HANDLING OF MISSING DATA IN THE 1996 INTEGRATED COVERAGE MEASUREMENT

Michael M. Ikeda, Anne T. Kearney, Rita J. Petroni, Bureau of the Census*
Anne T. Kearney, Bureau of the Census, Statistical Research Division, 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 1996 Integrated Coverage Measurement (ICM). 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 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. Unlike the Census imputation procedures, the ICM procedures do not include any data editing. Third, some persons will have an unresolved final residence, match, or enumeration status. Probabilities for the final statuses are calculated for these persons based on a modeling procedure outlined in Section C-3. The procedures in Sections C-1 through C-3 are similar to those used for the 1990 Post Enumeration Survey (PES). See [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 1996 Community Census was conducted in three sites: Chicago ,IL; Fort Hall, ID, and Acoma, NM. All blocks in the 1996 Community Census were also in ICM.

There were three separate rosters involved in the ICM missing data processing: the R-Sample, the P-Sample, and the E-Sample. Each roster was created for all three sites. The R-Sample was used in Census Plus estimation. Census Plus tries to obtain a "true" roster from the ICM blocks. Census Plus estimates are calculated based on the assumption that the R-Sample is the "truth" for the ICM blocks. The P and E-Samples were used in Dual System Estimation (DSE). 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. Further details on DSE and Census Plus estimation can be found in [12]. Final 1996 DSE and Census Plus estimates are given in [11].

In 1996, 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. In Census Plus there was a panel using administrative records (roughly half of the site). The preliminary Census roster used by Census Plus in this panel also included persons added by administrative records. Due to problems in obtaining the records, administrative records were not available for major portions of the Fort Hall and Acoma sites. 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 which does not include persons added by administrative records to form the E-Sample. An overview of the 1996 ICM operations is given in [15].

R-Sample: The R-Sample is the Resolved Roster of persons for Census Plus. 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 are 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 are added during ICM interviewing are also referred to as R-Sample housing units.

P-Sample: The P-Sample is created from the independent roster of persons. The P-Sample is used to estimate persons missed in the Census. The independent roster is collected from I-Sample housing units. I-Sample units are those housing units from an independent listing that are confirmed to exist on Census day. The P-Sample consists of those persons in the independent roster who are residents of I-Sample housing units 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 from the Census file. The extract is taken before the Census edit and imputation because of timing concerns. The E-Sample is used to estimate persons erroneously enumerated in the Census.

C. OUTLINE OF PROCEDURES

C.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.

The main noninterview adjustment is at the block cluster x type of place level. The type of place categories are collapsed into four categories for the adjustment: single-family, 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 type of place category is spread among the interviewed housing units in the same block cluster x type of place category. Special procedures are used for noninterviews with missing type of place.

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 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 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 stratum.

Missing Type of Place: Noninterviewed housing units with missing type of place are treated specially in the R-Sample. Their weight is spread over all interviewed housing units in the block cluster.

If the number of noninterviewed units with missing type of place 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 units in the same stratum.

C.2. Characteristic Imputation

Some persons in interviewed households had missing characteristics. Missing characteristics were filled in using a hot-deck imputation procedure. Characteristic imputation was performed on all three samples. Similar procedures were used for the R, P, and E-Samples. The variables imputed were tenure, sex, age, race, and Hispanic origin. These are the variables needed to create population estimates. Unlike the Census, the ICM imputation procedures did not include editing of data. The race imputation was performed on the five main race categories. Age imputation was performed on four age categories (0-17, 18-29, 30-49, 50+). Tenure imputation was performed on owner/renter. Hispanic origin imputation was performed on Hispanic/non-Hispanic. Imputation was done separately for each site. An overview of the characteristic imputation procedure is given in [13].

Tenure was imputed from the nearest previous unit with the same structure type (type of place is structure type for R and P-samples). Race was imputed from the distribution of race within the household or, if the whole household was missing race, from the distribution in the nearest previous household with nonmissing race. Hispanic origin was imputed from the distribution of Hispanic origin within the household or, if the whole household was missing Hispanic origin, from the distribution in the nearest previous household with 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.

The most complicated imputation procedure was for sex. For a reference person (spouse present) or spouse of reference person, the person with a missing value of sex was assigned the sex opposite to that of their spouse. If both reference person and spouse had sex missing, then we imputed sex for the reference person based on the distribution of sex for reference persons with spouse present and assigned the spouse the sex opposite to the sex assigned to the reference person. The same procedure was followed if the spouse of reference person had sex missing but there was no reference person in the household (except that sex was not actually assigned to a reference person). For one-person households, sex was imputed based on the distribution of sex for one-person households.

