all households to pass all edits following the edit and blanking process is evidence of an error condition in the edit rules themselves that must be fixed before proceeding with the imputation.

1.2 Nearest-neighbor Imputation

Second, households with missing characteristics are imputed from distributions of households that are close in both a geographic sense and in the similarity of characteristics.

2. Proposed System Theory and Design

The proposed imputation system employs a multivariate imputation using a saturated log-linear model (Bishop et al. 1975; Little and Rubin 2002; Sands and Griffin 2006) represented as a $3n$-dimensional contingency table of counts of households. Each table cell is indexed by the household vector of *n* persons' REL, AGE and SEX values.

Little and Rubin (2002) describes this contingency table model for handling incomplete categorical data assuming an ignorable missing-data mechanism. The authors describe a maximum likelihood (ML) estimation of the missing data probabilities where the cells of the table form a multinomial distribution. For general patterns of missing data, i.e. where the data can be missing in a household in any combination of characteristics, the ML estimator has no closed-form solution. For this reason, an iterative procedure is used to create the ML estimator for the probabilities.

2.1 Household Characteristic Vector

A household vector of $3n$ characteristics: REL_1 AGE_1 SEX_1 REL_2 AGE_2 SEX_2 … REL_n AGE_n SEX_n is created for each census household of n persons. The REL (15 categories) and AGE (33 categories) characteristics can have the values shown in Tables 1a and 1b, respectively. The characteristic SEX has the values: **M**, **F**.

Table 1a. REL

01	**HH** (Householder)	04	**AC** (Adopted Child)	07	**FM** (Parent)	10	**SL** (Son/Daughter-in law)	13	**HR** (House/Room mate)
02	**SP** (Spouse)	05	**SC** (Step-child)	08	**GC** (Grandchild)	11	**OR** (Other relative)	14	**UP** (Unmarried partner)
03	**CH** (Biological Child)	06	**BS** (Sibling)	09	**PL** (Parent-in-law)	12	**RB** (Roomer/Boarder)	15	**NR** (Other Non-relative)

Table 1b. AGE

01	**0 yrs**	06	**11-13 yrs**	11	**26-28 yrs**	16	**41-43 yrs**	21	**56-58 yrs**	26	**71-73 yrs**	31	**86-88 yrs**
02	**1 yrs**	07	**14-16 yrs**	12	**29-31 yrs**	17	**44-46 yrs**	22	**59-61 yrs**	27	**74-76 yrs**	32	**89-91 yrs**
03	**2-4 yrs**	08	**17-19 yrs**	13	**32-34 yrs**	18	**47-49 yrs**	23	**62-64 yrs**	28	**77-79 yrs**	33	**92+ yrs**
04	**5-7 yrs**	09	**20-22 yrs**	14	**35-37 yrs**	19	**50-52 yrs**	24	**65-67 yrs**	29	**80-82 yrs**		
05	**8-10 yrs**	10	**23-25 yrs**	15	**38-40 yrs**	20	**53-55 yrs**	25	**68-70 yrs**	30	**83-85 yrs**		

The household vectors are partitioned into two types. First, completely classified or donor vectors represent households whose characteristics are fully reported and whose values pass all edits. Second, partially classified or donee vectors represent households for which at least one characteristic is missing due either to nonresponse or to edit failure. Using this setup, all completely classified households, of size n, are represented as a $3n$-dimensional cross-classification wherein each cell holds the count of households with a particular vector of characteristic values. Further, the partially classified households form supplemental marginals of the cross-classification (De Waal et al. 2011).

For example, in a simple hypothetical distribution of two-person households, there are the list of counts, in Figure 1 below. The top four rows are completely classified (donor) household vectors. The first of these represents 1,000 donor households with a female householder 44-46 years old and a male biological child 0 years old. The bottom three rows of Figure 1 are partially classified (donee) household vectors. Each of these donee vectors represents a particular missing data pattern. The first of these represents 50 donee households with a male child in which the age is missing.

