HANDLING OF MISSING DATA IN THE 1995 INTEGRATED COVERAGE MEASUREMENT SAMPLE

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

Key Words: Noninterview Adjustment, Imputation, Modeling

A. INTRODUCTION

This paper gives an overview of the methods used to handle missing data in the 1995 Integrated Coverage Measurement (ICM) sample. It also provides an evaluation of the likely importance of any effect of the ICM missing data methods on the final results.

Data needed for ICM estimation is missing in some cases. First, we are unable to obtain adequate interviews from some households. A noninterview adjustment procedure, that is outlined in Section C-1, was used to account for whole household noninterviews. Second, there may be missing characteristics for some persons in interviewed households. The missing characteristics were filled in using a hot-deck imputation procedure outlined in Section C-2. Third, some persons will have an unresolved final residence, match, or enumeration status. Probabilities (for the final status) are calculated for these persons based on a logistic regression procedure outlined in Section C-3. The procedures in Sections C-1 and C-3 are similar to those used for the 1990 Post Enumeration Survey (PES). Documentation of the 1990 PES procedures can be found in [1].

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

B. GENERAL BACKGROUND

The 1995 Census Test was conducted in three sites: Oakland, CA; Paterson, NJ; and Northwest Louisiana. The ICM sample was selected separately for each site. The sampling units were block clusters (single blocks or groups of blocks, generally with 30 or more housing units). The block clusters were stratified by size and race/ethnic composition in the 1990 Census. Small (<3 housing units) block clusters were not stratified by race/ethnic composition. An overview of the ICM sample design (and of the evaluation results) can be found in [4].

There are three separate rosters involved in the ICM missing data processing: the R-Sample, the P-Sample, and the E-Sample. The R-Sample was created for all three sites and was used in Census Plus estimation. Census Plus estimates are calculated based on the assumption that the R-Sample is the "truth" for the ICM blocks. The P and E-Sample were only created for the Oakland and Paterson sites and were used in Dual System Estimation (DSE). DSE estimates are calculated based on the assumption that the P-Sample is collected independently of the E-Sample. Further details on the DSE and Census Plus estimation methods can be found in [8]. Details on the actual DSE and Census Plus estimates for the 1995 ICM can be found in [7].

In 1995, the information for both DSE and Census Plus was collected in a single interview. An independent roster was collected and then matched during the interview to a preliminary Census roster. Census Plus combined the preliminary Census roster and the independent roster into a final household roster. DSE used the independent roster to form the P-Sample and used the final Census roster to form the E-Sample. An overview of the 1995 ICM operations is given in [9].

R-Sample: The R-Sample contains all persons who should have been counted as residents in the Census in the ICM block clusters. The ICM produces a list, called the Enhanced Listing, of housing units that are confirmed to exist in the ICM block clusters on Census day. The R-Sample includes all persons who were residents on Census day of either housing units in the Enhanced listing or housing units added during ICM person interviewing. Housing units that are either in the Enhanced listing or were added during ICM interviewing are also referred to as R-Sample housing units.

P-Sample: The P-Sample is created from the Pristine (before matching to Census) Independent roster of persons. It is used to estimate persons missed in the Census. The P-Sample consists of those persons in the Pristine Independent roster who were residents of I-Sample housing units on Census day. I-Sample units are those housing units from an independent listing that are confirmed to exist on Census day. Housing units in the I-Sample are also referred to as P-Sample housing units.

E-Sample: The E-Sample consists of persons enumerated in the Census in the ICM block clusters. The E-Sample is an extract (before Census edit and imputation) from the Census file. It is used to estimate persons erroneously enumerated in the Census.

C. OUTLINE OF PROCEDURES

1. **Noninterview Adjustment**: Whole-household noninterviews are accounted for using a noninterview adjustment. The noninterview adjustment procedures are almost identical in the R-Sample and P-Sample. Noninterview adjustment is not applied to the E-Sample (although the proportion of Census whole-person imputations is incorporated into the DSE adjustment factor).

The main noninterview adjustment is at the block cluster x (recoded) type of place level. The type of place categories are collapsed into five categories for the adjustment: single-family attached, single-family detached, apartments, other, missing. Type of place is never missing in the P-Sample. The weight for noninterviewed housing units with nonmissing type of place in a given block cluster x recoded type of place category is spread among the interviewed housing units in the same block cluster x recoded type of place category. Special procedures are used for noninterviews with missing type of place and for noninterviews in the small block cluster stratum.

