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

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

Key Words: Noninterview Adjustment, Imputation, Modeling

I. Introduction

This paper outlines the Integrated Coverage Measurement (ICM) missing data procedures that will be used for the Census 2000 Dress Rehearsal. A noninterview adjustment procedure, outlined in Section III, is used to account for whole-household nonresponse. A characteristic imputation procedure, outlined in Section IV, 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 V. The procedures are generally similar in effect to those used for ICM in the 1996 Community Census and the 1990 Post-Enumeration Survey (PES). Changes from the 1990 PES and/or 1996 Community Census procedures are summarized in this document. Information on the prevalence of missing data in the 1990 PES, 1995 ICM, and 1996 ICM is presented and results of research related to ICM missing data procedures is summarized. Methodologies and analysis of procedures are documented in [2], [4], and [9], respectively for the 1990 PES, 1995 ICM, and 1996 ICM.

II. General Background

The Census 2000 Dress Rehearsal is being conducted in three areas: Sacramento, CA; Menominee, WI; and Columbia, SC. The South Carolina site was divided into two subsites for the purposes of ICM sample selection and ICM missing data processing. The ICM sample was selected separately for each site and the two subsites. An overview of the ICM sample design for the Dress Rehearsal can be found in [5].

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 is 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 inmovers in the P-Sample poststratum estimates and uses outmovers to obtain poststratum estimates of match probability for inmovers. Further details on DSE estimation for the Dress Rehearsal can be found in [16].

III. Noninterview Adjustment

Noninterview adjustment is only performed on the P-Sample. Unlike 1996, there will be two noninterview adjustments. Two noninterview adjustments are needed because of the use of PES C estimation in the Dress Rehearsal. The two noninterview adjustments are basically identical to each other, except for the reference date. One noninterview adjustment will be based on housing unit status as of Census Day. The other noninterview adjustment will be based on housing unit status as of the day of ICM interview. The procedures are similar to the 1996 noninterview adjustment procedure. Each noninterview adjustment spreads the weights of noninterviewed units over interviewed units in the same block cluster and similar type of basic address (called type of place in 1996). 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. The definitions of interviews and noninterviews are similar to those used for 1996. Nonmovers and outmovers are used to determine Census Day housing unit status. Nonmovers and 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 (as of the given reference date) housing unit that is not an interview is a noninterview.

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

Changes from the 1990 PES: The only major change from the noninterview methodology for the 1990 PES is the use of two noninterview adjustments. Other minor changes were made due to minor changes in the data collected during listing.

Summary of Research Results: One alternative that has been suggested is the use of Census demographic data to help define noninterview adjustment cells. We decided against doing this because research using the 1995 [12] and 1996 [14] ICM data suggested it would not have much effect on the estimates, even when the noninterview rate was high. Results from 1990 data [15] do suggest that completely dropping the noninterview adjustment would have important effects on the estimates. The noninterview rates for 1990, 1995, and 1996 are given in Table 1 at the end of the document.

IV. Characteristic Imputation

P-Sample characteristic imputation for the Dress Rehearsal will be similar to characteristic imputation for the 1996 ICM and the 1990 PES. In a change from the 1996 methodology, we will use the demographic information from the Dress Rehearsal Census edited file (CEF) for the Dress Rehearsal E-Sample. This means that the only ICM imputation that needs to be done in the E-Sample is for E-Sample persons that could not be matched to the CEF. The methodology for any remaining E-Sample ICM imputation is 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). The use of Hispanic origin to help impute race (and vice versa) is a change from the 1996 methodology. 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 will be 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.

Changes from the 1990 PES

1) The most important change is the use of CEF demographic data for E-Sample persons. We do this because the E-Sample is a sample from the Census. It therefore makes sense (and should reduce random noise) to use the final Census demographic data for E-Sample persons.

2) In the Dress Rehearsal, we will use Hispanic origin to help impute race (in cases where a whole household is missing race) and vice versa. This was not done in 1990. We decided to make this change because our experience with ICM data suggests Hispanic origin should be helpful in predicting race (and vice versa). In addition, the Census edit/imputation system will use Hispanic origin to help impute race (and vice versa) in the Dress Rehearsal.

3) In 1990, relationship to reference person was also imputed. We did not impute this variable in 1995 and 1996, and are not planning to impute this variable in the Dress Rehearsal. It is not necessary to impute this variable in order to calculate poststratum estimates. Type of structure was also imputed in 1990. The

corresponding variables will not have missing values in the Dress Rehearsal.

4) There were minor changes in the imputation procedure for sex (e.g. reference persons from multi-person households with no spouse present are now imputed using their own distribution).

5) Marital status was used in 1990 in the imputation process (and was also one of the imputed variables). Marital status is not available in the Dress Rehearsal.