Figure 1. Hypothetical Distribution of Two-Person Donor and Donee Households

```
                                                  C
 R    A      S     R    A      S                  O
 E    G      E     E    G      E                  U
 L    E      X     L    E      X                  N
 1    1      1     2    2      2                  T

HH 44-46 yrs F    CH    0 yrs M              1,000
HH 44-46 yrs F    CH    0 yrs F              1,050
HH 44-46 yrs F    CH  5-7 yrs M              4,000
HH 44-46 yrs F    CH  5-7 yrs F              3,950

HH 44-46 yrs F    CH -------  M                 50
HH 44-46 yrs F    CH   0 yrs  -                  5
HH 44-46 yrs F    CH -------  -                 10
```

Table 2 shows how the Figure 1 distribution is represented in a contingency table.

Table 2. Contingency Table of Hypothetical Two-Person Households
(All households have HH 44-46 years F with a CH)

CH	M	F		
0 years	1,000	1,050	2,050	5
5-7 years	4,000	3,950	7,950	
	5,000	5,000	10,000	
	50			10

All census households for each household size $n = 1\ to\ 8$ persons are treated the same way in the current study. While Table 2 shows only a few values in two dimensions of the six dimensions in a two-person household, the census households are actually represented in three dimensions for the one person household up through 24 dimensions for the eight-person household.

In implementation, following Horowitz and Sahni (1987), these high-dimensional contingency tables are each realized as a sparse matrix implemented as a data set of non-

zero cell observations. Both donor vectors and donee vectors are handled this way (Figure 1). Schafer (1997) and Sands and Asiala (2003) applied the sparse matrix approach for the efficient storage of high-dimensional categorical data.

2.2 EM Algorithm Illustration

The iterative procedure selected to generate the ML estimator is the expectation-maximization (EM) algorithm that dates back, at least, to work by Deming and Stephan (1940) for the 1940 Census. In the current use of the EM algorithm, the partially classified counts (shown in red in Table 2) are distributed into the main (green) portion of the table using the completely classified probabilities iteratively until convergence. These probabilities are calculated using the appropriate green cell count divided by the corresponding uncolored cell marginal total.

The steps of the EM algorithm are shown in Tables 3a through 3c using the example from Table 2. For the first iteration, the adjusted counts are calculated as follows:

$$1013.44 = 1000 + \frac{1000}{5000}50 + \frac{1000}{2050}5 + \frac{1000}{10000}10$$

$$1053.61 = 1050 + \frac{1050}{2050}5 + \frac{1050}{10000}10$$

$$4044.00 = 4000 + \frac{4000}{5000}50 + \frac{4000}{10000}10$$

$$3953.95 = 3950 + \frac{3950}{10000}10$$

For the second iteration, the adjusted counts are calculated as follows:

$$1013.48 = 1000 + \frac{1013.44}{5057.44}50 + \frac{1013.44}{2067.05}5 + \frac{1013.44}{10065.00}10$$

$$1053.59 = 1050 + \frac{1053.61}{2067.05}5 + \frac{1053.61}{10065.00}10$$

$$4044.00 = 4000 + \frac{4044.00}{5057.44}50 + \frac{4044.00}{10065.00}10$$

$$3953.93 = 3950 + \frac{3953.95}{10065.00}10$$ and so on.

It takes five iterations to reach convergence in this example. Convergence is defined here as no difference (to eight decimal places) present between each cell value in the current iteration compared to the corresponding cell value in the previous iteration for all cells in the table.

Table 3a. Contingency Table of Hypothetical Two-Person Households (Table 2 repeated) (All households have HH 44-46 years F with a CH)

CH	M	F		
0 years	1,000	1,050	2,050	5
5-7 years	4,000	3,950	7,950	
	5,000	5,000	10,000	
	50			10

Table 3b. First Iteration EM Algorithm On Contingency Table of Hypothetical Two-Person Households (All households have HH 44-46 years F with a CH)

CH	M	F		
0 years	1,013.44	1,053.61	2,067.05	5
5-7 years	4,044.00	3,953.95	7,997.95	
	5,057.44	5,007.56	10,065.00	
	50			10

Table 3c. Fifth Iteration EM Algorithm On Contingency Table of Hypothetical Two-Person Households (All households have HH 44-46 years F with a CH)

CH	M	F		
0 years	1,013.48	1,053.59	2,067.07	5
5-7 years	4,044.00	3,953.93	7,997.93	
	5,057.48	5,007.52	10,065.00	
	50			10

In the current study, for each household size $n = 1\ to\ 8$ persons, the EM algorithm is used to calculate ML estimates using the partially classified vectors from all patterns of missing data.