If the number of noninterviewed units in the given block cluster x recoded type of place category is more than twice the number of interviewed units, then the weight of the noninterviewed units is instead spread among the interviewed housing units in the same ICM sample selection stratum x type of place category.

If the number of noninterviewed units in the given block cluster x recoded type of place category is more than twice the number of interviewed units in the ICM sample selection stratum x recoded type of place category, then the weight of the noninterviewed units is instead spread among the interviewed housing units in the same block cluster.

If the number of noninterviewed units in the given block cluster x recoded type of place category is more than twice the number of interviewed units in the block cluster, then the weight of the noninterviewed units is instead spread among the interviewed housing units in the same ICM sample selection stratum.

Missing Type of Place: Noninterviewed housing units with missing type of place are treated specially in the R-Sample. A portion of their weight is spread over all interviewed housing units in the block cluster. The rest of their weight is spread over the vacant/delete units in the block cluster. The portion assigned to vacant/delete units is the estimated proportion of vacant/deletes in the block cluster. The collapsing criteria and collapsing sequence for units with missing type of place are the same as for other units except that the sequence starts at the block cluster level.

Small Block Cluster Stratum: For noninterviewed units with nonmissing type of place in the small block cluster stratum, the weight of the noninterviewed units in a given block cluster x recoded type of place category is spread among interviewed units in the same recoded type of place category in the small block cluster stratum (in the same site). The collapsing criteria and collapsing sequence for units in the small block cluster stratum are the same as for other units except that the sequence starts at the stratum x recoded type of place level and ends at the site level. Units with missing type of place are treated the same in the small block cluster stratum as they are in other strata except that the collapsing sequence ends at the site level.

2. **Characteristic Imputation**: Some persons in interviewed households will have missing characteristics. Missing characteristics were filled in using a hot-deck imputation procedure. Characteristic imputation is performed on all three samples. Similar procedures are used for the R, P, and E-Samples. The variables imputed are tenure, sex, age, race, and hispanic origin. These are the variables needed to create population estimates. The race imputation is performed on the five main race categories. The main portion of the imputation is performed using the Flexible matching imputation procedure developed by Todd Williams, Lynn Weidman and Kimberly Long. Details on this procedure are in [10] and [11].

The first stage of the characteristic imputation procedure imputes for tenure and for the sex of married householders and spouses of householders. Tenure is imputed from the nearest previous unit with a nonmissing value of tenure. Married householders and spouses of householders with a missing value of sex are assigned the sex opposite of their spouses. The E-Sample also imputes age based on computed age (age computed from year of birth).

The next step is the modeling portion of the flexible matching program. This portion finds the variables that are matched on in the matching phase of the program. The matching variables used to impute age are found by using stepwise linear regression. The matching variables used to impute sex, race, and hispanic origin are found by using stepwise logistic regression. Modeling is done separately for each site.

The final step is the matching portion of the flexible matching program. This portion imputes missing values by finding persons who match on the matching variables that were found in the modeling portion of the program. Matching is done separately for each site.

3. **Modeling of Probabilities**: Some persons will have an unresolved final residence, match, or enumeration status. The modeling of probabilities (for the final status) for these persons is done using a hierarchical logistic regression program for the R, P, and E-Samples. The programs are modified versions of the program used to model match probabilities for the 1990 PES. All sites are modelled together for each sample.

Probabilities for persons with unresolved final status are calculated using a model fitted on persons with resolved final status. The model contains both general parameters (fitted using all persons) and group parameters (fitted using persons in the given group). Persons are assigned to groups based on their initial status in combination with other variables. The model parameters (both group and overall) were generally similar to the parameters used in the 1990 PES. Residence status probability is modeled for the R-Sample, match probability is modeled for the P-Sample, correct enumeration probability is modeled for the E-Sample. A complication for the P and E Samples is that roughly half of the persons needing followup were sampled out of DSE followup. Persons needing followup but sampled out are considered to be unresolved.

There were also some P-Sample persons with unresolved residence status after DSE followup. The probability of residence for P-Sample persons with unresolved residence status was calculated to be the proportion of residents among those persons with resolved residence status who were sent to followup. The calculation was done separately for Oakland and Paterson. Persons needing followup but sampled out are considered to have unresolved residence status.