Summary of Research Results: A different methodology was used for characteristic imputation in the 1995 ICM. A simulation study on 1990 data [3] supported the use of methodology similar to the 1990 methodology. Comparison of estimates obtained using different P-Sample imputation methods on the 1995 data [6], [8] suggest that the exact choice of the P-Sample imputation method is not particularly important. Results from 1990 [15] do suggest that simply removing all persons with missing data would have important effects on the estimates. The study, however, removed persons with missing data from both the P-Sample and the E-Sample. Comparison of estimates obtained using different E-Sample imputation methods on the 1995 data suggest that the E-Sample imputation adds a substantial amount of seemingly random noise to some estimates [6], [8]. This supports the use of CEF demographic data for the E-Sample. Item imputation rates for the 1990, 1995, and 1996 P-Samples are given in Table 2. The corresponding E-Sample imputation rates are given in Table 3.

V. 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.

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 nonmovers and outmovers with resolved final residence status. The Census Day residence probability for inmovers is irrelevant to estimation and will be set to 0. Note that the residence probability as of the date of ICM interview for inmovers and nonmovers is assumed to be 1 (except that infants born after Census Day are not considered to be ICM interview day residents).

Some nonmovers and outmovers will have unresolved match status. The match probability for these persons is the proportion of matches among nonmovers and 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 inmovers is irrelevant to estimation and will be 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 NRFU and 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. If an E-Sample person was verified to belong in a surrounding block and was also duplicated with a surrounding block person in the NRFU or UAA sample, then the E-Sample correct enumeration "probability" was set to one minus the NRFU or UAA weight of the surrounding block person.

Changes from 1990 PES

1) In 1990, match and enumeration probability were modeled using hierarchical logistic regression. Residence probability was not specifically modeled in

the P-Sample, although there was an adjustment for fictitious P-Sample persons. More information on the 1990 PES model can be found in [1] and [2]. In 1996, we modeled residence probability in the P-Sample and used a simple ratio estimate for estimating match probability. In the Dress Rehearsal we are using simple ratio estimates for match, residence and enumeration probabilities. The reasons for the changes between 1990 and 1996 are the following (see [4] and [9]):

First, almost everyone sent to followup in 1995 (and 1996) had resolved final match status (2 persons in 1995 and 3 persons in 1996 who were sent to followup had unresolved final match status).

Second, followup never changed a match to nonmatch, and almost never changed a nonmatch to a match (8 persons in 1995 and 11 persons in 1996 were changed from nonmatch to match by followup).

Third, even with sampling for followup, a substantial majority of the persons with unresolved final match status were persons with insufficient information for matching. The way that the 1990 and 1995 logistic regression model calculated match probability for these persons basically amounts to a simple ratio (plus random noise).

We expect the Dress Rehearsal and 2000 Census will provide similar results because of changes in procedures since 1990.

2) In a change from 1996, the logistic regression models for residence and enumeration probability are being replaced by simple ratios. This is due to research results [7], [11], [13] that suggested that the use of logistic regression models has little effect on the estimates.

3) We are using PES C in the Dress Rehearsal. This means that we match P-Sample nonmovers and outmovers. The match status of inmovers is irrelevant to estimation in the Dress Rehearsal, and is not obtained. In 1990, we matched nonmovers and inmovers and did not collect outmovers. Note that in 1995 and 1996, we matched nonmovers and outmovers and did not collect inmovers.

Summary of Research Results: The results generally suggest that the exact form of the logistic regression model is not very important. Results [10] from the 1995 data indicated that the age/race-ethnic/sex interactions have almost no effect on either the poststratum estimates or the individual fitted probabilities. As a result, the interaction parameters were dropped from the 1996 models. More recent results from the 1995 data [7], [11], [13] suggest that even completely dropping the logistic regression models and replacing them with simple ratios will not have major effects on the poststratum estimates. The results actually tend to overstate any effects, since they are based on 1995 data. In 1995, roughly half of the persons needing followup were sampled out of followup. We are not planning to sample for followup in either the Dress Rehearsal or the 2000 Census.

Prevalence of Unresolved Status

In both 1995 and 1996 (see [4] and [9]):

1) Roughly 20% of the P-Sample and E-Sample needed followup. Roughly half of these persons were sampled out of followup.

2) Roughly 10% of P-Sample persons sent to followup had unresolved final residence status. Almost none of the P-Sample persons sent to followup had unresolved final match status.

3) Roughly 20% of E-Sample persons sent to followup had unresolved final correct enumeration status.

1990 P-Sample results are not comparable because of differing procedures. In the 1990 E-Sample [2], about 15% of E-Sample persons needed followup, and 7.4% of the E-Sample persons sent to followup had unresolved final correct enumeration status.

