Alternatives of the A.C.E. Missing Data Evaluation

Anne T. Kearney*, Donald H. Keathley*, Thomas R. Belin, UCLA, Rita J. Petroni*, *U.S. Census Bureau
Anne T. Kearney, U.S. Census Bureau, Room 1107-2, Mail Stop 9200, Washington, DC. 20233-9200
Anne.Theresa.Kearney@census.gov

Keywords: noninterview adjustment; late data; logistic regression; nonignorable missingness

1. BACKGROUND[1]

The U.S. Census Bureau conducted Census 2000 to determine the count of housing units and persons in the United States. As in any census or survey operation, Census 2000 is subject to coverage error. Historically certain minority groups are undercounted. To measure the coverage error in Census 2000, the Census Bureau fields a coverage measurement survey after finishing the census. The coverage measurement survey is called the Accuracy and Coverage Evaluation (A.C.E.). The A.C.E. is a national sample of approximately 300,000 housing units.

One sample is the population sample or the P-Sample. We collect data for the P-Sample during A.C.E. field operations. The census enumerations in A.C.E. blocks are called the enumeration sample or the E-Sample. Two samples are matched to estimate the number of people missed in the census, and to measure the number of erroneous enumerations in the census. Results from matching the two lists to measure the error are used in dual system estimation to measure coverage error (Childers, 2001).

The A.C.E. computes dual system estimates (DSEs) at the post-stratum level. We form post-strata groups defined by demographic and operational variables to decrease heterogeneity bias. Post-stratum level DSEs can then be added to form higher level estimates. See Griffin (2000) for details on dual system estimation.

As in most surveys, there are missing data in the A.C.E. from whole household noninterviews and item nonresponse. Specifically, the missing data include:

- Noninterviews for P-Sample households
- Interviews with some or all of the following:
  - missing demographic characteristics for P-Sample persons - imputation for E-Sample persons was not necessary
  - unresolved match and/or resident status for P-Sample persons
  - unresolved enumeration status for E-Sample persons

The unresolved status cases result from either incomplete data on residence or enumeration status such as where the person was living on census day, or incomplete data such that we were not able to discern with certainty if a P-Sample and an E-Sample person matched.

To account for these missing data we implement a set of missing data procedures. The A.C.E. missing data procedures include the following operations. First, spread the noninterviewed household weights over P-Sample interviewed households in the same adjustment cell. Second, use distributions, hot decks and combinations of these to impute race, ethnicity, sex, age, tenure. Finally, use an imputation cell estimation procedure to impute missing resident, match, and enumeration status probabilities (Cantwell, 2001b).

2. METHODS

We wanted to evaluate the A.C.E. production missing data procedures with a sensitivity analysis. This analysis involved the calculation of DSEs using alternative missing data procedures on the same A.C.E. data. The resulting range of DSEs would give us an indication of how sensitive the DSEs were to changes in one or more of the missing data procedures. For an analysis of the potential bias see Spencer, et al., (2002).

Note, we did not consider all alternatives as possible candidates for the A.C.E. production missing data procedures, so the spread in the DSEs from the alternatives may be larger than if we only used candidate alternatives. In particular, the non-ignorable missingness alternatives discussed below implement assumptions about how observed data differs from unobserved data from very limited information from the 1990 Post Enumeration Survey. But we do not have an objective way to prove these assumptions since we do not have data on the unobserved people. Therefore, our original thinking was the nonignorable missingness alternatives are useful for evaluation and sensitivity analysis, but not for point estimates. This conclusion was substantiated by our analyses (see Keathley, et al 2002).