2.3 Donor Probability Calculation

After the EM algorithm is completed for each household size, the actual imputations can be done.

Using our example, referring to the M column of Table 3c, the 50 households with male biological child will be assigned the AGE:

0 years: $\Pi = \frac{1013.48}{5057.48} = 0.20$

5-7 years: $\Pi = \frac{4044.00}{5057.48} = 0.80.$

Next, referring to 0 years row of Table 3c, the 5 households with a 0-year-old biological child will be assigned SEX:

Male: $\Pi = \frac{1013.48}{2067.07} = 0.49$ Female: $\Pi = \frac{1053.59}{2067.07} = 0.51.$

Finally, referring to the entirety of Table 3c, the 10 households with a child will be assigned AGE, SEX (probabilities are control rounded, to two decimal places, to sum to 1.0):

Male, 0 years: $\Pi = \frac{1013.48}{10065.00} = 0.10$ Female, 0 years: $\Pi = \frac{1053.59}{10065.00} = 0.11$

Male, 5-7 years: $\Pi = \frac{4044.00}{10065.00} = 0.40$ Female, 5-7 years: $\Pi = \frac{3953.93}{10065.00} = 0.39$

3. Proposed System Simulation

Based on the design just described, this section documents the simulation of the proposed characteristic imputation system performed on data collected from the 2010 Census housing unit population.

3.1 Identify Legacy System Edit Rules

The first step in this simulation of the proposed system was to identify the edit rules for REL, AGE and SEX employed in the legacy system. After a close examination of the legacy system documentation, 14 edit rules were identified. These 14 rules concern potential inconsistencies among, exclusively, the REL, AGE and SEX characteristics within a household. The set of rules collectively define the permissible structure of a household as defined by the three characteristics.

For this simulation, in the 2010 production census data set, all person characteristics that had non-zero edit flag values, i.e., either the characteristic was missing or the characteristic was not valid as reported, were blanked. In the current set-up, following the edit process, all households (with or without blanked characteristics) do not violate any of the 14 edits.

The proposed system began the imputation of new values only when the edit process had completed. Consequently, the imputation methodology is independent of any particular set of edit rules. The edit rules themselves as well as the choice of characteristic(s) to blank upon an edit failure were left to the demographic experts.

3.2 Create Household Characteristic Vectors

The second step in the simulation of the proposed imputation system was to form the household vectors of $3n$ characteristics: REL_1 AGE_1 SEX_1 REL_2 AGE_2 SEX_2 … REL_n AGE_n SEX_n for each 2010 Census household of $n = 1 - 8$ persons. As described previously in section 2.1, this was done by using the 15 relationships for REL, 33 age categories for AGE and two sexes for SEX. Characteristics that were blank were designated as such. The count of households that had each particular combination of values was attached to each household *vector*. This vectorization of the households enabled a significant improvement in processing efficiency.

3.3 Partition Household Vectors into Donor and Donee

The third step in the simulation of the proposed system was to separate the household vectors not requiring imputation from those that did. The household vectors without blank characteristics are referred to as completely classified or donor vectors. Conversely, the household vectors with one or more blank characteristics are referred to as partially classified or donee vectors.

3.4 Match Donor and Donee Vectors

The fourth step was to link each donee vector with all possible donor vectors by matching on the characteristics (if any[2]) reported in the donee household vector. This accomplishes two things. First, since all donee and donor vectors pass all edits, the selection of a donor household vector is guaranteed to result in a valid imputation in a single step. Second, all characteristics missing in the donee vector will be replaced at the same time. This is not only efficient, but it also maintains the multivariate distribution of the three characteristics among donee households.

In practice, the subsets of donor household vectors linked to each donee vector are not only numerous and often large but frequently overlap with one another. For this reason, a system of indexing using pointers (Horowitz and Sahni, 1987) is embedded in the implementation of the proposed system so that both computer run times and memory use are minimized. Although this data processing issue is "under the hood" and does not alter the logical design of the proposed system, it is noted when discussing the processing of some 116.7 million occupied census households in a reasonable time.