Acknowledgements

The authors wish to thank Yves Thibaudeau and Lynn Weidman for helpful comments on earlier drafts of this manuscript.

References

[1] T. Belin, G. Diffendal, S. Mack, D. Rubin, J. Schafer, and A. Zaslavsky (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," Journal of the American Statistical Association, 88, 1149-1159.

[2] 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.

[3] S. Dorinski, R. Petroni, M. Ikeda, and R. Singh (1996), "Comparison and Evaluation of Alternative ICM Imputation Methods", 1996 Proceedings of the Section on Survey Research Methods, American Statistical Association, 299-304.

[4] 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).

[5] 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.

[6] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Comparison of Using 1996 ICM Characteristic Imputation Methodology and the 1996 Census Characteristic Imputation Methodology on the 1995 ICM P and E-Sample Data, DSSD Census 2000 Dress Rehearsal Memorandum Series A-21 (M. Ikeda, author)," December 11, 1997.

[7] 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.

[8] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Using the 1996 ICM Characteristic Imputation and Probability Modeling Methodology on the 1995 P and E-Sample Data, DSSD Census 2000 Dress Rehearsal Memorandum Series A-20 (M. Ikeda, author)," December 11, 1997.

[9] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Handling of Missing Data in the 1996 Integrated Coverage Measurement Sample, DSSD Census 2000 Dress Rehearsal Memorandum Series A-26 (M. Ikeda, author)," January 5, 1998.

[10] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Eliminating the Age/Sex/Race Interaction Parameters in the R-Sample Residence Status and the E-Sample Correct Enumeration Status Probability Models, DSSD Census 2000 Dress Rehearsal Memorandum Series A-27 (M. Ikeda, author)," January 5, 1998.

[11] 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.

[12] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Using Data from Matching HERF Households to Define Noninterview Adjustment Cells for the 1995 ICM, DSSD Census 2000 Dress Rehearsal Memorandum Series A-25 (M. Ikeda, author)," January 5, 1998.

[13] 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.

[14] Bureau of the Census internal memorandum from R. Petroni to E. Vacca, "Effect of Using Data from Matching CEF Households to Define Noninterview Adjustment Cells for the 1996 ICM, DSSD Census 2000 Dress Rehearsal Memorandum Series A-25 (A. Kearney, author)," December 11, 1997.

[15] R. Petroni, A. Kearney, M. Town, and R. Singh (1995), "Should We Account for Missing Data in Dual System Estimation?", Proceedings of the Sixth International Workshop on Household Survey Nonresponse, Statistics Finland, 166-176.

[16] Bureau of the Census internal memorandum from R. Singh to M. Lynch, "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)," March 5, 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.

Table 1: P-Sample Noninterview Rates (%)

1990	1995			1996	
U.S.	Oakland	Paterson	Chicago	Fort Hall	Acoma
1.51	15.06	8.49	9.28	3.04	2.11

The rates in the table are the (unweighted) percentage of occupied housing units that are noninterviews. Note that the 1995 noninterview adjusted rates are artificially high because of problems with the 1995 CAPI instrument.

Table 2: P-Sample Item Imputation Rate (%)

	1990	1995			1996	
	U.S.	Oakland	Paterson	Chicago	Ft Hall	Acoma
Tenure	2.39	2.82	1.42	0.14	0.21	0.87
Sex	0.53	0.15	0.13	0.14	0.08	0.04
Age	0.73	2.84	1.89	1.77	1.02	2.08
Hisp Orig	2.29	0.90	0.71	0.95	0.12	0.30
Race	0.94	1.05	0.93	3.80	0.41	0.15

1995 and 1996 rates are for residents and possible residents from interviewed households. Note that the 1990 P-Sample included GQ persons while the 1995 and 1996 P-Sample excluded them.

Table 3: E-Sample Item Imputation Rates (%)

	1990	1995			1996	
	U.S.	Oakland	Paterson	Chicago	Ft Hall	Acoma
Tenure	3.05	1.20	0.91	4.34	1.38	1.61
Sex	1.19	1.50	1.18	3.22	0.84	1.19
Age	2.58	8.25	8.54	5.77	1.25	3.73
Hisp Orig	10.55	6.10	6.15	19.14	46.27	60.21
Race	3.50	6.16	6.05	10.86	2.20	4.19

Note that Hispanic origin was not used in the poststratification for Fort Hall and Acoma. Race was not used in the poststratification for Acoma. We suspect the extremely high proportion of missing Hispanic origin in Fort Hall and Acoma was due to respondents not viewing the question as relevant.

The age imputation rates for 1995 have been recalculated to make them comparable with the 1996 age imputation rates. They therefore differ somewhat from the imputation rates given in [4].