ANALYSIS OF THE MISSING DATA ALTERNATIVES FOR THE 2000 A.C.E.[1]

Don Keathley, Tom Belin, William Bell, Anne Kearney, Rita Petroni
U.S. Bureau of the Census, PRED, Mail Stop 9200, Washington, D.C. 20233-9200

Key Words: missing data alternatives; dual system estimates; standard deviations; graphical and regression analysis.

1.0 BACKGROUND AND INTRODUCTION

1.1 Background

The Accuracy and Coverage Evaluation Survey (A.C.E.) relied on dual system estimation to estimate coverage in Census 2000. The A.C.E. computed dual system estimates (DSE) at the post-stratum level (post-strata were defined on race/ethnicity, tenure, MSA/TEA, Census 2000 questionnaire return rate, and age/sex group). Post-stratum level DSEs could then be added to form higher level estimates. See Griffin (2000) for details on dual system estimation.

As in most surveys, missing data in the A.C.E. resulted from non-interviews and item non-response. The A.C.E. had to account for these missing data to calculate the DSEs. It did this by implementing a set of missing data procedures. These data included:

A. Noninterviews for P-Sample[2] households

B. Interviews with some or all of the following:

   i. missing demographic characteristics (race, ethnicity, sex, age, tenure) for P-Sample persons - imputation for E-Sample wasn't necessary (see Cantwell (2001)).

   ii. unresolved match and resident status for P-Sample persons

   iii. unresolved enumeration status for E-Sample persons.

The A.C.E. production operations accounted for these missing data in various ways. It spread non-interviewed household weights over P-Sample interviewed households.

It used national distributions and hot decks to impute for the missing demographic characteristics. Finally, it used imputation cell procedures to impute missing resident, match, and enumeration status probabilities - these probabilities were the mean within-cell proportions of residents, matches, and correct enumerations, respectively, for persons with the appropriate resolved status. See Kearney, et. al. (2002) for summaries of the missing data procedures; see Cantwell (2001) for the details.

1.2 Introduction

We wanted to examine the effects alternative missing data procedures would have on the DSEs. Table 1. shows the procedures we used for this analysis. Note that we didn't include any demographic characteristic imputation alternatives. Our thinking was that the estimates of A.C.E. sampling variance would account for the variation associated with these imputations, with some minor adjustments to the variance procedures, if necessary (see Kearney, et. al. (2002)).

Every alternative procedure in Table 1. contains two levels. One level is using the alternative procedure (level=alt), the other level is using the procedure that was used in A.C.E. production (level=A.C.E.). So, we had 2^7 = 128 alternative procedure combinations (combinations).

Note that all 128 combinations were possible. For example, one combination set level=alt for late data and logistic regression and level=A.C.E. for the other five alternatives. Another combination set level=A.C.E. for the nearest neighbor non-interview adjustment alternative and level=alt for the remaining alternative procedures.

Sections 2.0 through 4.0 describe our analyses. Section 2.0 shows which alternatives and combinations of alternatives had significant effects on the DSEs. Section 3.0 explores the accuracy of two of the non-ignorable

[1] 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.

[2] See Kearney, et. al. (2002) for a brief description of the P- and E-Samples; see Childers (2001) for details.

missingness procedures. Section 4.0 presents the ranges of DSEs among all 128 combinations of alternatives as well as two sub-sets of combinations - these ranges gave us an indication of how sensitive the DSEs were to changes in one or more of the A.C.E. missing data procedures. Section 5.0 gives our conclusions based on these analyses.

Table 1. Alternative Missing Data Procedures

Alternative Procedure	Description [3]	Motivation for using the Alternative Procedure
Alternative Non-interview Adjustment (NIA) Cell Definitions	Use different NIA cells. These alternative cells were defined on type of basic address, race/ethnicity/tenure, census division, state within division, type of enumeration area, and household size	Household characteristics may be more homogeneous within these alternative NIA cells.
Nearest-Neighbor NIA Procedure	Add each non-interviewed household's (donor) weight to the nearest (in an NIA cell or collapsed NIA cell) interviewed household's (donee) weight. Each donee would receive no more than one donor's weight.	More homogeneity may result between donor and donee household characteristics when compared to spreading weights over many interviewed households.
Late Data	Assign non-interviewed household weights to late-arriving household interviews only; use the same late-arriving interview information in imputing for probabilities	Late-arriving interview data may more accurately reflect non-interviews and persons with unresolved match, resident, and enumeration status
Logistic Regression	Assign resident, match, and enumeration probabilities to unresolved cases using a logistic regression model	Logistic regression allows for the use of an extensive number of explanatory variables in predicting probabilities
Non-ignorable Missingness for Enumeration Probability	Lower the imputed enumeration, match, and resident probabilities for the corresponding unresolved cases.	Estimated enumeration, match, and resident rates using resolved-only cases may overstate the corresponding rates for unresolved cases.
Non-ignorable Missingness for Match Probability		
Non-ignorable Missingness for Resident Probability		

[3] See Kearney, et. al. (2002) for a more detailed summary; see Cantwell (2001) for explicit details.