Further, in the simulation, especially for larger household sizes, there were a significant number of cases where donee vectors could not be matched to a donor vector even at the U.S. level. An additional sub-process was implemented to create "synthetic" donor vectors for each donee vector lacking a match to actual donor vector. A sample of approximately 100 synthetic donors of all possible synthetic donor vectors was created for each non-matched donee vector. Each synthetic vector was given a "count" of 10^{-12} that was then weighted by the relative occurrence of the REL(s), assigned to the synthetic vector, in the 2010 Census.

3.5 Calculate Donor Vector Probabilities

The fifth step in the simulation of the proposed system was to calculate the probabilities for the household donor vectors linked to each donee vector. This step included the implementation of the EM algorithm, described in Section 2.2, that distributed, until convergence, the donee vector counts into the appropriate donor counts within a particular geographic area, i.e. tract or U.S. The ML probabilities were then produced for all household sizes up to eight persons as described in Section 2.3.

3.6 Impute Donee Vectors

The sixth step in the 2010 Census simulation was to make the actual imputations for the donee vectors. Random draws were made from the ML probabilities to choose a donor vector for a donee vector.

3.6.1 Tract-level Processing

First, steps 3.4 and 3.5 were conducted at the tract level. The individual donee households in a particular tract were matched to individual donor(s) in the same tract. In general, donee vectors with more common missing data patterns could be matched to donor

[2] If a donee vector has all blank characteristics then any donor vector the same size is a match.

vectors at the tract level. The imputations were then made for the matched donees from the tract-level donor vector distributions.

3.6.2 U.S.-level Processing

Following the tract-level processing, steps 3.4 and 3.5 were repeated at the U.S.-level for all donee households still requiring imputation. Imputations were then made for these donees from the U.S.-level donor distributions. In addition, the U.S.-level donor distributions included any required synthetic donor vectors to impute donee vectors previously identified as having no matching donor vectors in the U.S. No synthetic donor vectors were used at the tract-level.

3.7 Match Donee Households to "Truth" Deck

The final step in the 2010 Census simulation was to match the actual imputations for the donee vectors to the truth deck formed from respondents in the 2010 CCM survey (Viehdorfer 2011).

The CCM was a completely independent (of the 2010 Census operations) enumeration of an approximately $\frac{1}{750}$ sample of the U.S. housing unit population. The CCM enumeration of persons (also known as the P sample) was conducted by personal visit and collected the same demographic information as the 2010 Census. The CCM also collected information to determine the Census Day residence of the P-sample person. The P-sample persons were then linked to the census person roster using a matching system that has, historically, a 99.8% true match rate (Bean 2001). Consequently, the CCM provides an independent verification of person characteristics in the census.

Persons in the CCM P sample that were i) matched, by CCM, to 2010 Census persons and ii) reported in the CCM the characteristic that was missing in the census formed the subset in the truth deck comparison of characteristics imputed, respectively, by the two systems.

4. Proposed System Simulation Results

The proposed imputation system was simulated using the 2010 Census housing unit person data set. After the edit and imputation phases of the proposed system were complete, various distributions of the legacy (2010 Census production) and proposed system imputations, of the three characteristics, were examined and compared. Next, the agreement of the imputations of the two systems, separately, with the truth deck were examined.

4.1 Total Household Counts

There were 116.7 million total occupied households in the 2010 Census. Of these, 116.2 million households contained from one through eight persons.

The 116.2 million households in this simulation were subdivided into two groups:

i) 105.5 million donor households (all characteristics reported and passed all edits)

ii) 10.7 million donee households (at least one characteristic missing or blanked by an edit failure).

Less than 50,000 donee households could not be processed by the proposed system. These fell into one of two categories. The first relatively small category contained residual edit failures. The second category contained, predominantly, larger households that had rare missing data patterns for which there were no donors and, further, the number of possible synthetic donors was greater than 10 million. The problem concerning processing larger households and the combinatorial issues with the proposed system is discussed in section 5.

4.2 Donee Household Results

The 10.7 million donee households were further divided into two groups:

i) 2.5 million donee households with all characteristics missing (whole household substitutions)
ii) 8.2 million donee households with one or more characteristics reported

4.3 Donee Whole-Household Substitutions

The 2.5 million donee households with all characteristics missing (whole-household substitutions) were nearly all (99.98%) processed at tract level. This undoubtedly occurred because these donee households, with no characteristics reported, only needed a "matching" donor household in the tract that had the same number of persons. The remaining 0.02% of whole-household substitutions were processed at the U.S. level because, in these relatively few instances, the particular tract contained no donor households of the same size.