For a reference person (no spouse present) in a multi-person household, sex was imputed from the distribution of sex for reference persons with no spouse present in multi-person households. For other persons with non-missing relationship (except for spouse of reference person) from multi-person households we imputed sex based on the distribution of sex for persons (excluding reference persons and spouses of reference persons) with nonmissing relationship from multi-person households. For persons with missing relationship from multi-person households, sex was imputed based on the distribution of sex for persons (excluding householders) from multi-person households.

C.3. Modeling of Probabilities

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

Probabilities for persons with unresolved final status were calculated using a model fit on persons with resolved final status. Resolved final status was determined from a field followup of persons designated as requiring followup. The model contained both general parameters (fit using all persons) and group parameters (fit using persons in the given group). Persons were assigned to groups based on their before followup status in combination with other variables. The model parameters (both general and group) were generally similar to the parameters used in the 1990 PES. Residence status probability was modeled for the R-Sample and P-Sample, and correct enumeration probability was modeled for the E-Sample. A complication for the P and E Samples was that roughly half of the persons needing followup were sampled out of followup. Persons needing followup but sampled out were considered to have unresolved final status.

There were also some P-Sample persons with unresolved match status after followup. These persons were possible matches or had insufficient information for matching. The match probability for P-Sample persons with unresolved match status was calculated to be the proportion of matches among those persons with resolved match status (excluding confirmed nonresidents). The calculation was done separately for each site.

R-Sample Residence Status Groups: In the R-Sample, group parameters were fit within residence status group. The before followup residence status groups were based on the residence status code assigned in the ICM interview and whether a person was classified as an outmover or not. The groups were:

1. Unresolved Residence Status, not outmover
2. Unresolved Residence Status, outmover
3. Resolved Residence Status

P-Sample Match Code Groups: In the P-Sample, group parameters were fit within match code groups. The match code groups were based on the before-followup match codes, before followup whole/partial household match code, address code, and person followup flag for persons who needed followup. The definitions of the groups are given below:

1. Possible matches and matches needing followup.
2. Nonmatches in partial household nonmatch needing followup.
3. Nonmatches needing followup in whole household nonmatches where address is matched.
4. Nonmatches needing followup in whole household nonmatches where address is not matched.
5. Matches and nonmatches where followup is not needed.
6. Insufficient information for matching

The predicted residence probabilities for persons in group 6 were calculated by taking a weighted average of the probabilites that were assigned to groups 1-5. The weighting was by the frequency of groups 1-5, with groups 1-4 double weighted to account for sampling for followup.

E-Sample Match Code Groups: The E-Sample match code groups are based on the before-followup match codes, whole/partial (before followup) household match code, address code from HU matching, and followup flag. The definitions of the groups are given below:

1. Possible Matches and Matches sent to followup.
2. Nonmatches in partial household nonmatch sent to followup.
3. Nonmatches sent to followup in whole household nonmatches where address is matched.
4. Nonmatches sent to followup in whole household nonmatches where address is not matched.
5. Matches and Nonmatches where followup is not needed.
6. Insufficient information for matching.

Persons in group 6 are given a probability of correct enumeration equal to 0.

D. MISSING DATA RESULTS

D.1. Source of R-Sample Persons

Persons in the 1996 R-Sample could come from three sources: ICM interview, Census enumeration, administrative records. If a person came from more than one source, the ICM interview data took priority over other data and the administrative records data had the lowest priority. Over 97% of R-Sample persons (that is, confirmed and possible residents in interviewed households) in each site were from the ICM interview: 17730 out of 18236 in Chicago, 5332 out of 5466 in Fort Hall, and 2698 out of 2731 in Acoma. The number of residents added by administrative records was only 48 in Chicago, 10 in Fort Hall, and 3 in Acoma (unresolved persons who were only from administrative records were not included in the production R-Sample).

D.2. Noninterview Adjustment

Table 1 gives the noninterview rate by site for the R- and P-samples (note that all tables in this paper are based on unweighted counts). Note the relatively high noninterview rates in Chicago.

Table 1: Noninterview (NI) Rates (%)

	R-Sample		P-Sample	
	NI Rate(%)	Occ HU	NI Rate(%)	Occ HU
Chicago	9.53	7848	9.28	7470
Fort Hall	2.84	1690	3.04	1644
Acoma	2.08	626	2.11	615

Occ HU is the total number of occupied housing units.

Analysis in [2], [9], [14] from the 1995 ICM found that the estimates were fairly robust to different methods of handling noninterviews. In addition, [3] suggests that the choice of noninterview adjustment method did not have a major effect on the 1996 DSE estimates. The results suggest that the noninterview adjustment method did not produce a substantial effect on either Census Plus or DSE estimates.

