MISSING DATA PROCEDURES FOR THE CENSUS 2000 ACCURACY AND COVERAGE EVALUATION SAMPLE

Michael Ikeda and Patrick J. Cantwell, Bureau of the Census*
Michael Ikeda, Statistical Research Division, Bureau of the Census, Washington, DC, 20233
michael.m.ikeda@census.gov

Key Words: Noninterview Adjustment, Characteristic Imputation, Imputation Cell Estimation

I. Introduction

This paper outlines the procedures for handling missing data in the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) sample. Section II gives some general background. A noninterview adjustment procedure, outlined in Section III, is used to account for whole-household nonresponse. Two separate adjustments are needed: one based on the interview status of the household as of Census Day, the other based on the interview status of the household as of the day of A.C.E. interview. A characteristic imputation procedure, outlined in Section IV, is used to assign values for specific missing demographic variables. Depending on the variable, we used hot-deck imputation, imputation from conditional distributions, or a combination of the two. Finally, persons with unresolved match, residence, or enumeration status have probabilities assigned based on the procedure outlined in Section V. Persons with unresolved status were assigned a probability based on the status of resolved persons in the same imputation cell. Full details of the missing data procedures can be found in [2]. An overview of results from A.C.E. missing data can be found in [7].

II. General Background

Census 2000 is conducted for the entire nation. There are two separate A.C.E. samples: one for the U.S. (50 states and the District of Columbia), and a second sample for Puerto Rico. Sampling units are block clusters (blocks or groups of blocks). The U.S. and Puerto Rico A.C.E. samples are processed separately by the A.C.E. missing data system. The A.C.E. missing data procedures are almost identical for the U.S. and Puerto Rico samples. For simplicity, this document will focus primarily on the U.S. sample, with the few differences noted.

The A.C.E. uses dual system estimation (DSE) to calculate estimates. With DSE we try to obtain a roster from the A.C.E. block clusters independently of the Census. The independent roster (P Sample) and the Census roster from A.C.E. block clusters (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.

One complicating factor is that some people move between Census Day and A.C.E. interview day. To handle movers, A.C.E. uses a DSE method called Mover Procedure C. Procedure C uses person inmovers to obtain estimates of movers (as it is easier to collect information on inmovers) and uses person outmovers to obtain estimates of mover match rates (as it is easier to match outmovers). In the 1990 Post-Enumeration Survery (PES), person inmovers were used to obtain the estimates of both movers and the mover match rate.

The missing data procedures are similar to those used for the Integrated Coverage Measurement (ICM) sample in the Census 2000 Dress Rehearsal. An outline of the missing data procedures for the Dress Rehearsal ICM and a summary of related research is given in [4]. An overview of the Dress Rehearsal missing data results is given in [5]. Additional background related to A.C.E. missing data can be found in [6]. An overview of A.C.E. operations can be found in [1]. For the 1990 PES, the 1990 operation analogous to A.C.E., an overview of the missing data procedure and results can be found in [3].

III. Noninterview Adjustment

Noninterview adjustment is only performed on the P Sample. The noninterview adjustment procedure is almost identical to the procedures used in the Census 2000 Dress Rehearsal and is similar to the procedure used in the 1990 PES.

Mover Procedure C requires estimates of both inmovers and outmovers. Therefore, two noninterview adjustments are needed: one based on housing-unit status as of Census Day, the other based on housing-unit status as of the day of A.C.E. interview. The two noninterview adjustments are identical to each other, except for the reference date for housing-unit status.

The noninterview adjustment based on Census Day status is used to adjust the weights of person nonmovers and person outmovers. The noninterview adjustment based on A.C.E. Interview Day status is used to adjust the weights of person inmovers.