We did not include any demographic characteristic imputation alternatives based on judgement that this contribution to uncertainty would be small relative to A.C.E. sampling variability. We used a hot-deck procedure for race, ethnicity, and tenure. Race and ethnicity also depend on within household empirical

[1] 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.

distributions. For these three variables, over 91% of the imputations had donor and donee in the same primary sampling unit (cluster) so that production variance estimates account for most of the associated variation. Imputation of age and sex was based on random draws from a national distribution. The sampling error in the distribution is relatively small, due to the large sample size of the A.C.E., and thus may be ignored. The level of variance that would arise from imputation in a simple random sample would be accounted for by the Bureau's estimate of sampling variance, which does not include finite population corrections (Cochran 1977, Section 13.10). The actual variance will be somewhat higher in practice, due to block-clustering that induces a correlation (plausibly, positive) among errors in the imputations within a block cluster (Cochran 1977, Section 13.11). However, the fraction of the sample that receives age and sex imputation is so small – under 2.5% in the A.C.E. (Kostanich (2001)) – that the effects on variance must be small as well. Thus, the estimates of sampling variance will account to a sufficient degree for the variance from imputation of age and sex.

First, we will describe the production methodology, and after that we describe each alternative.

2.1 Production Procedure for Noninterview Adjustment (NIA)

In A.C.E. production, the adjustment cells for noninterviewed housing units were defined on *block cluster* × *type of basic address*. A block cluster was a group of contiguous blocks or possibly a single block. The three types of basic address were single-family units, apartments, and all others. Generally, weights of noninterviewed units were spread uniformly over interviewed units in the same cell. When there were more than twice the number of noninterviews compared to the number of interviews, we spread the weights over a larger category of interviewed units. The next larger category was recoded sampling stratum (largely defined by demographic-tenure make up of the cluster) by type of basic address. An entire cluster belonged to one and only one recoded sampling stratum. Next, if there still were not enough interviews, the adjustment category was block cluster (without regard to type of basic address); next, recoded sampling stratum (without regard to type of basic address), and the final adjustment category was across all clusters in the state. Note that cells were never collapsed together. Rather, weights of non-interviewed units in a cell with too few interviews were spread over a broader category, but weights of noninterviewed units in all other cells were still spread only within their cell. See Cantwell (2001b) for a complete description of the production noninterview adjustment procedure.

2.2 Production Imputation of Match, Residency, and Correct Enumeration Probabilities

To impute for unresolved match and residence status (for P-Sample people) and enumeration status (for E-Sample people), the A.C.E. production used imputation cell estimation. First, all people were placed in cells formed from a few relevant operational and demographic characteristics maintaining acceptable cell sizes; different imputation cells were used for match, residence, and correct enumeration statuses. Then the proportion matched (or resident or correct enumeration) in each cell was calculated from the people in the cell with resolved status, and this proportion was assigned as a probability to each person in the cell with unresolved status. See Cantwell (2001b) for a complete description of imputation cell estimation.

2.3 Alternative Noninterview Adjustment Cell Definitions

As stated in Section 2.1 above, the production NIA used block cluster and type of basic address (along with collapsing rules) to form adjustment cells. As an alternative to this, we performed the same type of NIA (spreading the weights of the noninterviews to the interviews in the same cell) using different cell definitions. We hypothesized that cells formed by demographic and operational variables may be more homogeneous with respect to whole household response propensity than the variables used in production.

The variables used to define the alternative cells in order of priority were: type of basic address; a recoded variable combining race, Hispanic origin indicator, and tenure of householder; census division; state within division; type of enumeration area; and household size. So, if there were not enough interviews compared to noninterviews in a cell, we spread the weights of the noninterviews over a broader category. Note, like production, cells were never collapsed together. If there were too many noninterviews in the cell, we dropped a variable off the list, and spread the weight of the noninterviews of that cell over the interviews in the collapsed cell. For example, the first variable to drop out was household size. Then we compared the number of noninterviews in the original cell to the number of interviews in the new cell. If there still were not enough interviews then type of enumeration area dropped out, and if there were enough interviews we spread the weight of the problem cell over type of basic address, recoded race/Hispanic origin/tenure of householder, census division and state within division.

The variables race, Hispanic origin, tenure, type of enumeration area, and household size were obtained by matching the P-Sample housing unit file to the census Hundred Percent Census Edited File to pick up census information about the matching household. Special codes were assigned when a P-Sample housing unit matched to a vacant census housing unit, and also when the P-Sample housing unit did not match to the census which kept all housing units of these types in the same cells (Kearney, 2002a).