There were 5.7 million persons that had all three characteristics; REL, AGE and SEX imputed in these substitution households. Since the persons in these households reported no characteristics they could not be matched to the CCM P sample. Therefore, these cases could not be evaluated by the truth deck. This left only the overall distributions imputed by the two systems available for comparison.

The mean age imputed by the proposed system for these persons was 2 years higher than the mean age imputed by the legacy system. Further, the number of females imputed by the proposed system was 51.5% versus 49.5% female imputed by the legacy system. Finally, the number of spouses imputed by the proposed system was 0.2 million higher than the legacy system.

Since it is, historically, unlikely that the census is undercounting older females and spouses it would appear that the legacy system outperformed the proposed system in the whole household substitution cases.

4.4 Donee One or More Characteristics Reported Households

The 8.2 million donee households with one or more characteristics reported had 4.3 million persons imputed REL, 10.0 million persons imputed AGE, and 0.8 million

persons imputed SEX. A particular person may be included in any combination of these three counts.

The subset of donee households, with one or more characteristics reported, who were matched to CCM P-sample persons and who reported the particular characteristic, was much smaller than the full set of donee households due to the relatively small sample in CCM. For these CCM-matched donee households there were: 3,150 persons imputed REL, 7,192 persons imputed AGE, and 450 persons imputed SEX. Tables 4a through 4c compare the distributions of REL, AGE, and SEX respectively, for the two imputation systems to the CCM reported distribution.

4.4.1 REL Imputation

Table 4a. Un-weighted Percent Distribution of REL for CCM Reported, Legacy Imputed, and Proposed Imputed for 3,150 Matched Persons in Donee Households

REL	CCM REPORTED PERCENT	LEGACY IMPUTED PERCENT	PROPOSED IMPUTED PERCENT
SP	16.52	14.96	15.87
CH	46.09	46.63	46.48
AC	1.18	0.64	1.56
SC	2.35	1.81	3.46
BS	2.29	3.43	2.89
FM	2.32	1.43	1.24
GC	7.34	10.20	6.98
PL	0.60	0.73	0.48
SL	1.02	0.70	0.73
OR	6.96	8.23	6.35
RB	1.24	1.56	2.03
HR	3.21	2.86	3.84
UP	3.53	2.10	3.40
NR	5.37	4.73	4.70

The proposed system appeared to get as close or even closer to the CCM distribution of REL than did the legacy system. For the four largest REL values (see Table 1a for key to abbreviations): SP, CH, GC, and OR, the proposed system was closer to the true number than was the legacy system. The proposed system seemed to be off relative to the legacy system for SC. Both systems undercounted FM, SL, and NR.

4.4.2 AGE Imputation

Table 4b. Un-weighted Mean AGE for CCM Reported, Legacy Imputed, and Proposed Imputed for 7,192 Matched Persons in Donee Households

CCM REPORTED AGE	LEGACY IMPUTED AGE	PROPOSED IMPUTED AGE
36.59	39.59	40.74

Neither the legacy system nor the proposed system did an accurate job of imputing AGE for CCM matched persons. The legacy system outperformed the proposed system by imputing a mean age 3 years higher than the truth whereas the proposed system is 4 years higher.

4.4.3 SEX Imputation

Table 4c. Un-weighted Percent Distribution of SEX for CCM Reported, Legacy Imputed, and Proposed Imputed for 450 Matched Persons in Donee Households

SEX	CCM REPORTED PERCENT	LEGACY IMPUTED PERCENT	PROPOSED IMPUTED PERCENT
M	42.89	45.78	49.78
F	57.11	54.22	50.22

Neither the legacy system nor the proposed system did an accurate job of imputing SEX for the matched truth deck persons. The legacy system outperformed the proposed system in this case by underestimating the percentage of females by only 3% versus the proposed system's underestimate of females of 7%. It is noted that the majority (57%) of matched CCM persons reporting SEX (but missing SEX in the census) are female. This seems to indicate not only a not-missing-at-random (NMAR) condition but also an undercount of females by the two imputation schemes.