Person nonmovers and person outmovers are used to determine Census Day housing-unit status. Person nonmovers and person inmovers are used to determine A.C.E. 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 two demographic characteristics) 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 (for a given reference date) spreads the weights of noninterviewed units equally over interviewed units in the same noninterview adjustment cell. Noninterview adjustment cells are defined as the block cluster crossed with the type of basic address category. The noninterview adjustment uses three categories for type of basic address: single-family unit, apartment, and other. Mobile homes are an example of "other" units.

If the number of noninterviewed units in a cell is more than twice the number of interviewed units, then the weights of the noninterviewed units are spread out over a broader category of interviewed units. If the number of noninterviewed units in the cell is more than twice the number of interviewed units in the current broader category, then we go to the next category of interviewed units.

Note that cells are not collapsed together: weights of noninterviewed units in a problem cell are spread over a broader category, but weights of noninterviewed units in non-problem cells are still spread only within their cell.

The categories above the adjustment cell (in the order they would be used) are as follows:

1) Adjustment Stratum x type of basic address category
2) Block Cluster
3) Adjustment Stratum
4) State (DC and Puerto Rico are considered states for the noninterview adjustment)

Adjustment Stratum is a classification of block clusters within each state for the purpose of the noninterview adjustment. Most clusters are grouped into Adjustment Strata based on the expected demographic/tenure makeup of the cluster. The small cluster sampling stratum and the American Indian Reservation sampling stratum are each separate Adjustment Strata. This is because sampling rates in these two sampling strata can be very different from the sampling rates for other clusters.

In the U.S., the weights of noninterviews in 65 cells needed to be spread over a broader category in the noninterview adjustment for Census Day, compared with 13 cells in the adjustment for A.C.E. interview day (in both cases, out of roughly 15,000 cells that contained housing units). For 64 of these Census Day cells and all 13 of these A.C.E. interview day cells, the Adjustment Stratum x type of basic address category had enough interviewed units. One Census Day cell used the Block Cluster category. We did not need to use the Adjustment Stratum category or the State category.

IV. Characteristic Imputation

P-Sample characteristic imputation for the Census 2000 A.C.E. is basically identical to characteristic imputation for the Dress Rehearsal ICM and is similar to the method used in the 1990 PES. For the Census 2000 E Sample we use the demographic information from the Census 2000 Hundred-Percent Census Edited File (HCEF). Because all E-Sample persons matched to the HCEF, no A.C.E. imputation needed to be done in the E Sample. In the 1990 PES there was a separate E-Sample imputation. We decided to use HCEF data in the 2000 E Sample since the E Sample is a representative sample from the Census.

The variables imputed in the A.C.E. are tenure, race, Hispanic origin, age, and sex. P-Sample person mover status is not considered when imputing characteristics. However, persons from the P-Sample whole-household outmover interview path are considered to be a separate household for imputation purposes. Age and sex distributions are calculated separately for the U.S. and Puerto Rico. Imputation for a specific missing characteristic is not affected by the imputation for other missing characteristics. The sort for imputation, important for the hot-deck procedures described below, is essentially a geographic sort, except that block clusters with similar demographic characteristics will tend to be grouped together within a state.

Tenure: Tenure (collapsed to owner vs. non-owner) is imputed from the nearest previous household with recorded tenure and the same type of basic address category. The type of basic address categories are single-family unit, apartment, and other--the same ones used for the noninterview adjustment.

Race: Race is imputed using the race distribution (that is, selecting a person at random) in the same household or a previous household. The distribution used depends on whether the whole household is missing race and, if so, whether the whole household is also missing Hispanic origin.

1) If there is at least one person in the household with race reported, then race is imputed using the distribution of race in the household.

2) If everyone in the household has a missing value of race but at least one person has a reported value of Hispanic origin, then race is imputed using the distribution of the nearest previous household with reported race and same Hispanic origin. The Hispanic origin categories used for this procedure are Non-Hispanic vs. Hispanic.

3) If everyone in the household has a missing value of both race and Hispanic origin, then race is imputed using

the distribution of the nearest previous household with reported race.