D.3. Characteristic Imputation

Table 2 gives the item imputation rates for Chicago for the five variables that were imputed. The imputation rates are generally fairly low for the R and P-Samples. The R-Sample rates are somewhat higher than the P-Sample rates. The E-Sample imputation rates are higher than the rates for the R- and P-samples.

Table 2: Item Imputation Rates (Percent)

Chicago	R-Sample	P-Sample	E-Sample
Tenure	0.21	0.14	4.34
Sex	0.22	0.14	3.22
Age	2.65	1.77	5.77
Hispanic Origin	1.54	0.95	19.14
Race	4.04	3.80	10.86

The R-Sample and P-Sample rates are for residents and possible residents from interviewed households. The E-Sample rates are for all E-Sample persons.

The R-Sample and P-Sample imputation rates are generally low and therefore would not be likely to have an important effect on the estimates. Results from the 1995 ICM ([6],[7],[10]) generally support this assumption.

The E-Sample imputation rates are higher but 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 from the 1995 ICM ([6],[7],[10]) suggest the imputation procedures have no important affect on the comparison between Census Plus and DSE.

D.4. Modeling for Unresolved Status

D.4.a General Overview

Table 3 gives information on the proportion of persons with unresolved status. Most of the persons with unresolved enumeration status in the E-Sample and unresolved residence status in the P-Sample are due to being sampled out of DSE followup. 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. Also note that before followup matches sampled out of followup are assumed to be matches and before followup nonmatches sampled out of followup are assumed to be nonmatches.

The rate of unresolved residence status in the R-Sample is highest in Chicago, although it does not seem high enough to have a major impact on the Census Plus results. In general, it does not appear likely that the modeling of probabilities has any major effect on eiter Census Plus or DSE estimates. This is supported by results from the 1995 ICM [4],[5],[8].

Table 3: Unresolved Residence Status by Site (Percent)

	R-Sample	P-Sample	E-Sample
Chicago	1.90	13.47	10.84
Ft Hall	0.86	7.32	7.12
Acoma	0.37	11.03	7.11

R-Sample and P-Sample percentages are percentages of residents and possible residents from interviewed households. E-Sample percentages are percentages of E-Sample persons.

R-Sample and P-Sample percentages refer to persons with unresolved final residence status. E-Sample percentages refer to persons with unresolved final enumeration status.

D.4.b. R-Sample

Table 4 gives summary statistics on estimated residence probabilities for selected variables. The most important variable in the R-Sample modeling seems to be relationship to reference person. Unresolved persons only from the Census (and to a lesser extent persons with missing relationship) generally have substantially lower estimated residence probabilities.

Table 4: Summary Statistics for Estimated Residence Probabilities for Unresolved Persons from Interviewed Households

Site	N	Mean	Std. Dev.*
Chicago	347	0.5374	0.3478
Ft Hall	47	0.4151	0. 3751
Acoma	10	0.2656	0.2524
Tenure	N	Mean	Std. Dev.*
Owner	145	0.3874	0.3254
Renter	259	0.5886	0.3474
Relationship	N	Mean	Std. Dev.*
Ref Person	155	0.7214	0.2140
Spouse	34	0.7071	0.2578
Child	34	0.7342	0.2617
Sibling	12	0.7172	0.1500
Parent	2	0.8104	0.0898
Other Rel	17	0.7214	0.2131
Nonrelative	9	0.5923	0.3192
Census Only	123	0.0924	0.1538
Missing	18	0.4787	0.2469

* These are standard deviations for the elements, not the mean.

D.4.c. P-Sample

There does not appear to be any single variable that strongly drives the estimated P-Sample residence probabilities. In fact, most of the variables do not seem to be strongly affecting the residence probabilities. Fort Hall tends to have somewhat lower estimated probabilities than the other sites. Table 5 gives summary statistics of the estimated residence probabilities for a sample of the variables.

Table 5: Summary Statistics for Estimated Residence Probabilities for Unresolved Persons from Interviewed Households

Site	N	Mean	Std. Dev.*
Chicago	2262	0.8769	0.0932
Ft Hall	379	0.7773	0.1221
Acoma	291	0.9349	0.0483
Tenure	N	Mean	Std Dev*
Owner	1165	0.8765	0.0942
Renter	1767	0.8653	0.1067
Relationship	N	Mean	Std Dev*
Ref Person	982	0.8952	0.0922
Spouse	319	0.8974	0.0775
Child	847	0.8593	0.1053
Sibling	100	0.8672	0.0952
Parent	47	0.8699	0.0734
Other Rel	370	0.8298	0.1211
Nonrelative	242	0.8439	0.0781
Missing	25	0.7250	0.1630

* These are standard deviations for the elemets, not the mean.