2.0 Effects of the Alternatives on the DSEs

2.1 Methods

We used regression analysis and graphical examinations to assess the effects of the alternative procedures on the DSEs.

We ran regressions using DSE as the dependent variable and the alternative procedures as the

independent variables. We used factor effects (+1, -1) as the regressors.

For example, for the main effects, a +1 indicated that an alternative procedure was present in a given combination of alternative procedures, while a -1 indicated that it wasn't. All of the models contained an intercept term.

We ran both stepwise and full-model regressions. We started with main-effects models only and worked our way to four-way interaction models. In the end, we had the results from eight regression runs, four stepwise and four full-model.

A graphical analysis allowed us to compare the results of the regression output to a plot of DSEs.

2.2 Results

In both the stepwise and full-model regressions, the same five main effects and one two-way interaction (from the two-, three-, and four-way interaction models only) separated themselves from the pack. These six were the "primary" effects:

a. logistic regression
b. non-ignorable missingness (NIM) for correct enumeration probability
c. alternative non-interview adjustment (NIA) cell defintions
d. NIM for match probability
e. late data ⊗ logistic regression
f. NIM for resident probability

The p-values for all six primary effects were far below 0.0001 whereas the p-values for all other effects were above 0.0001. The R^2 for just these six effects was 0.8611 - the maximum R^2 from any of the other models was 0.9262 (where the model with the six primary effects had far larger adjusted R^2s than the other models).

There was some evidence of model bias, however: Mallow's C_p for this six-effect model was 28.4, while the number of parameters, including the intercept, was 7 (the expected value of Mallow's C_p is the number of parameters, including the intercept, in the model). We added some effects to the six-effect model so that Mallow's C_p more closely approximated the number of parameters in the model. So, we added five more effects, four two-way interactions (from the two-, three-, and four-way interaction models only) and a main effect (not significant in the main-effects only model). These five were the "secondary" effects:

g. logistic regression ⊗ nearest neighbor NIA
h. late data ⊗ NIM for correct enumeration prob.
i. NIM for match prob. ⊗ NIM for resident prob.
j. logistic regression ⊗ alternative NIA cell defin's
k. late data

The model containing both the primary and secondary effects resulted in a Mallows $C_p = 10.5$; the number of parameters in this model equaled 12. Thus, this model showed minimal evidence of model bias. The p-values for the five secondary effects were between 0.0001 and 0.05; the p-values for the primary effects were still far below 0.0001. The R^2 for the model using all eleven effects was 0.8883.

A graphical analysis emphasized the influence of the primary effects on the DSEs; it emphasized to a lesser extent the influence of the secondary effects. Note that this graph contained too many points to render itself legible in black-in-white on 8.5 x 11 paper. Please contact the author(s) for a copy of this graph.

Some additional analysis on the graph revealed the following:

- the highest DSEs were associated with combinations that included the NIM for match probability
- the lowest DSEs were associated with combinations that included the late data ⊗ logistic regression interaction
- of the eight NIM combinations (using none, one, two, all three), the lowest DSEs occured when we used the NIMs for correct enumeration and resident probabilities

3.0 Accuracy of the Alternatives

The only alternatives we were able to evaluate were the non-ignorable missingness alternatives for enumeration and resident status. We used data from the Measurement Error Reinterview (MER; see Krejsa (2001) for details) for this assessment. See Kearney et. al. (2002) for details on the implementation of these two procedures.

The MER was a sub-sample of the Evaluation Follow-Up (Keathley (2001)), which in turn was a 1-in-5 sample of the A.C.E. (so, the MER is a subsample of a subsample of the A.C.E.) The MER was able to resolve enumeration, match, and resident statuses for persons with the corresponding unresolved A.C.E. statuses. There were too few persons whose match status changed from unresolved in the A.C.E. to resolved in the MER, however, for any meaningful analysis. So, we could examine enumeration and resident status, only.

For our assessment, we compared the average A.C.E. imputed correct enumeration and resident probabilities versus the corresponding MER resolved rates for the above persons. The non-ignorable missingness procedures systematically decreased the imputed A.C.E. correct enumeration and resident probabilities. So, the observed MER correct enumeration and resident rates for these persons should be lower than their average imputed A.C.E. probabilities.