All 63 different combinations of the 6 basic race categories are imputed (the 6 categories being: White, Black, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Other). All 63 categories are treated the same in the imputation; there are no special procedures for categories or groups of categories.

Hispanic Origin: Hispanic origin is collapsed to Non-Hispanic and Hispanic for imputation purposes. Hispanic origin is imputed using the Hispanic origin distribution in the same household or a previous household. The distribution used depends on whether the whole household is missing Hispanic origin and, if so, whether the whole household is also missing race.

1) If there is at least one person in the household with reported Hispanic origin, then Hispanic origin is imputed using the distribution of Hispanic Origin in the household.

2) If everyone in the household has a missing value of Hispanic origin but at least one person has a reported value of race, then Hispanic origin is imputed using the distribution of the nearest previous household with reported Hispanic origin and similar race. The race categories used here are: white; other, or white and other; all remaining nonmissing categories.

3) If everyone in the household has a missing value of both Hispanic origin and race, then Hispanic origin is imputed using the distribution of the nearest previous household with reported Hispanic origin.

Age: We impute the age category, not the raw age. The age categories imputed are the same ones used in the poststratification: 0-17, 18-29, 30-49, 50+. Age category is imputed using age category distributions of persons with reported age. The distribution used depends on household size (single vs. multi-person), relationship category, and, for selected relationship categories, age category of reference person (if the household contains a reference person).

1) Age category in one-person households is imputed from the distribution of age in one-person households.

2) For several relationship categories in multi-person households we are willing to assume a strong correlation between the age of the given person and the age of the reference person. Therefore for spouse, child, sibling, and parent of the reference person, we impute age category from the distribution of age category for persons (in multi-person households) with the same relationship category and the same age category of the reference person. Imputed age category of reference person is not used to calculate age distributions, but it is used to determine which age distribution is used in this imputation.

3) For the remaining reported relationship categories in multi-person households (reference person, other relative, nonrelative) we impute the age category using the distribution of age category for persons in the same relationship category in multi-person households.

4) If the person has reported (non-missing) relationship but the household lacks a reference person, we impute using the distribution of age category for persons in the same relationship category.

5) For persons with missing relationship in multi-person households, we impute from the distribution of age category in multi-person households, excluding reference persons.

Sex: The sex imputation is the most complicated. Reference person (spouse present) and spouse are imputed from each other. Other persons are imputed from sex distributions of persons with reported sex. The distribution used depends on household size (single vs. multi-person), relationship category, and, for some relationship categories, whether a spouse is present in the household.

Sex distributions for multi-person households are calculated after the imputation of sex of spouse from reference person and sex of reference person (spouse present) from spouse (assuming at least one of the two has reported sex).

1) For one-person households, sex is imputed from the distribution of sex in one-person households.

2) If one of the reference person or spouse is missing sex, then 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.

3) 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.

4) For the reference person (with no spouse present) of a multi-person household, the distribution of sex for reference persons (with no spouse present) of multi-person households is used.

5) For persons with reported relationship (except reference persons and spouses) from multi-person

households, sex is imputed from the distribution of sex for persons with reported (non-missing) relationship from multi-person households (excluding reference persons and spouses). This distribution is also used to impute sex for persons with missing relationship in a household where a spouse is present.

6) For persons with missing relationship from multi-person households where no spouse is present, sex is imputed from the distribution of sex for persons (excluding reference persons) from multi-person households.

V. Assigning Residence, Match, and Correct Enumeration Probabilities

Probabilities for persons with unresolved final Census Day residence (P Sample), final match (P Sample), or final enumeration (E Sample) status are estimated by the imputation cell estimation (ICE) method. With ICE we calculate weighted ratios based on persons with resolved final status. The weights used in ICE include all applicable stages of sampling, but do not include the P-Sample noninterview adjustment. The Dress Rehearsal also used ICE to estimate all three probabilities. However, we define cells at a more detailed level in the Census 2000 A.C.E. The ICE cells for Puerto Rico are collapsed versions of the cells for the U.S.