4.4.4 Kappa Measure of Agreement

The Kappa statistic (Agresti 2002) was used as a measure of inter-rater agreement where, one "rater" was the imputed characteristic value in the 2010 Census and the other "rater" was the value reported by the same person in the CCM survey (the truth deck). Kappa can take any value between -1 and 1 inclusively. A score of 1 means perfect agreement, 0 means chance, and -1 designates perfect disagreement. Although mathematically possible, negative κ values, and especially those near -1, are unlikely in any practical situation.

The two formulae below are for the unweighted Kappa (1) and weighted Kappa (2). Weighted Kappa takes into consideration the distance between (ordinal) categories. In the 2×2 case $(I = 2)$, κ and κ_w would give the same result.

$$\kappa = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}} \qquad (1)$$

$$\kappa_w = \frac{\sum\sum w_{ij}\pi_{ij} - \sum\sum w_{ij}\pi_{i+}\pi_{+j}}{1 - \sum\sum w_{ij}\pi_{i+}\pi_{+j}} \ , w = 1 - \frac{(i-j)^2}{(I-1)^2} \ , I \ values \qquad (2)$$

Table 5a shows the calculation of the Kappa measure of agreement for the proposed system SEX imputation versus the truth deck.

Table 5a. CCM Matched Persons by Proposed System Imputed SEX and CCM Reported SEX

CCM Reported SEX	Proposed System Imputed SEX		
	M	F	Total
M	101	92	193
F	123	134	257
Total	224	226	450

$$\kappa = \frac{\frac{101+134}{450} - \frac{193 \cdot 224 + 257 \cdot 226}{450^2}}{1 - \frac{193 \cdot 224 + 257 \cdot 226}{450^2}} = \frac{0.5222 - 0.5003}{1 - 0.5003} = 0.0439$$

Table 5b shows the truth deck diagonal agreement and Kappa measures for the legacy system and proposed system for each of the three characteristics. These results show the Kappa scores to be virtually identical for the two systems for all three characteristics. From the standpoint of the "truth" as measured by these matched persons in the CCM who did not respond in the census, the two imputation systems perform the same.

Table 5b. Measures of Agreement of the REL, AGE, and SEX Values Imputed by the Proposed and Legacy Systems with the Reported Values for CCM Matched Persons

Characteristic	n	Imputation System	π_{ii}	κ	κ C.I.
REL	3,150	Legacy	0.5304	0.3664	(0.3458, 0.3870)
		Proposed	0.5231	0.3582	(0.3382, 0.3781)
AGE	7,192	Legacy	0.1623	0.4568*	(0.4443, 0.4692)
		Proposed	0.1623	0.4514*	(0.4388, 0.4640)
SEX	450	Legacy	0.5133	0.0148	(-0.0774, 0.1071)
		Proposed	0.5222	0.0439	(-0.0476, 0.1353)

* κ_w adjusts the agreement using the distance between ordinal AGE categories

5. Discussion

The 2.5 million donee households with all characteristics missing (whole-household substitutions) contained 5.7 million persons that had all three characteristics; REL, AGE and SEX imputed. Nearly all these persons were imputed with tract-level donors. Even though this local area of tract was used (and not the entire U.S.) it would appear that the tract was not local enough. By using, literally, the nearest-neighbor donor household, the legacy system correctly imputed fewer older females and spouses for the whole household cases than the proposed system. By using donor household distributions at the block-level, the proposed system might more accurately capture the nearest-neighbor mechanism. Perhaps the addition of an explicit nearest-neighbor variable like nearest-neighbor type of household that was used in the 2010 Census count imputation system (Thibaudeau 2002; Sands and Griffin 2006) would do a better job.

For the 8.2 million donee households with one or more characteristics reported, the CCM matched distribution of 3,150 persons imputed REL, 7,192 persons imputed AGE, and 450 persons imputed SEX was used to gauge the performance of the two imputation systems.

In this simulation, it appears that neither system can impute SEX properly. It has been shown in a previous study by the author (Sands 2004) that the only reliable way to impute SEX is from the person's first name ($\kappa = 0.90$). Any other random approach is not demonstrably different from chance. Regarding the imputation of AGE, the two methods do better than they did for SEX; with measures of agreement about halfway in between chance and perfection. When the distribution of AGE imputed by the two methods for the 7,192 truth deck cases is compared, the mean age for the legacy and proposed systems is high by 3 and 4 years, respectively. Since both imputation systems are imputing too old and both assume that donor households have the same distribution as donee households

suggests a not-missing-at-random (NMAR) missing-data mechanism (Little and Rubin 2002) is present with AGE. In the case of AGE, it appears that missingness is related to the AGE of the person. A similar phenomenon could also be happening with SEX.

