The Effect of the Method of Imputation for Missing Probabilities of Match, Residence, and Correct Enumeration on Dual System Estimates

Maria Cupples Hudson and Kara Morgan Clarke
Maria Cupples Hudson, US Census Bureau, Washington, DC, 20233-7600

Key Words: Missing Data, Logistic Regression, Ratio Estimation, Census

1. Background

After the 1990 Census, the Census Bureau conducted a coverage improvement program called the Post-Enumeration Survey (PES). The Bureau selected and re-interviewed a nationwide sample of block clusters. Persons enumerated in those PES blocks during the census are called E-Sample persons. Persons interviewed in the PES are called P-Sample persons. Clerks matched the samples to each other to determine who the census missed or counted in error. Using Dual Systems Estimation (DSE), the Bureau derived a coverage factor for each of 357 post-strata.

Because several months passed between the census enumeration and the PES interview, the Bureau had to deal with movers, people who moved between those two dates. We call persons who moved between these two dates inmovers or outmovers. We call persons who did not move nonmovers.

Each person in the census enumeration has a probability of correct enumeration in the census. If an E-Sample nonmover matched to a P-Sample person, then we considered that person correctly enumerated and set the correct enumeration probability to one. We followed up unmatched persons to determine if they were erroneously enumerated. If so, we set the correct enumeration probability to zero. Similarly, we assigned a match probability of one to each person in the P Sample who matched an E-Sample person and of zero to those who did not match an E-Sample person. However, for persons with an unresolved probability of match or correct enumeration we had to assign a value between zero and one. This paper focuses on what effect, if any, the exact method of assigning these probabilities has on the Dual Systems Estimates.

2. Methodology

We ran three imputation routines for missing probabilities of match and correct enumeration in the 1990 PES data. The first routine was a simple imputation cell estimation (ICE) program, the second was logistic regression, and the third a more complex ICE routine. We measured the change resulting from the different imputation methods by the percent differences in the Dual System Estimates.

$$100 * \frac{DSE_{LRM} - DSE_{ICE}}{DSE_{LRM}}$$

Where DSE_{LRM} were the DSEs derived from using LRM and DSE_{ICE} were the DSEs derived from using ICE. We chose this measure because if a change would be made in the census count, it would show up in the DSEs. We examined these percent differences to find any systematic differences between the methods.

Next we performed a cross-validation analysis to compare the differences between the methods and the "actual" outcome to see which method came closer to the truth. A caveat to this type of analysis is that it could only say how well the methods work for resolved cases. If unresolved cases were inherently different from the resolved cases within a particular cell then it could not tell us how well the methods worked on unresolved cases.

2.1 Accounting for Changes Since 1990

Due to differences in procedures between 1990 and the present, we made some accommodations with the data. First, in 1990 logistic regression was also used to calculate a probability of correct geocoding for the E Sample and a probability of fictitious inclusion in the P Sample. Unable to duplicate the probabilities, we chose to mimic the 1998 procedures for assigning final enumeration status based on a person's match codes. We proceeded in the following order.

- We gave all persons with a final match code of "insufficient information for matching" a final enumeration status of zero.
- We gave all persons erroneously geocoded a final enumeration status of zero.
- We gave all persons with unresolved geocoding status a final enumeration status of unresolved.
- We gave all persons with match codes M, CE, and C1 (match/correct enumeration/college dorm) a final enumeration status of 1/(1+k) where k is the number of times the person was duplicated in the E Sample.
- We gave all persons with match codes EE, EF, and DE (erroneous/fictitious/duplicate) a final enumeration status of zero.
- We gave all other persons a final enumeration status of unresolved.

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.

Using the above criteria, the total number of persons with unresolved enumeration status was 5,825 (1.5%).

In the 1990 P Sample, logistic regression was used to give each person a probability of fictitious inclusion in the P Sample. This is not analogous to the residence probability to be used in Census 2000 missing data processing because it only looks at fictitious persons, not real persons who were not residents on Census Day (for example college students who were in dormitories on April 1, 1990). Again, this probability will be incorporated into the match codes for Census 2000. Thus we created a match status variable using final match codes as follows:

- We gave all persons with match codes M and Q3 (match/ match in surrounding block) a match status of one.
- We gave all persons with match codes N1, N2, N3, N4 (nonmatches), L (rejected form),Y3 (erroneous), NG (not located), PF (fictitious), DP (duplicate), RP (removed), S1, S2, and SP (group quarters, Puerto Rico, or Alaska) a match status of zero.
- We gave all others a match status of unresolved.

The number of persons with unresolved match status was 6,492 (1.7%).

For our first imputation cell estimation run, we needed to divide the data into mutually exclusive and exhaustive cells. Since we intended to mimic the Census 2000 Dress Rehearsal procedure, we used the before-followup group (BFUGP) variable for the E Sample and mover status (MOVER) for the P Sample. We combined groups 5 and 6 into one group due to data constraints. The formation of the before-followup groups is given in Figure 1 below. Mover status simply indicates if a person was a mover or nonmover.