2.4 Nearest Neighbor Noninterview Adjustment Procedure

The production noninterview adjustment procedure spread the weights of the noninterviews to all the interviews in the appropriate noninterview adjustment cell. If there were not enough interviews compared to noninterviews in the cell, the weight of the noninterview was spread to a broader category. Depending on the size of the adjustment cell, the interviews getting the weight of the noninterviews may be geographically distant from each other, possibly increasing the chance that they would be different with respect to response propensity. As an alternative, we implemented the nearest neighbor procedure. This procedure gives the whole weight of a noninterview in the cell to only one interview, the nearest previous neighbor. If there were more noninterviews than interviews in a cell, then we donated the weight to an interview in a broader category, working our way back up the list. In the other noninterview adjustment alternatives, we went to a broader category when there are more than twice the number of noninterviews compared to interviews. In this alternative, so as not to give any housing units too much influence, we allow an interviewed housing unit to take the weight of only one noninterview. Moving up the list in the broader category to find a donee for the noninterviewed housing unit's weight puts distance between the donor and the donee, but potentially less distance than implied by the production procedure since we stop proceeding up the list of housing units when we find an acceptable donor. The production method spreads the weight from the top to the bottom of the category (Keathley, 2002).

2.5 Late Data Alternative

2.5.1 Whole Household Noninterviews Adjustment Using Late Data

We use a housing unit level file for the whole household noninterview adjustment.

The production noninterview adjustment procedure spreads the weights of the noninterviews to all interviews within an adjustment cell. The interviews in the cells may differ on certain characteristics. For example, during the data collection process, there has to be a time when we stop our attempts to collect data from the hard to reach households. Towards the end of the data collection process interviewers try to get information from people who may not be home much, who refuse to participate, or who for some other reason are difficult to contact and interview. If we had unlimited resources and time, we could conceivably continue our attempts until we interviewed every sampled housing unit. It is possible that the interviews obtained late in the interview process are more like the noninterviews than the interviews obtained early in the process. (See Bates and Creighton, (2002).) Therefore, in this alternative, we separate the "late" interviews from the "non-late" interviews, and spread the weights of the noninterviews over only the late interviews leaving the weights of the non-late interviews unchanged.

Late and non-late interviews were classified as follows:

a. Housing units completed by telephone are not considered late interviews. The A.C.E. released the telephone phase before the personal visit phase. The portion of the A.C.E. sample that went to the telephone phase was the early responders to Census 2000 who reported their telephone number on their census form. From the start, this set of respondents have been cooperative. The telephone interviewers were instructed to avoid hard core refusals by accepting refusals earlier than they typically would. Then the telephone refusals were held over to be interviewed during the personal visit phase. (Telephone refusals are not considered late.) So, even though some telephone interviews may have been difficult to obtain, we judged that in general the interviews obtained during the telephone phase are from cooperative respondents, and these respondents are probably unlike the typical late responder, or even further unlike a noninterviewed household in terms of response propensity.
b. Nearing the end of the field work, there is an operation designed to try to convert housing units that were currently noninterviews to interviews. All housing units that were part of the nonresponse conversion operation were considered late respondents.
c. For the remaining interviewed housing units, the last 30 percent of the interviews in each local census office were considered late respondents. The remaining 70 percent were treated as non-late respondents. We selected a cutoff that would give us adjustment cells of adequate size, but would still capture some of the differences between late and non-late respondents.
d. All whole household noninterviews were included with the late respondents.
e. Vacant and deleted housing units were excluded from the late respondents (Kearney, 2002b).

