HANDLING OF MISSING DATA IN THE CENSUS 2000 DRESS REHEARSAL INTEGRATED COVERAGE MEASUREMENT SAMPLE

Anne Kearney, and Michael Ikeda, Bureau of the Census*
Anne Kearney, Statistical Research Division, Bureau of the Census, Washington, DC, 20233

Key Words: Noninterview Adjustment, Imputation, Modeling

A. Introduction

This paper outlines procedures used to handle missing data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement (ICM) sample. It also provides a summary of the results of missing data processing. A noninterview adjustment procedure, outlined in Section C, is used to account for whole-household nonresponse. A characteristic imputation procedure, outlined in Section D, is used to assign values for specific missing demographic variables. Finally, persons with unresolved match, residence, or enumeration status have probabilities assigned based on a procedure outlined in Section E. The missing data procedures are generally similar in effect to those used for ICM in the 1996 Community Census and the 1990 Post-Enumeration Survey (PES). Methodologies and analysis of procedures are documented in [1] for the 1990 PES, in [5] for the 1995 ICM, and in [4] for the 1996 ICM. Differences between the Dress Rehearsal ICM missing data procedures and those for 1990, 1995, and 1996 are outlined in [3].

Section B gives some general background. Section F includes results from missing data processing and discussion of their implications. Section G contains conclusions.

B. General Background

The Census 2000 Dress Rehearsal is conducted in three areas: Sacramento, CA; Menominee, WI; and Columbia, SC. The South Carolina site is divided into two subsites for the purposes of ICM sample selection and ICM missing data processing. The ICM sample is selected separately for each site and the two subsites. An overview of the ICM sample design for the Dress Rehearsal can be found in [6]. A general overview of ICM operations in the Dress Rehearsal can be found in [2].

The Dress Rehearsal uses Dual System Estimation (DSE) to calculate estimates. DSE tries to obtain a roster from the ICM blocks independently of the Census. The independent roster (P-Sample) and the Census roster (E-Sample) are matched and the results of the matching are used to estimate the number of persons missed by both rosters. Estimates are calculated separately for population subgroups called poststrata. Poststratum estimates are summed to marginal totals which are used to calculate the final estimates. The Dress Rehearsal uses a DSE method called PES C. PES C uses person inmovers in the P-Sample poststratum estimates and uses person outmovers to obtain poststratum estimates of match probability for person inmovers. Further details on DSE estimation for the Dress Rehearsal can be found in [7].

C. Noninterview Adjustment

Noninterview adjustment is only performed on the P-Sample. The noninterview adjustment procedure is similar to the procedures use in the 1990 PES and the 1995 and 1996 ICM. However, there are two noninterview adjustments in the Dress Rehearsal because of the use of PES C estimation. The two noninterview adjustments are basically identical to each other, except for the reference date. One noninterview adjustment is based on housing unit status as of Census Day. The other noninterview adjustment is based on housing unit status as of the day of ICM interview. Each noninterview adjustment spreads the weights of noninterviewed units over interviewed units in the same block cluster and similar type of basic address. There are collapsing rules if the number of interviewed units (in the block cluster x type of basic address category) is too small compared to the number of noninterviewed units. Person nonmovers and person outmovers are used to determine Census Day housing unit status. Person nonmovers and person inmovers are used to determine ICM interview day housing unit status.

Interview: A unit is an interview (for the given reference date) if there is at least one person (with name and at least one demographic characteristic) who possibly or definitely was a resident of the housing unit on the given reference date.

Noninterview: An occupied housing unit (as of the given reference date) that is not an interview is a noninterview.

The noninterview adjustment based on Census Day is used to adjust the weights of person nonmovers and person outmovers. The noninterview adjustment based on day of ICM interview is used to adjust the weights of person inmovers.

D. Characteristic Imputation

P-Sample characteristic imputation for the Dress Rehearsal is similar to characteristic imputation for the 1990 PES and the 1996 ICM. In a change from both 1990 and 1996, we use the demographic information from the Dress Rehearsal Census edited file (CEF) for the Dress Rehearsal E-Sample. Edits and imputation are performed on this file. All E-Sample persons matched to the CEF in the Dress Rehearsal. Because of this, no ICM imputation was done in the Dress Rehearsal E-Sample. If we had needed to do ICM imputation in the E-Sample, the methodology would have been basically the same as the P-Sample methodology.

The variables imputed in the Dress Rehearsal are race, Hispanic origin, sex, tenure, and age. P-Sample person mover status is not considered when imputing characteristics. However, persons from a P-Sample whole-household outmover interview are considered to be a separate household for imputation purposes. Age and sex distributions are calculated separately by site.