Table 2. shows the average imputed A.C.E. probabilities and the corresponding resolved MER rates. These data do not support the notion that these alternatives accurately predict correct enumeration and resident rates for persons with the corresponding unresolved A.C.E. statuses. For enumeration status, there was a decrease from the average imputed probability (0.767) to the resolved rate (0.754). This decrease was not large enough, however, to firmly support the NIM for correct enumeration as an accurate predictive method. For resident status, not only was the observed resident rate (0.828) an increase over the average imputed probability (0.704), this increase shows some evidence of there being non-ignorable missingness in the opposite direction.

Table 2. Conversion Results

Imputed/ Resolved Status	Sample Size (MER conve-rsion rate) [4]	Avg. Imp. Prob. (A.C.E.)	Resolved Proportions (MER)
Correct Enumer.	3,475 (0.517)	0.767	0.754
Resident	1,309 (0.263)	0.704	0.828

[4] The rates in parentheses are the percent of MER cases with unresolved enumeration and resident status that the MER resolved.

We weren't able to turn this lack of support into a firm conclusion, however. First, only a fraction of all cases with an imputed A.C.E. probability were resolved in the MER. The conversion rates within the MER itself were only 51.7% and 26.3% for enumeration and resident status, respectively. Second, we were unable to make comparisons for match status.

4.0 DSE Ranges

We looked at the ranges of DSEs for the following three sets of combinations of procedures:

1 the 16 combinations that contained none of the NIM procedures
2. the 16 combinations that contained all three of the NIM procedures
3. all 128 procedures.

We looked at the ranges in set 1because of the results from section 3.0 - we worked under the assumption that the NIM procedures were not as reflective of the true nature of the applicable missing data as the other four alternative procedures. For set 2, we anticipated that when all three of the NIM procedures were used together, their effects would tend to cancel each other out.

Thus, we expected the smallest range of DSEs to be in set 1, the largest range in set 3, and something in between for set 2. The fact that we used the NIM procedures made us think that set 2 would exhibit a larger range of DSEs than set 1 - even with the anticipated cancelling-out effect of using all three NIM procedures.

Table 3 shows the ranges and the standard deviations for the these three sets of combinations.

Table 3. DSE Ranges

Set	Range of DSEs	Standard Deviation
1. no NIM procedures	1,266,317	384,115
2. all 3 NIM procedures	1,300,959	402,284
3. all 128 combinations	2,628,487	531,569
A.C.E. sampling error		378,222

Table 3. shows that sets 1 and 3 did have the smallest and largest ranges of DSEs. The magnitude of these two ranges wasn't surprising either, as using one or two NIM procedures by themselves produced the largest

and smallest DSEs (section 2.0). The range for set 2 was a bit smaller than we expected, not significantly different for the set 1 range.

The standard deviations for sets 1 and 2 for the DSEs compare favorably with the A.C.E. sampling error. The standard deviation for set 3 was expectedly high compared to the A.C.E. sampling error.

We did some additional analysis on the ranges in sets 1 and 2. The line graph in Figure 1 shows the DSEs for both sets. DSEs are on the y-axis. The 16 non-NIM procedure combinations are on the x-axis, where

AC = alternative NIA cell definitions
NN = nearest neighbor NIA
LR = logistic regression
LD = late data

The top line shows the DSEs for set 1, the bottom line for set 2. The horizontal line shows the actual A.C.E. DSE.

For example, the left-most entry on the x-axis is AC. The set 1 DSE in the graph for AC represents the DSE for the combination that used alternative NIA cell defintions only; the set 2 DSE for AC represents the DSE for the combination that used alternative NIA cell definitions and all three of the NIM alternative procedures. Similarly, the set 1 DSE for the (NN LR) x-axis entry shows the DSE for the combination that used the nearest neighbor NIA and logistic regression procedures only.

The graph shows a consistent downward shift in DSEs from set 1 to set 2. Because of the results in section 3.0, we might assume that using all three NIM procedures produces a downward bias in DSEs.

5.0 Conclusions

Considerable non-sampling variability arose from the use of these alternative missing data procedures.

Section 2.0 shows that five of the seven alternative procedures (as main effects) and the logistic regression $\otimes$ late data interaction had highly significant effects on the DSEs. The section also mentions that the late data alternative procedure, as a main effect, plus four other interactions, had moderately significant effects on the DSEs. As a main effect, only the nearest neighbor NIA showed no signs of significantly influencing the DSEs.