2.5.2 Unresolved Match, Residence, and Enumeration Status Probabilities Using Late Data

The P-Sample person level late data includes:

a. All persons with unresolved residence or match status
b. All persons in late housing units

The E-Sample person level late data includes:

a. All persons with unresolved enumeration status
b. All persons in housing units who had their census data collected through the Census 2000 Nonresponse Followup (NRFU) operation

For residency and match status imputation we used the production imputation cells to assign probabilities. See Tables A.1 and A.2 for residence and match probability cells, respectively. With the E-Sample definition of late data, one of the production imputation cells for enumeration status contained zero resolved people, so we collapsed followup groups 5 and 6 before computing probabilities. See Table A.3 for the production correct enumeration imputation cells.

2.6 Logistic Regression to Impute Match, Residence, and Enumeration Status

As described in Section 2.2, the production probabilities were calculated using a few variables to form imputation cells. It was necessary to limit the number of predictor variables to maintain acceptable cell sizes. The logistic regression alternative finesses the problem of the size of imputation cells by assuming that cases are exchangeable given a detailed set of covariates whose relationship to the logistic transformed probability of occurrence of the outcome of interest is summarized by a linear combination of predictors (Belin, 2001). The extensive information available on many unresolved cases makes it possible to include dozens of predictors in such a logistic regression model (see Cantwell, 2001a for the list of available variables).

2.7 Nonignorable Missingness

All of the missing data alternatives above assume that the missing data is ignorable at least within the adjustment cells. However, it may be nonignorable. This section describes the procedure used to adjust for nonignorability.

Fitting a logistic regression model to the cases that were resolved during follow-up as described in Section 2.6 or using the imputation cell estimation methodology used in production to predict probabilities for unresolved cases implicitly assumes that the logistic-transformed outcome probability or the imputation cell probability depend only on observed covariates, and not residually on the fact that the case remained unresolved. Such a model is known as an ignorable model (Rubin, 1976). There is no evidence in the data from the A.C.E. survey by itself to contradict an assumption of ignorable unresolved status, but it is possible that evaluation follow-up[2] (EFU) data would suggest that ignorable model predictions are either too high or too low. That is, EFU data might indicate that ignorable-model predictions exhibit some systematic bias. To explore the possible extent of such effects, we also implemented nonignorable alternatives.

The logic underlying the nonignorable alternatives implemented here joins a conceptual model with empirical evidence from previous coverage-measurement survey experience. The logistic regression models developed here control for dozens of predictors, all of which are entered as binary variables. Suppose we consider a binary predictor that is not controlled that explains the bias in model predictions. A question of practical interest is what the magnitude of the logistic-regression coefficient of such a predictor might plausibly be.

We proceed to anchor an estimate of how large such a coefficient might be using 1990 Post Enumeration Survey (PES) data. Belin, et al. (1993) contrasted 1990 PES predictions with 1990 EFU findings and found that the overall average PES predicted probability of 0.322 for unresolved cases that were resolved in the 1990 EFU compared favorably to the overall 1990 EFU match rate of 0.316 for those same cases. In addition, Belin, et al. (1993) cited 12 comparisons for various subgroups between average predicted probabilities from a logistic regression model and observed 1990 EFU match rates where the number of available 1990 EFU cases was at least 15. The largest difference was between an average predicted probability for cases with no PES follow-up attempted (0.693) and an observed enumeration rate of 0.500 among 18 cases in the 1990 EFU. The next largest difference was between an average predicted probability of 0.773 for cases with predicted PES probability between 0.67 and 1.00 and an observed enumeration rate of 0.615 among 52 such 1990 EFU cases. Here, we summarize the magnitude of a logistic regression coefficient that could plausibly have produced the discrepancy between 0.773 and 0.615. We choose to focus on this case in part because the most extreme case is apt to be less predictable, in part because the larger discrepancy was based on only 18 cases, and in part because it is common in statistics to develop procedures that account for usual rather than atypical deviations.