Tenure is imputed from the previous household with a similar type of basic address (structure code in the E-Sample) with tenure recorded. Missing race is imputed from the distribution of race in the same household. If no one in the household has a nonmissing value of race, then the distribution of the nearest previous household with reported race and similar Hispanic origin is used. Hispanic origin is imputed from the distribution of Hispanic origin in the same household (or the nearest previous household with reported Hispanic origin and similar race if no one in the household has nonmissing Hispanic origin). Age is imputed from the distribution of age for persons with similar relationship to reference person, and age of reference person. For one-person households, age is imputed from the distribution of age in one-person households.

Sex of reference person (with spouse present) or spouse of reference person is imputed by assigning the person with a missing value for sex the sex opposite to that of their spouse. If both reference person and spouse have sex missing, then sex for the reference person is imputed from the distribution of sex for reference persons with spouse present. The spouse is then assigned the sex opposite to that of the reference person. For one-person households, sex is imputed from the distribution of sex in one-person households. For the reference person (with no spouse present) of a multi-person household, the distribution of sex for reference persons of multi-person households with no spouse present is used. For persons (except reference persons and spouses) from multi-person households with non-missing relationship, sex is imputed from the distribution of sex for persons (excluding reference persons and spouses) from multi-person households. For persons from multi-person households with missing relationship, sex is imputed from the distribution of sex for persons (excluding reference persons) from multi-person households.

E. Assigning Match, Residence, and Correct Enumeration Probabilities

Probabilities for persons with unresolved final Census Day residence (P-Sample), final match (P-Sample), or final correct enumeration (E-Sample) status are estimated by calculating weighted ratios based on persons with resolved final status. Ratios are calculated separately for each site and use the ICM sampling weights. The use of ratios to estimate all three probabilities is new for the Dress Rehearsal. In 1996, hierarchical logistic regression was used to model residence and correct enumeration probability and in 1990 and 1995 hierarchical logistic regression was used to model match and correct enumeration probability.

For Census Day residence status, three separate ratios are calculated. The residence probability for unresolved persons needing followup is the proportion of persons needing followup who are residents. The residence probability for unresolved persons who did not need followup is the proportion of persons not needing followup who are residents. The residence probability for persons with insufficient data for matching is the proportion of all persons who are residents. The proportions are based on person nonmovers and person outmovers with resolved final residence status. The Census Day residence probability for person inmovers is irrelevant to estimation and was set to 0. Note that the residence probability as of the date of ICM interview for person inmovers and person nonmovers is assumed to be 1 (except that infants born after Census Day are not considered to be ICM interview day residents).

Some person nonmovers and person outmovers have unresolved match status. The match probability for these persons is the proportion of matches among person nonmovers and person outmovers with resolved final match status (excluding confirmed Census Day nonresidents). The match probability is set to 0 for confirmed Census Day nonresidents. The match probability for person inmovers is irrelevant to estimation and was set to 0.

For E-Sample persons with unresolved enumeration status, the correct enumeration probability is the proportion of correct enumerations (among persons with resolved enumeration status) in the given match code group. E-Sample match code groups are defined by before-followup match code, whole/partial match code,

address code (HU match status from HU matching), and DSE followup status.

Special Cases

Large clusters were subsampled in the Dress Rehearsal. If an E-Sample person is duplicated with K persons subsampled out of the E-Sample, then the initial correct enumeration probability is multiplied by 1/(K+1), since we do not know which person is the "real" person.

A surrounding block search was done in a small number of outlier clusters. Surrounding blocks in Sacramento were generally eligible for Nonresponse Followup (NRFU) and Undeliverable As Addressed vacant (UAA) sampling. If a P-Sample person matched to a surrounding block person from the NRFU or UAA sample, then the match "probability" of the P-Sample person was set equal to the NRFU or UAA weight of the surrounding block person. There were no E-Sample persons duplicated in a surrounding block in the Dress Rehearsal. If an E-Sample person had been verified to belong in a surrounding block and also to be duplicated with a surrounding block person in the NRFU or UAA sample, then the E-Sample correct enumeration "probability" would have been set to one minus the NRFU or UAA weight of the surrounding block person.

F. Results

All counts in this document are unweighted counts. Certain tables display results for only Sacramento. This is in the interest of space and because the results from the other sites were similar.

1. Noninterview Adjustment

Table 1 gives the noninterview rate by site for Census Day interview status and ICM interview day interview status. Noninterview rates based on Census Day status tend to be higher than noninterview rates based on ICM interview day status because all person nonmovers and inmovers (except for persons born after Census Day) are assumed to be ICM interview day residents, while there are residence questions and other operations that can make person nonmovers and outmovers Census Day nonresidents.