In the 1990 PES, logistic regression was used to assign match and correct enumeration probabilities. Residence probability was not specifically modeled in the 1990 PES because of the different treatment of movers. We decided to use ICE in A.C.E. because of concerns about the feasibility of implementing logistic regression and because our research suggested minimal difference between ICE and logistic regression [4], [6].

Roughly 20% of the A.C.E. block clusters were targeted extended search (TES) clusters. In TES clusters, the search area was extended to blocks that touch the cluster. Persons in all clusters were classified as "TES persons" or "non-TES persons". In general, non-TES persons are people who would not be eligible for TES even if they were in a TES cluster. For TES persons, the weight used in ICE incorporates TES sampling as well as all earlier stages of sampling. The weight used in ICE for non-TES persons does not incorporate TES sampling. TES person status is also used to determine the E-Sample imputation cells.

There was one important late change made to the cells for P-Sample residence probability and E-Sample correct enumeration probability. Certain persons are sent to person followup to obtain further information on residence, match, or enumeration status. The information from person followup is used by clerical matchers to assign final status. In a separate keying operation, the answers to the person followup questions were keyed into various letter codes. The information from this keying operation was originally intended for evaluation purposes and was not initially expected to be available in time for production.

Due to accelerated work by those involved in the operation, the information from the keying operation was available just in time to be used in production. The information was used to classify some persons as "potentially fictitious" and others as "said to be living elsewhere". Persons who were classified as "potentially fictitious" or "said to be living elsewhere" were placed into newly defined imputation cells for residence and correct enumeration probability. Unresolved persons in the new cells had low estimated probabilities, while the estimated probabilities for unresolved persons were raised in the other cells that contained persons needing followup.

Residence Probability. For P-Sample persons with unresolved residence status, the residence probability is the weighted proportion of residents (among persons with resolved residence status) in the given imputation cell. The imputation cells for the U.S. for estimation of P-Sample residence probability are defined in Table 1. Each internal table cell is an imputation cell. Match code group uses the before-followup match status, sufficient information status, and, for some cases, information from the followup interview.

Table 1. Imputation Cells for Residence Probability

Match Code Group	Owner				Renter			
	NH White		Others		NH White		Others	
1=Matches needing followup								
2=Possible matches								
3=Partial household nonmatches needing followup	3a	3b	3a	3b	3a	3b	3a	3b
4=Whole household nonmatches needing followup, not conflicting households								
5=Nonmatches from conflicting households								
6=Resolved before followup								
7=Insufficient information for matching	Weighted column average of groups 1-5,8		Weighted column average of groups 1-5,8		Weighted column average of groups 1-5,8		Weighted column average of groups 1-5,8	
8=Potentially fictitious or said to be living elsewhere								

Group 3 was split into two subgroups. Group 3a includes persons aged 18-29 in match code group 3 who are children of the reference person. Group 3b includes all other persons in match code group 3. NH White means Non-Hispanic White. The imputation cells for residence probability in Puerto Rico are the U.S. cells collapsed over NH White/Others and over 3a and 3b.

The Census Day residence probability for person inmovers is irrelevant to estimation and is set to 0.

Match Probability. For unresolved match status, the match probability for persons with unresolved match status is the weighted proportion of matches in the same imputation cell among persons with resolved final match status (excluding confirmed Census Day nonresidents). Most persons with unresolved match status (over 98 percent) are persons with insufficient information for matching. Persons with insufficient information lack a valid name or lack most person characteristics. All persons with unresolved match status also have unresolved residence status.

The imputation cells for the U.S. estimation of P-Sample match probability are defined in Table 2. Each internal table cell is an imputation cell.