Because the selection of cases into the subgroup with PES probability between 0.67 and 1.00 is based on being in a restricted range in one tail of the distribution of predicted PES probabilities, the discrepancy between the average PES probability and the EFU match rate for such cases can be expected to be due in part to a regression-to-the-mean phenomenon. That is, it would be entirely expected for the average EFU match rate to be lower for these cases than the average PES match probability among these cases, just as it would be entirely expected for the average EFU match rate to be higher than the average PES match rate for the cases with predicted PES probability between 0 and 0.33. (For the 190 cases in the 1990 PES with predicted probability between 0 and 0.33 that were resolved in the 1990 EFU, the average PES predicted probability was 0.125 and the observed EFU match rate was 0.195.) Thus, if we are able to account for the observed difference in a nonignorable model, it is arguable that we would be more than adequately accounting for potential nonignorable effects.

Alternative procedures described in greater detail in Belin (2001) suggested estimating the magnitude of a parameter

[2] Evaluation Followup (EFU) is a subsample of production block clusters. A sample of people in EFU clusters were recontacted to measure error in assignment of residence, match, and enumeration status.

for potential nonignorable effects using the 90th percentile value of the order statistics of the absolute values of the logistic regression parameter. That is, since the 90th percentile order statistic among the absolute values of the logistic regression parameters from the 1990 PES explained the observed difference between 1990 PES predictions and 1990 EFU findings, we assumed that the 90th percentile order statistic from 2000 A.C.E. logistic regression might similarly explain differences between 2000 A.C.E. predictions and 2000 EFU findings.

We took possible nonignorable missingness into account by lowering the match, residence, or correct enumeration probabilities for all persons with imputed status.

3. RESULTS

For a description of the results, see Keathley, et al. (2002).

4. REFERENCES

Bates, N. and Creighton, K. (2000) "The Last Five Percent: What Can We Learn from Late Interviews?" in 2000 Section on Government and Social Statistics Proceedings, pp. 120-125. American Statistical Association: Alexandria , VA.

Belin, T.R., Diffendal, G.J., Mack, S., Rubin, D.B., Schafer, J.L., Zaslavsky, A.M. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation" (with discussion), *Journal of the American Statistical Association*, 88:1149-1166.

Belin, Tom R. (2001), *Evaluation of Unresolved Enumeration Status in 2000 Census Accuracy and Coverage Evaluation Program.* Document prepared by Datametrics, Inc. for the Bureau of the Census (contract number 50-YABC-7-66021)

Cantwell, P.J. (2001a), DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM Q-40, from Patrick Cantwell for Maureen Lynch dated January 19, 2001, Accuracy and Coverage Evaluation Survey: Specifications for Data Requirement for Missing Data Input and Output Files. Internal Census Bureau Memorandum.

Cantwell, P. (2001b) DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM Q-62 from Patrick Cantwell for documentation dated 7/9/01, Accuracy and Coverage Evaluation Survey: Specifications for the Missing Data Procedures; Revision of Q-25.

Childers, D.R. (2001) DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES, Chapter S-DT-1, Revised, Accuracy and Coverage Evaluation: The Design Document dated [illegible]

Cochran, W. G. (1977) *Sampling Techniques.* 3rd ed. New York: Wiley.

Griffin, Richard (2000), *Accuracy and Coverage Evaluation Survey: Dual System Estimation.* Internal Census Bureau memorandum, DSSD Census 2000 Procedures and Operations Memorandum Series Q-20.

Kearney, A.T. (2002a), PLANNING, RESEARCH, AND EVALUATION DIVISION TXE/2010 MEMORANDUM SERIES: CM-MD-S-03-R1, Specifications for the Pseudo-Block Cluster Alternative Cell Definition for the Noninterview Adjustment in Evaluation of the A.C.E. Missing Data Procedures, dated March 6, 2002

Kearney, A.T. (2002b), PLANNING, RESEARCH, AND EVALUATION DIVISION TXE/2010 MEMORANDUM SERIES: CM-MD-S-04, Definition and Processing Specifications for the Late Data Alternative Noninterview Adjustment and Definition and Creation of Input Files for the Late Data Imputation Cell Estimation Procedure in the Evaluation of the A.C.E. Missing Data Procedures, internal memorandum dated March 6, 2002, from Petroni, R.J. for documentation.