Table 1: Noninterview Rates

	Census Day Status		Interview Day Status	
	NI Rate (%)	Occ HU	NI Rate(%)	Occ HU
Sacramento	5.07	15087	2.18	15217
Rural SC	4.04	7377	1.39	7391
Columbia	6.23	8417	2.02	8398
Menominee	1.68	416	0.17	595

NI Rate is the noninterview rate. Occ HU is the total number of occupied housing units.

2. P-Sample Characteristic Imputation

Table 2 gives the item imputation rates for Sacramento for the five variables that were imputed. Rates are given for three sets of persons. The first set consists of all persons that are included with nonzero weights somewhere in the P-Sample portion of the Dress Rehearsal estimate. Specifically, this includes person nonmovers who are Census Day residents or possible residents from interviewed households based on Census Day interview status, person inmovers from interviewed households based on ICM interview day interview status, and person outmovers from interviewed households based on Census Day interview status. The second set consists of those person inmovers included in the first set. The third set consists of those person outmovers included in the first set.

In general, the imputation rate for inmovers is slightly higher than the overall rate, while the imputation rate for outmovers is substantially higher than the overall rate for age, Hispanic origin, and race. This is probably due to outmover data often being collected by proxy.

Table 2: Item Imputation Rates (Percent)

Sacramento	All	Inmovers	Outmovers
Tenure	0.66	1.39	0.56
Sex	0.41	0.55	1.86
Age	2.14	2.79	8.90
Hispanic Origin	1.29	1.82	11.04
Race	2.13	2.53	13.01
Total persons	37968	2368	1775

3. E-Sample Characteristic Imputation

The variables needed to assign poststrata (tenure, race, Hispanic origin, age, and sex) were obtained from the Census Edited file. Because of this, there was no missing data for these variables and no actual E-Sample imputation was done by the ICM missing data system.

4. Modeling for Unresolved Status

General Overview

Table 3 gives information on the proportion of persons with unresolved status. Note that P-Sample persons with insufficient information for matching are unresolved for both residence status and match status, as are P-Sample persons with a final code of possible match. The proportion of unresolved persons is fairly small. Results from the 1995 ICM [8], [9], [10] suggest that the method for modeling for unresolved status does not have a major

effect on the estimates. Note that there was a substantially higher proportion of unresolved persons in the 1995 ICM, since roughly half of the persons needing followup were sampled out of followup in 1995.

Table 3: Unresolved Status

Note that a few P and E-Sample persons identified as not needing followup have unresolved final status.

a. Percent Unresolved (Overall)

	P-Sample		E-Sample
	UR	Insuff	UR
Sacramento	3.10	1.17	3.64
Rural SC	1.55	0.66	1.47
Columbia	2.51	1.04	3.33
Menominee	1.54	1.43	0.50

P-Sample percentages are percentages of Census Day residents and possible residents from interviewed households (based on Census Day interview status). E-Sample percentages are percentages of E-Sample persons. P-Sample UR refers to persons with unresolved final residence status. E-Sample UR refers to persons with unresolved enumeration status before accounting for duplication with persons subsampled out of the E-Sample. Insuff indicates insufficient information for matching.

b. Persons Sent to DSE Followup

	P-Sample			E-Sample	
	Tot	%UR	%UR M	Tot	%UR
Sacramento	3306	20.7	0.5	5470	21.68
Rural SC	1144	13.1	0.6	2247	10.95
Columbia	1146	19.3	0.4	3205	17.72
Menominee	86	15.1	0.0	101	5.94

P-Sample Tot is the number of residents and possible residents sent to followup. P-Sample percentages are percentages of P-Sample Tot. P-Sample UR are unresolved residents, UR M refers to unresolved match status after followup. E-Sample Tot is the number of E-Sample persons sent to followup. E-Sample percentages are percentages of E-Sample Tot. E-Sample UR are unresolved enumeration status.

P-Sample

We see in Table 4 that DSE followup in the Dress Rehearsal resolved the match status of almost all persons sent to followup in Sacramento. We also see that DSE followup almost never changed a before followup match to a nonmatch (except for before followup matches to surrounding blocks) and rarely changed a before followup nonmatch to a match. Possible matches could become either matches or nonmatches (but more frequently became matches). Note that confirmed nonresidents are not in the table. DSE followup confirmed 551 persons as nonresidents in Sacramento.