Table 2. Imputation Cells for Match Probability

Mover Status	Address Code			
	Housing Unit Match		Housing Unit Nonmatch or Conflicting Households	
Nonmover	0 imputes	1+ imputes	0 imputes	1+ imputes
Outmover	0 imputes	1+ imputes		

1+ imputes means that at least one of the five imputed variables (tenure, race, Hispanic origin, age, or sex) had missing data in the A.C.E. The imputation cells for match probability for Puerto Rico are the U.S. cells collapsed over 0/1+ imputes.

The match probability is set to 0 for confirmed Census Day nonresidents. The match probability for person inmovers is irrelevant to estimation and is set to 0.

Correct Enumeration Probability. For E-Sample persons with unresolved enumeration status, the correct enumeration probability is the weighted proportion of correct enumerations (among persons with resolved enumeration status) in the given imputation cell.

The imputation cells for the estimation of E-Sample correct enumeration probability in the U.S. are defined in Table 3. Each internal table cell is an imputation cell. Note that the few unresolved persons in group 9 (insufficient information for matching) are assigned a correct enumeration probability of 0. Match code group uses the before-followup match status, sufficient information status, TES person status, and, in some cases, information from the followup interview.

Table 3. Imputation Cells for Correct Enumeration Probability

Match Code Group	0 Imputes		1+ Imputes	
1=Matches needing followup				
2=Possible matches				
3=Partial household nonmatches	3a	3b	3a	3b
4=Whole HH nonmatches where HU matched, not conflicting households	NH White	Others		
5=Nonmatches from conflicting HH, HU not in regular nonresponse follow-up (NRFU) in Census				
6=Nonmatches from conflicting HH, HU in regular NRFU in Census				
7=Whole HH nonmatches, housing unit did not match in HU matching	NH White	Others		
8=Resolved before followup	NH White	Others		
9=Insufficient information for matching				
10=TES persons				
11=Potentially fictitious				
12=Said to be living elsewhere				

1+ imputes means that at least one of the five imputed variables had imputed data in the Census. Group 3 is split into two subgroups 3a and 3b, defined for the E Sample as they were for the P Sample (see Residence Probability). NH White means Non-Hispanic White. The imputation cells for correct enumeration probability for Puerto Rico are simply the match code groups.

References

[1] Bureau of the Census Internal Memorandum from D. Childers to M. Ramos (2001), "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, Revised, Janurary 26, 2001.

[2] Bureau of the Census Internal Memorandum from P. Cantwell to D. Kostanich (2001), "Accuracy and Coverage Evaluation Survey: Specifications for the Missing Data Procedures; Revision of Q-25," DSSD Census 2000 Procedures and Operations Memorandum Q-62, July 9, 2001.

[3] Bureau of the Census Internal Memorandum from G. Diffendal and T. Belin (1991), "Results of Procedures for Handling Noninterviews, Missing Characteristic Data, and Unresolved Enumeration Status in 1990 Census/Post-Enumeration Survey," STSD Decennial Census Memorandum Series #V-112, July 1, 1991.

[4] M. Ikeda, A. Kearney, and R. Petroni (1998), "Missing Data Procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," *1998 Proceedings of the Section on Survey Research Methods, American Statistical Association*, 617-622.

[5] A. Kearney and M. Ikeda (1999), "Handling of Missing Data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," *1999 Proceedings of the Section on Survey Research Methods, American Statistical Assocation*, 468-473.

[6] Bureau of the Census Internal Memorandum from D. Kostanich to H. Hogan (1999), "Accuracy and Coverage Evaluation Survey: Overview of Missing Data for P & E Samples," DSSD Census 2000 Procedures and Operations Memorandum Series Q-3, September 23, 1999.

[7] M. Zelenak, D. McGrath, N. Nguyen, and P. Cantwell (2001), "Missing Data Results for the Census 2000 Accuracy and Coverage Evaluation Sample," *2001 Proceedings of the Section on Survey Research Methods, American Statistical Association*, in press.

* This paper reports the results of research and analysis undertaken by Census Bureau Staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.