Keathley, D.H. (2002) PLANNING, RESEARCH, AND EVALUATION DIVISION TXE/2010 MEMORANDUM SERIES: CM-MD-S-05, Documentation of the Programming for the Nearest Neighbor Non-Interview Alternative Procedure in the Analysis of Missing Data for the Accuracy and Coverage Evaluation, 3/14/02

Keathley, D., Belin, T., Bell, W., Kearney, A.T., Petroni, R. (2002) Analysis of the Missing Data Alternatives for the 2000 A.C.E., Survey Research Methods Section of the Proceedings of the 2002 Joint Statistical Meetings of the American Statistical Association, forthcoming.

Kostanich, D. (2001) DSSD Census 2000 Procedures and Operations Memorandum B-7* from Donna Kostanich to Howard Hogan dated February 28, 2001, Accuracy and Coverage Evaluation: Missing Data Results.

Rubin, D.B. (1976), "Inference and Missing Data," *Biometrika*, 63:581-592.

Spencer, B.D., Kearney, A.T., Keathley, D.H., Petroni, R., Belin, T., Mulry, M.H., Quantifying Bias from Missing Data Procedures in the 2000 A.C.E. dated August 1, 2002 Draft

Table A.1. Imputation Cells Used to Assign Production Resident Probabilities to P-Sample People in the United States (36 cells)

Follow-Up Group	Owner				Non-Owner			
	Non-Hispanic White		**Other**		**Non-Hispanic White**		**Other**	
1 = Matches needing follow-up								
2 = Possible matches								
3 = Nonmatches needing follow-up from partial household nonmatches	3a. 18-29 child of reference person	3b. Other	3a. 18-29 child of reference person	3b. Other	3a. 18-29 child of reference person	3b. Other	3a. 18-29 child of reference person	3b. Other
4 = Nonmatches needing follow-up from whole household nonmatches (not conflicting households)								
5 = Nonmatches from conflicting households needing follow-up								
6 = People resolved before follow-up								
7 = People with insufficient information for matching	weighted average over groups 1 - 5 and 8		weighted average over groups 1 - 5 and 8		weighted average over groups 1 - 5 and 8		weighted average over groups 1 - 5 and 8	
8 = After Follow-Up Code (potentially fictitious or potentially lived elsewhere on Census Day)								

Table A.2. Imputation Cells Used to Assign Match Probabilities to P-Sample People in the United States (7 cells)

Mover Status	Address Code			
	ADDCDE = 1 (HU Match from initial matching)		ADDCDE = 2 or 4 (HU Nonmatch or conflicting household)	
Non-mover (MOVERPER =1)	AMTIMP = 0	AMTIMP = 1	AMTIMP = 0	AMTIMP = 1
Out-mover (MOVERPER = 3)	AMTIMP = 0	AMTIMP = 1	Combine two imputation groups to form this cell	

Table A.3. Imputation Cells Used to Assign Production Enumeration Status to E-Sample People in the United States (29 cells)

Follow-Up Group	AMTIMP = 0		AMTIMP = 1 or 2	
1 = Matches needing follow-up				
2 = Possible matches				
3 = Nonmatches from partial household nonmatches	3a. 18-29 child of reference person	3b. Other	3a. 18-29 child of reference person	3b. Other
4 = Nonmatches from whole household nonmatches (where housing unit matched); not conflicting households	Non-Hispanic White	Other		
5 = Nonmatches from conflicting households; housing unit not in regular NRFU				
6 = Nonmatches from conflicting households; housing unit in regular NRFU (NRU = 3)				
7 = Nonmatches from whole household nonmatches; housing unit did not match during housing unit matching	Non-Hispanic White	other		
8 = People resolved before follow-up	Non-Hispanic White	other		
9 = People with insufficient information for matching				
10 = Targeted extended search people (TESPER=1) not in BFUGP 11 or 12				
11 = Potentially fictitious people				
12 = People who potentially lived elsewhere on Census Day				