R-Sample Residence Status Groups: The R-Sample residence status groups are based on the Computer residence status code (residence status code assigned in the ICM interview) and the person outmover status. (Person outmover status was mainly determined from responses to coverage probes. Evaluation of these responses in [2] suggests that the coverage probes had serious shortcomings. This could have prevented us from identifying a substantial number of outmovers.) Persons with resolved Computer residence status are in residence status group 3. Persons with unresolved Computer residence status are in group 2 if they are person outmovers and in group 1, otherwise.

P-Sample Match Code Groups: The P-Sample match code groups are based on the before-followup (BFU) match codes, BFU whole/partial household match code, address code from housing unit matching, followup flag, and DSE followup sampling flag. Persons with insufficient information for matching are in match code group 8. Other persons not needing followup are in match code group 4. Persons sent to followup are in match code groups 1-3 (persons from whole-household nonmatches where address is not matched are in group 3, persons from other whole-household nonmatches are in group 2, other persons sent to followup are in group 1). Persons needing followup but sampled out of followup are in groups 5-7. The definitions for groups 5-7 otherwise correspond to, respectively, the definitions for groups 1-3.

Persons in groups 5-7 are not used to fit the model. Their estimated match probabilities are calculated as if they were in, respectively, groups 1-3. The estimated match probabilities for persons in group 8 are calculated by taking a weighted average of the probabilities that would have been assigned for groups 1-4. Weighting is by the frequency of groups 1-4, with groups 1-3 double weighted (since the sample rate is 1/2).

E-Sample Match Code Groups: The E-Sample match code groups are based on the BFU match codes, initial BFU followup) whole/partial household match code, address code from HU matching, followup flag, and DSE followup sampling flag. The definitions of the E-Sample groups are the same as the P-Sample groups. Persons in groups 5-7 are not used to fit the model. Their estimated correct enumeration probabilities are calculated as if they were in, respectively, groups 1-3. Persons in group 8 are given a probability of correct enumeration equal to 0 (except for four cases where such persons had a final match code different from their BFU match code).

D. MISSING DATA RESULTS

1. **Noninterview Adjustment**: The noninterview rates were highest in Oakland (8.54% R-Sample, 15.06% P-Sample), lowest in NW LA (0.94% R-Sample), with Paterson in the middle (2.18% R-Sample, 8.49%

P-Sample). Note that the main reason for the difference between the R-Sample and P-Sample noninterview rates is that in a number of housing units (6.7% of occupied P-Sample housing units in Oakland and 5.01% of occupied P-Sample housing units in Paterson) a roster was collected but none of the persons on the roster were collected independently of the Census.

As part of the evaluations of the 1995 Test, a sample of the noninterviews in Oakland was followed up and new estimates were calculated based on the results of the noninterview followup. Analysis in [3] and [5] found that the noninterview followup estimates were generally similar to the production estimates for both Census Plus and DSE (although average household size appeared to be smaller for converted noninterviews than for production interviews and there were differences for some Census Plus estimates). The results suggest that the noninterview adjustment did not generally have a major effect on the estimates and did not produce any substantial effect on the comparison between Census Plus and DSE.

2. **Characteristic Imputation**: The R-Sample imputation rates are generally low (less than 4% of persons in interviewed R-Sample households for all imputed variables in all three sites) and therefore would not be likely to have an important effect on the estimates. Age and tenure had the highest imputation rates, sex had the lowest. The imputation rates were slightly higher in Oakland than in the other two sites. Results in [6] suggest that the differences between Census and R-Sample imputation had little effect on the Census Plus estimates.

The P-Sample rates are somewhat lower than the R-Sample rates (less than 3% of P-Sample persons for all imputed variables in both DSE sites). Age and tenure had the highest imputation rates, sex had the lowest. The imputation rates were slightly higher in Oakland than in Paterson. Results in [6] suggest that the differences between Census and P-Sample imputation had little effect on the DSE estimates.

The E-Sample imputation rates are higher than the R and P sample rates (except for tenure). The imputation rates for age are about 10% (of E-Sample persons) and for race and hispanic origin are just over 6%. The imputation rates for sex and tenure are less than 2%. The effect of imputation on the final DSE estimates will partially cancel out since it affects both the numerator and the denominator of the DSE adjustment factor. Results in [6] do suggest that the E-Sample imputes fewer older renters compared to the Census. This tended to increase the estimated correct enumeration probabilities for some categories of renters in the oldest (50+) age category compared to the probabilities if Census data were used. Generally, however, the results in [6] suggest that the effects of the imputation procedures on the comparison between Census Plus and DSE were minor.