Section 4.0 depicts the standard deviations for the three sets of combinations of alternatives. At best, for set 1, whose combinations excluded the NIM alternatives, the standard deviations was comparable, but not smaller, than the A.C.E. sampling error (see Table 3.).

In section 3.0, there was some evidence that the NIM alternative procedures for correct enumeration and resident probability did not accurately predict the true correct enumeration and resident probabilities, respectively, for cases who had missing enumeration and/or resident status. These results were based on only a subsample of cases, however - an analysis on all A.C.E. cases with missing enumeration, resident, and match status, had they been available, might have resulted in a different conclusion.

6.0 References

Cantwell, P. J. (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

Childers, D.R. (2001), “Chapter S-DT-1, Revised, Accuracy and Coverage Evaluation: The Design Document” DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1

Griffin, R. (2000), “Accuracy and Coverage Evaluation Survey: Dual System Estimation” DSSD Census 2000 Procedures and Operations Memorandum Series Q-20.

Kearney, A.T., Belin, T.R., Keathley, D.H.,Petroni, R.J. (2002) “ Alternatives of the A.C.E.Missing Data Evaluation” Proceedings of the American Statistical Association, Section on Survey Research Methods (forthcoming)

Keathley, D. (2001) “EFU Sample Design, Stratification, Selection, and Weighting” Planning, Research, and Evaluation Division TXE/2010 Memorandum Series: CM-GES-S-02-R2

Krejsa, E. (2001) “Measurement Error Reinterview Sample Selection” Planning, Research, and Evaluation Division TXE/2010 Memorandum Series: CM-MER-S-01

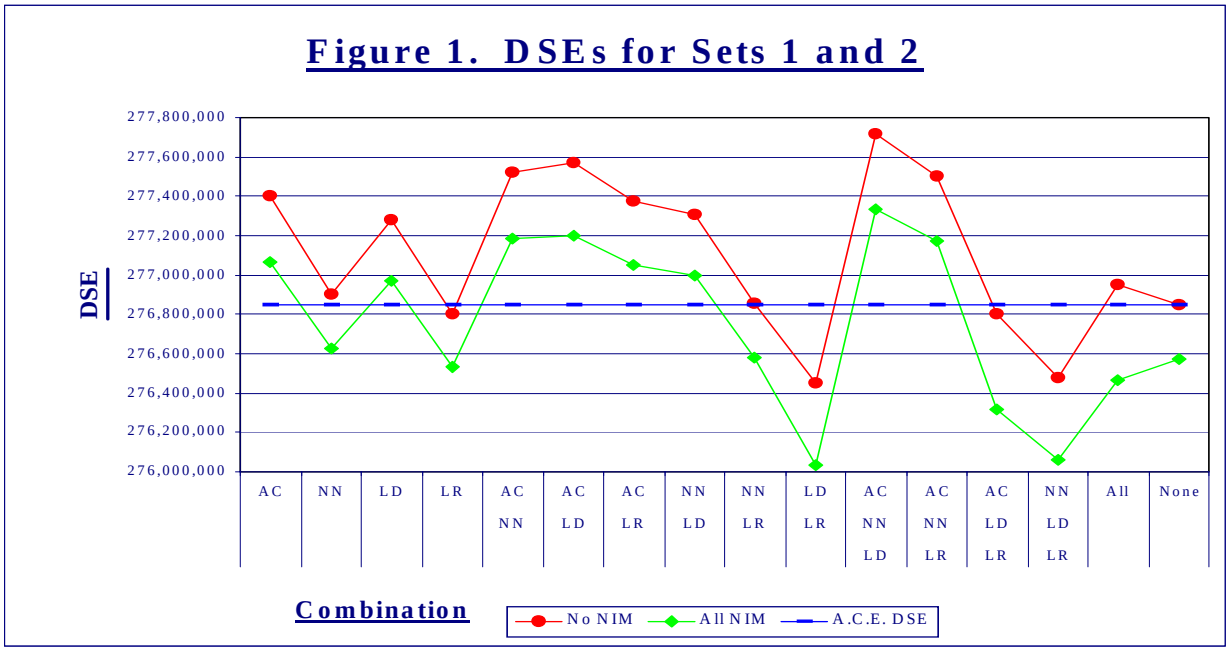
Figure 1. DSEs for Sets 1 and 2
DSE
277,800,000
277,600,000
277,400,000
277,200,000
277,000,000
276,800,000
276,600,000
276,400,000
276,200,000
276,000,000
AC
NN
LD
LR
AC NN
AC LD
AC LR
NN LD
NN LR
LD LR
AC NN LD
AC NN LR
AC LD LR
NN LD LR
All
None
Combination
No NIM
All NIM
A.C.E. DSE