Table 4: Before Followup Match Code and Final Match Code for P-Sample Persons Sent to Followup (Except for Confirmed Nonresidents)

	Sacramento Final Match Code							
BFU Match Code	MR	MS	MU	NR	NU	P	KP	Total
Match (M)	202	0	71	0	1	0	0	274
Match Sur Bl (MS)	0	2	0	0	9	0	0	11
Nonmatch (NP)	33	0	0	2167	570	1	1	2772
Poss Match (P)	165	0	6	53	9	16	0	249
Total	400	2	77	2220	589	17	1	3306

MR is matched resident, MS is matched resident, matched to person in surrounding block, MU is matched with unresolved residence status, NR is nonmatched resident, NU is nonmatched with unresolved residence status, P is possible match, KP is match not attempted due to incomplete or invalid name

Table 5 shows the estimated residence probabilities assigned to persons with unresolved residence status in each site.

Table 5: Estimated Residence Probabilities

	Site			
Followup Status	Sacramento	Rural SC	Columbia	Menominee
Sent	0.826	0.808	0.742	0.771
Not Sent	0.991	0.988	0.989	0.985*
Insuff Info	0.976	0.976	0.972	0.969

* No unresolved persons in this category

For illustration purposes, Table 6 contains the counts of confirmed residents, confirmed nonresidents, and unresolved persons by match code group for Sacramento. All persons included in the residence probability calculations are included in Table 6. The proportion resident in BFUGP 1 (matches and possible matches sent to followup) tends to be somewhat higher than for other persons sent to followup; the proportion resident in BFUGP 3 (whole household nonmatches) tends to be somewhat lower.

Table 6: Residence Status by Match Code Group

	Sacramento			
BFUGP*	Resident	Nonres	Unresolved	% Resident of Resolved
1	422	18	112	95.91
2	1495	237	284	86.32
3	705	296	288	70.43
4	31779	276	3	99.14
5	0	0	421	---

*BFUGP is the P-Sample Match Code Group. BFUGP 1-3 are sent to followup. BFUGP 1 are matches and possible matches. BFUGP 2 are partial household nonmatches. BFUGP 3 are whole household nonmatches. BFUGP 4 are persons resolved before followup. BFUGP 5 are persons who have insufficient information for matching (before followup status).

Table 7 gives a further breakdown of residence status for BFUGP 3 (whole household nonmatches) for Sacramento. A conflicting household is where the housing unit matched and both the P-Sample and E-Sample collected persons but none of the persons in either the P-Sample or E-Sample households were matches or possible matches. The proportion resident for persons from conflicting households tends to be lower than for the other persons from BFUGP 3.

Table 7: Residence Status for Whole Household Nonmatches

	Sacramento			
Address Code	Resident	Nonres	Unres	% Resid of Resolved
HU Matched	215	29	101	88.11
HU Not Matched	57	6	66	90.48
Conflicting HH	433	261	121	62.39

The "HU Matched" row excludes persons from conflicting households.

Table 8 contains the estimated match probabilities assigned to persons with unresolved match status. Most of the persons with unresolved match status are persons with insufficient information for matching (the others are possible matches).

Table 8: Estimated Match Probabilities

	Sacramento	Rural SC	Columbia	Menominee
Est Match Prob	0.780	0.727	0.843	0.829

Table 9 contains the counts of confirmed matches, confirmed nonmatches, and persons with unresolved match status by person mover status for Sacramento. Most persons with unresolved match status have insufficient information for matching. Persons in Table 9 are Census Day residents or possible residents. The proportion matched tends to be slightly lower for person outmovers than for person nonmovers.

Table 9: Match Status by Person Mover Status

	Sacramento			
	Match	Nonmatch	Unres	% Match of Resolved
Nonmover	26528	7045	256	79.02
Outmover	901	591	183	60.39

E-Sample

Table 10 shows the estimated initial correct enumeration probabilities assigned to persons with unresolved initial enumeration status in each site. Initial correct enumeration probabilities are later modified to account for duplication with persons subsampled out of the E-Sample. Confirmed erroneous enumerations could have had their probabilities modified for duplication in surrounding blocks if there had been any such duplication.

Table 10: Estimated Initial Correct Enumeration Probabilities

	Site			
BFUGP**	Sacramento	Rural SC	Columbia	Menominee
1	0.940	0.879	0.943	0.475
2	0.874	0.849	0.864	0.774
3	0.741	0.732	0.800	0.384
4	0.848	0.714	0.971	0.897*
5	0.951	0.888	0.959	0.944*
6	0.000*	0.000*	0.000*	0.000*

* No unresolved persons in this category.