3. Modeling for Unresolved Status

R-Sample: The rate of unresolved residence status is highest in Oakland (5.22% of 22,086 persons from interviewed R-Sample households) and lowest in NW Louisiana (3.39% of 10,096). In Paterson, 4.05% of 21,769 persons had unresolved residence status. Although the rates are somewhat higher than hoped, they do not seem high enough to have a major impact on the pattern of Census Plus results, especially since the estimated probabilities were so strongly influenced by relationship to reference person. Relatives with unresolved residence status (1593 persons) always had high estimated residence status probabilities: the lowest estimated probability for a relative was .762. Persons who were nonrelatives, only in the Census, or had a missing relationship code were grouped together in the modeling (We will refer to this group of persons as nonrelative/unknown). The estimated residence status probabilities for the 782 unresolved persons in this group varied considerably, with a maximum of .993 and a minimum of .119.

There is also an interaction between the residence status group and relationship to reference person. Persons in the nonrelative/unknown relationship category with unresolved residence status tend to have a lower estimated residence status probability if they are in residence status group 1 (nonoutmovers with unresolved computer residence status code). The average estimated probability for persons in the nonrelative/unknown category is .3759 in group 1, .8828 in group 2, and .8347 in group 3 (group 2 is outmovers with an unresolved computer residence status code, and group 3 is all persons with a resolved computer residence status code). This result is expected, since a majority of the resolved persons in the nonrelative/unknown category from residence status group 1 are nonresidents (2578 nonresidents out of 4211 resolved) while most of the resolved persons in the nonrelative/unknown category from the other two residence status groups are residents (177 residents out of 191 resolved in group 2, 1566 residents out of 1759 resolved in group 3) .

The nonrelative/unknown category collapses together three different groups of people: nonrelatives, persons whose relationship is unknown because they were only in the Census, and persons from ICM whose relationship is unknown because relationship was not collected in ICM. Persons only in the Census turn out to be quite different from the rest of the nonrelative/unknown category. They are much more likely to be confirmed nonresidents. Most of the resolved persons who were only in the Census and are from residence status group 1 are nonresidents (2465 nonresidents out of 2740 resolved), while a majority of the resolved persons who were only in the Census are residents in the other two residence status groups (146 residents out of 157 resolved in group 2, 335 residents out of 528 resolved in group 3).

It appears that the estimated residence probability is largely driven by the relationship to reference person category (and its interaction with residence status group). There is one important exception to this: persons with resolved residence status from single-person households were always residents in the R-Sample and therefore the 262 persons from single-person households with unresolved residence status were given estimated probabilities very close to one (the minimum estimated probability for such persons was .9860). In the future, we should probably separate out persons who were only in the Census into their own category. It is also important to capture the interaction between residence status and outmover status for persons only in the Census.

P-Sample: The main result of the modeling for match probability is that the most important variable is the match status group. Persons with unresolved match status in match status groups composed entirely of before followup (BFU) nonmatches always get an estimated match probability close to zero (maximum estimated probability for these persons is 0.1156). Note that out of the 4052 persons with unresolved match status in the P-Sample almost all are due to either being sampled out of followup (3171 persons) or having insufficient information for matching (876 persons). Because of the strong relationship between BFU match status and final match status it does not seem likely that the P-Sample probability modeling had a major influence on the pattern of DSE estimates.

DSE followup in 1995 resolved the match status of almost all persons sent to followup (3526 out of 3528 persons sent to followup had resolved final match status) (Totals in this section exclude confirmed nonresidents unless otherwise indicated. DSE followup confirmed 236 persons as nonresidents). In addition, DSE followup never changed a BFU match to a nonmatch and almost never (only 8 out of 3182 BFU nonmatches were changed to matches by followup) changed a BFU nonmatch to a match. Possible matches could become either matches or nonmatches (118 out of 163 became matches). In 1995 we collapsed matches and possible matches with one category of nonmatches because of our worries over sample size. Looking at the results, we probably should not have done this collapsing. Because the DSE followup in 1995 resolved the match status of almost all persons sent to followup, it may be better to model the residence status for the P-Sample instead of the match status. Note that 339 of the persons sent to followup have unresolved residence status.