We see in Table 6, that followup in 1996 resolved the match status of almost all persons sent to followup. We also see that followup never changed a before followup match to a nonmatch and almost never changed a before followup nonmatch to a match. Possible matches could become either matches or nonmatches (but more frequently became matches). More than 10% of the persons sent to followup (excluding confirmed nonresidents) have unresolved residence status (the persons with final match codes of P, MU, or NU in Table 6). Note that followup confirmed 290 persons as nonresidents.

Table 6: Before Followup Match Code and Final Match Code for P-Sample Persons Sent to Followup (Except for Confirmed Nonresidents)

	Final Match Code						
BFU Match Code	M	MR	MU	NR	NU	P	Total
Match (M)	2	54	14	0	0	0	70
Nonmatch (NP)	0	10	1	1771	263	1	2046
Poss Match (P)	0	45	1	17	4	2	69
Total	2	109	16	1788	267	3	2185

M and MR are matched resident

MU is matched with unresolved residence status

NR is nonmatched resident

NU is nonmatched with unresolved residence status

P is possible match

D.4.d. 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. However, matches and possible matches tend to have somewhat higher estimated probabilities than other persons, while persons from whole household nonmatches where the housing unit did not match tend to have somewhat lower estimated probabilities than persons from other match code groups. Table 7 gives summary statistics of estimated correct enumeration probabilities. Acoma tended to have lower estimated probabilities than the other two sites.

Table 7: Estimated Correct Enumeration Probabilities for Unresolved Persons

Site	N	Mean	Std. Dev.*
Chicago	1982	0.7754	0.1295
Ft Hall	398	0.8622	0.0908
Acoma	185	0.6356	0.1836
Tenure	N	Mean	Std Dev*
Owner	932	0.7994	0.1531
Renter	1633	0.7670	0.1283
Relationship	N	Mean	Std Dev*
Ref Person	1017	0.7958	0.1210
Spouse	264	0.8119	0.1308
Child	652	0.7853	0.1346
Sibling	75	0.8335	0.1002
Parent	25	0.7483	0.1366
Other Rel	246	0.7503	0.1601
Nonrelative	212	0.7075	0.1468
Missing	74	0.6241	0.1726

* These are standard deviations for the elements, not the mean.

E. CONCLUSIONS

The 1996 ICM had three basic procedures for handling missing data.

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

- For Census Plus, we should distinguish persons who were only in the Census from other persons when we model residence status.
- We may want to continue to model residence status instead of match status in the P-Sample. The followup in 1996 resolved the match status of almost all persons (3 people were unresolved) sent to followup but was unable to resolve the residence status of 286 persons (1899 persons were resolved as residents by 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.

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-11, Results from the 1995 Census Test: Integrated Coverage Measurement Noninterview Followup - Evaluation Project 3 (P. Gbur, author)," March 14, 1996.

[3] 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 Census Dress Rehearsal Memorandum Series A-25 (A. Kearney, author)," December 11, 1997.

[4] 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.

[5] 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 Census 2000 Dress Rehearsal Memorandum Series A-23 (M. Ikeda, author)," January 5, 1998.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] Bureau of the Census internal memorandum from R. Singh for documentation, "1996 Community Census Memorandum Series IS-10, Estimation Review Results for the 1996 Community Census Test (E. Schindler, author)," October 31, 1997.

[12] Bureau of the Census internal memorandum from R. Singh to J.H. Thompson, "1996 Community Census Memorandum Series--IS#8 (Working Draft), Computer Specifications for ICM Site Level Estimation for the 1996 Community Census (E. Schindler, author)," February 25, 1997.

[13] Bureau of the Census internal memorandum from R. Singh to M. Lynch, "Overview of Characteristic Imputation, Noninterview Adjustment and Logistic Regression Programs for the 1996 ICM (Draft) (A. Kearney and M. Ikeda, authors)," July 16, 1997.

[14] R. Singh and R. Petroni (1997), "Handling of Household and Item Nonresponse in Surveys," presented at the Eighth International Workshop on Household Survey Nonresponse, Mannheim, Germany.

[15] Bureau of the Census internal memorandum from D. Whitford to R.A. Killion, "1996 Census Test Memorandum Series IP-D-22, The Design of the 1996 Integrated Coverage Measurement (D. Childers, author)," October 2, 1996.

* 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.