**BFUGP is the E-Sample Match Code Group. BFUGP 1-4 are sent to followup. BFUGP 1 are matches and possible matches. BFUGP 2 are partial household nonmatches. BFUGP 3 are whole household nonmatches where the address is matched. BFUGP 4 are whole household nonmatches where the address is not matched. BFUGP 5 are persons resolved before followup. BFUGP 6 are persons with insufficient information for matching (before followup status).

Table 11 gives a further breakdown of residence status for BFUGP 3 (whole household nonmatches where the housing unit matched in housing unit matching) for Sacramento. A conflicting household is where the housing unit matched and both the P-Sample and E-Sample collected persons but none of the persons in either the P-Sample or E-Sample households were matches or possible matches. The proportion correct among persons

from conflicting households in NRFU tend to be lower than for the other persons from BFUGP 3.

Table 11: Initial Correct Enumeration Status for Whole Household Nonmatches Where the HU Matched

	Sacramento			
	Corr	Erron	Unres	% Corr of Resolvd
HU Matched	1089	156	559	87.47
Conflict HH, Not NRFU	137	53	54	72.11
Conflict HH, NRFU	117	248	85	32.05

The "HU Matched" row excludes persons from conflicting households.

G. Conclusions

The Dress Rehearsal ICM Missing Data seems to be generally satisfactory. We probably want to put E-Sample persons from conflicting households in NRFU in their own match code group as these persons seem to have a lower probability of being correct. We may also want to split the remaining persons from BFUGP 3 (whole household nonmatches where HU matched in HU matching) into two match code groups: non-NRFU conflicting households and the remainder. On general principle, we may also want to put matches needing followup and possible matches needing followup into separate match code groups.

For the P-Sample, we may want to calculate P-Sample residence probabilities separately by match code group. We may also want to put P-Sample persons from conflicting households in their own match code group and put matches and possible matches needing followup into separate match code groups. In addition, we may want to calculate match probabilities for insufficient information people separately for movers and nonmovers.

H. References

[1] Bureau of the Census internal memorandum from G. Diffendal and T. Belin, "Results of Procedures for Handling Noninterviews, Missing Characteristic Data, and Unresolved Enumeration Status in 1990 Census/Post-Enumeration Survey," July 1, 1991.

[2] Bureau of the Census internal memorandum from D. Childers to M. Ramos, "DSSD Census 2000 Dress Rehearsal Memorandum Series F-DT-2, The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement," November 10, 1998.

[3] M. Ikeda, A. Kearney, and R. Petroni (1998), "Missing Data Procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," presented at the 1998 meetings of the American Statistical Association (ASA).

[4] M. Ikeda, A. Kearney, and R. Petroni (1998), "Handling of Missing Data in the 1996 Integrated Coverage Measurement," presented at the 1998 ASA Meetings.

[5] M. Ikeda and R. Petroni (1996), "Handling of Missing Data in the 1995 Integrated Coverage Measurement Sample," presented at the 1996 ASA Meetings (a shorter version of this paper appeared in the 1996 Proceedings of the Section on Survey Research Methods, American Statistical Association, 563-568).

[6] Bureau of the Census internal memorandum from D. Kostanich to M. Lynch "DSSD Census 2000 Dress Rehearsal Memorandum Series A-5, Computer Specifications for the Selection of the ICM Sample for the Census 2000 Dress Rehearsal (R. Sands, D. McGrath, and R. Zuwallack, authors) " November 15, 1997.

[7] Bureau of the Census internal memorandum from D. Kostanich to D. Stoudt, "DSSD Census 2000 Dress Rehearsal Memorandum Series A-38, Computer Specifications for ICM Site Level Estimation and Raking for the Census 2000 Dress Rehearsal (E. Schindler, author)," November 17, 1998.

[8] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Different Methods for Calculating Match and Residence Probabilities for the 1995 P-Sample Data, DSSD DSSD 2000 Census Dress Rehearsal Memorandum Series A-23 (M. Ikeda, author)," January 5, 1998.

[9] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Different Methods for Calculating Correct Enumeration Probabilities for the 1995 E-Sample Data, DSSD Census 2000 Dress Rehearsal Memorandum Series A-28 (M. Ikeda, author)," January 5, 1998.

[10] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Using Simple Ratio Methods to Calculate P-Sample Residence Probabilities and E-Sample Correct Enumeration Probabilities for the 1995 Data, DSSD Census 2000 Dress Rehearsal Memorandum Series A-30 (M. Ikeda, author)," January 28, 1998.

* This paper reports the general results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.