E-Sample: There does not appear to be any single variable that is strongly driving the estimated E-Sample correct enumeration probabilities. In fact, most of the variables do not seem to be strongly affecting the correct enumeration probabilities. Persons from housing units on the 1990 MAF (1990 Census address file) (5299 unresolved persons, average probability 0.899) did tend to have higher estimated probabilities than other persons (223 unresolved persons, average probability 0.636). In addition, persons from whole household before followup nonmatches where the housing unit did not match (307 unresolved persons, average probability 0.644) tended to have lower estimated probabilities than persons from other match code groups (5215 unresolved persons, average probability 0.902). There do not appear to be any major problems with the E-Sample modeling procedures nor does it appear likely that the modeling procedures had any important effect on the pattern of DSE estimates.

E. SUMMARY AND CONCLUSIONS

The 1995 ICM had three basic procedures for handling missing data. A noninterview adjustment procedure, that is outlined in Section C-1, was used to account for whole household noninterviews. Missing characteristics were filled in using a hot-deck imputation procedure outlined in Section C-2. Probabilities were calculated for persons with unresolved final status (residence status for R-Sample, match status for P-Sample, enumeration status for E-Sample) based on a logistic regression procedure outlined in Section C-3. The procedures in Sections C-1 and C-3 were similar to those used for the 1990 PES.

The analysis of the effects of the missing data procedures suggests the following:

- The missing data procedures for the 1995 ICM did not have important effects on the comparison between Census Plus and DSE.

- We should distinguish persons who were only in the Census from other persons when we model residence status. We also want to be able to capture any interaction between person outmover status and residence status, especially for persons only in the Census.
- We may want to model residence status instead of match status in the P-Sample. The DSE followup in 1995 resolved the match status of almost all persons (two people were unresolved) sent to followup but was unable to resolve the residence status of 339 persons (3189 persons were resolved as residents by DSE followup).
- The E-Sample modeling procedures appear to be satisfactory. We may wish to consider whether we should impute E-Sample data using data from the Census.

ACKNOWLEDGEMENTS

The authors wish to thank Tom Krenzke and Nick Alberti for helpful comments on earlier drafts of this manuscript.

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 R.A. Killion to J.H. Thompson, "DSSD 1995 Census Test Memorandum Series #U-2, Results from the 1995 Census Test: Responses to the Coverage Probes in the Integrated Coverage Measurement Person Interview - Project 4 (K. West, author)," January 29, 1996.

[3] Bureau of the Census internal memorandum from R.A. Killion to J.H. Thompson, "DSSD 1995 Census Test Memorandum Series #U-11, Results from the 1995 Census Test: Integrated Coverage Measurement Noninterview Followup - Evaluation Project 3 (P. Gbur, author)," March 14, 1996.

[4] Bureau of the Census internal memorandum from R.A. Killion to J.H. Thompson, "DSSD 1995 Census Test Memorandum Series #T-13, The 1995 Census Test: A Compilation of Results and Decisions (A. Vacca, M. Mulry, and R.A. Killion, authors)," April 1, 1996.

[5] R. Petroni, P. Gbur, and A. Kearney (1996), "Handling Noninterviews to Provide Equitable Comparisons of ICM Estimates," presented at the 1996 ASA annual meeting.

[6] R. Petroni, A. Kearney, and M. Ikeda (1996), "Imputation's Effect on 1995 Test Census Estimates," presented at the Seventh International Workshop on Household Survey Nonresponse, Rome, Italy.

[7] Bureau of the Census internal memorandum from R. Singh to R.A. Killion, "1995 Census Test Memorandum Series IS#18, Review and Evaluation of ICM Estimates for the 1995 Census Test (E. Schindler, author)," February 22, 1996.

[8] Bureau of the Census internal memorandum from R. Singh to J.H. Thompson, "1995 Census Test Memorandum Series IS#8, Computer Specifications for ICM Site Level Estimation for the 1995 Census Test--Revision 3 (E. Schindler, author)," June 18, 1996.

[9] Bureau of the Census internal memorandum from D. Whitford to R.A. Killion, "1995 Census Test Memorandum Series IP-MD-40, Documentation of the Design of the 1995 Integrated Coverage Measurement (D. Childers, author)," December 11, 1995.

[10] Bureau of the Census internal draft memorandum from T. Williams for Documentation, "Methodology used for the modeling of missing variables in the flexible matching imputation software," July 14, 1995.

[11] Bureau of the Census internal draft memorandum from T. Williams for Documentation, "Using the Flexible Matching Imputation Software," July 17, 1995.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.