Future research should concentrate on this NMAR issue, with AGE and SEX imputation, by using other data sets that are matched to the census such as the American Community Survey as truth decks. Other simulation methodologies employing multiple, synthetic truth decks could also be applied to this question. The author believes the main point here is to "spare no expense" in using the 2010 Census production and allied data sets to simulate all theories of missing data in order to improve and rationalize characteristic imputation in the census.

The combinatorial challenges for larger households provide a processing problem for the proposed system but this issue should diminish over time as computers continue to grow exponentially in size and power while households continue to shrink in size.

Finally, this project had as its goal to demonstrate what could be done with a simple yet powerful statistical model in comparison with the complicated but effective legacy imputation system. This was accomplished. Furthermore, the proposed system appears to perform comparably to the legacy system but could use some improvement for AGE. This discrepancy, in the proposed system, could probably be ameliorated by using a smaller area than tract to create household vector distributions or by inserting into the proposed system model a variable to capture the local area nearest-neighbor relationship. This last alternative, however, would make the model less "simplified".

Acknowledgements

I would like to thank my wife, Judy, for standing by me during this long endeavor. I also would like to thank my Census Bureau colleagues: Pat Cantwell, Inez Chen, Debbie Fenstermaker, Andy Keller, Ben Klemens, Rolando Rodriguez, and Yves Thibaudeau for their helpful comments regarding this project.

References

Agresti, A. (2002), *Categorical Data Analysis, Second Edition*, New York: Wiley-Interscience.

Bean, S.L. (2001), "ESCAP II: Accuracy and Coverage Evaluation Matching Error," A.C.E. REVISION II Memorandum Series #PP-5.

Bishop, Y.M.M., Feinberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: The MIT Press.

Brunsman, H.G. (1960), "Processing and Editing the data from the 1960 Census of Population," U.S. Department of Commerce, Bureau of the Census, *Presented at* May 1960 PAA Meetings Washington, DC.

Cardella, M.A. (2010), "Requirements for Editing 100% Items 2010 Census – Baseline v. 1.0," Internal Report, Washington, DC: U.S. Census Bureau.

Deming, W.E., and Stephan, F.F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known," *The Annals of Mathematical Statistics*, **11**, No. 4, 427-444.

De Waal, T., Pannekeok, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*, Hoboken, NJ: Wiley.

Fellegi, I.P., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of American Statistical Association*, **71**, 17-35.

Horowitz, E. and Sahni, S. (1987), *Fundamentals of Data Structures in Pascal*, *Second Edition*, Rockville, MD: Computer Science Press.

Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, New York: Wiley.

Netting, R.M., Wilk, R.R., and Arnould, E.J. (1984), *Household; Comparative and Historical Studies of the Domestic Group*, Berkeley: University of California Press.

Obenski, S., Farber, J., and Chappell, G. (2005), "Research to Improve Census Imputation Methods: Item Results and Conclusions," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3447-3454.

Sands, R.D. (2004), "Comparison of the Characteristic Imputation in 2000 Census to the Accuracy and Coverage Evaluation Survey for Matched Persons," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 4304-4310.

Sands, R.D. and Asiala, M.E. (2003), "A Simple and Efficient Approach to the Cross-Tabulation of Large SAS® Data Sets," *Conference Proceedings*, NESUG 2003.

Sands, R.D. and Griffin, R.A (2006), " 2010 Census Count Imputation – Research Results using Spatial Modeling," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3641-3647.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall.

Thibaudeau, Y. (2002), Model Explicit Item Imputation for Demographic Categories. *Survey Methodology*, **28**, 135-143.

Treat, J. (1994), "Summary of the 1990 Census Imputation Procedures for the 100% Population, Population and Housing Items," DSSD 1990 Census REX Memorandum Series #BB-11.

Viehdorfer, C. (2011), "The Design of the Coverage Measurement Program for the 2010 Census – REVISION #1," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-B-07-R1.