Figure 1. Definition of Before Followup Groups

BFU group	Description
1	Matches needing followup
2	Possible matches
3	Nonmatches from partial household matches
4	Nonmatches from whole household nonmatches where the housing unit matched
5	Nonmatches from conflicting households where the housing unit was not in regular nonresponse followup
6	Nonmatches from conflicting households where the housing unit was in regular nonresponse followup
7	Nonmatches from whole household nonmatches where the housing unit did not match
8	Resolved before followup
9	Insufficient information for matching

It should be noted that though we used the same data, we did not expect to obtain the same DSEs as the 1990 production. The operational and procedural changes made in our analysis made the resulting DSEs incomparable. Our goal was to compare the effects of Imputation Cell Estimation and Logistic Regression Modeling on the DSEs rather than recreate the 1990 results.

2.2 Logistic Regression

We ran logistic regression using the software package SUDAAN, which takes complex survey design into account when computing variance estimates. We chose Taylor series linearization as the variance estimation option and eliminated effects one at a time by removing the one with the largest p-value for the Wald F statistic.[1]

We ran logistic regression on the P and E sample data separately. Because the memory space available would not allow for any two-way interactions to be run on the entire data sets, we chose samples of the PES data. We sorted the data by state and then chose a systematic 20% sample. We also collapsed effects to get a minimal number of levels that still allowed us to include all two-way interactions. We show the variables, along with their levels, in Figure 2.

Figure 2. Definition of Variables for Logistic Regression

Variable	Definition	Levels
BFUGP2 (E Sample only)	Before-followup group (collapsed)	1 = BFUGP 1, 2 2 = BFUGP 3 3 = BFUGP 4, 5, 6 4 = BFUGP 7 5 = BFUGP 8
MOVERS (P Sample only)	Indicator for mover status	1=nonmovers 2=movers
TENURE	Indicator for tenure	1=owners 2=nonowners
RACE	Collapsed race	1=nonhispanic white 2=minority
REL	Collapsed relation-ship to house-holder	1=other 2=self
IMPUTE	Indicator for characteristic imputation	1=no imputes 2=at least one characteristic imputed
AGE	Collapsed age	1=less than 18 2=18 to 49 3=50 and older
REGION	Geographic region of the United States	1=NorthEast 2=MidWest 3=South 4=West

For the P sample, we chose a sample of 74,069 from 377,005 observations. The original model contained variables for the following effects and all possible two-way interactions: MOVERS, TENURE, RACE, REL, IMPUTE, AGE, and REGION. We removed effects using backward elimination leaving the following significant two-way interactions: REL*MOVERS, IMPUTE*MOVERS, IMPUTE*TENURE, REGION*TENURE, REGION*RACE. AGE did not have any significant two-way interactions, but was significant as an effect. We show the Wald statistics and p-values in Figure 3.

Figure 3. Indicators of Significance for Logistic Regression Variables for the P Sample[1]

Effect	Wald F Statistic	P-Value
AGE	14.91	0.0000
REL*MOVERS	4.38	0.0363
IMPUTE*MOVERS	9.04	0.0027
IMPUTE*TENURE	3.50	0.0613
REGION*TENURE	4.06	0.0069
REGION*RACE	3.26	0.0206

For the E sample, we chose a sample of 76,321 from 381,462 observations. We first modeled Correct Enumeration using the following variables with all two-way interactions: BFUGP2, TENURE, RACE, REL, IMPUTE, AGE, and REGION. After backward elimination of the insignificant interactions, the following significant interactions remained: TENURE*BFUGP2, RACE*BFUGP2, REL*BFUGP2, IMPUTE*BFUGP2, AGE*BFUGP2, AGE*REL, and AGE*IMPUTE. REGION did not have any significant two-way interactions, but was significant by itself. We show the Wald statistic and p-values in Figure 4.

Figure 4. Indicators of Significance for Logistic Regression Variables for the E Sample[1]

Effect	Wald F Statistic	P-Value
REGION	3.43	0.0164
TENURE*BFUGP2	4.34	0.0017
RACE*BFUGP2	3.36	0.0094
REL*BFUGP2	6.31	0.0000
IMPUTE*BFUGP2	5.34	0.0003
AGE*BFUGP2	4.99	0.0000
AGE*IMPUTE	4.97	0.0070
AGE*REL	2.97	0.0513

We calculated match and correct enumeration probabilities for all people with unresolved final enumeration or match status using the beta coefficients given by SUDAAN. First, we created indicator variables for each level of each variable. Then we calculated the probabilities by inserting the indicator variables and the beta coefficients into the final logistic regression model. We then used the probabilities obtained in the program that calculates DSEs and compared them to the DSEs from ICE.

2.3 Imputation Cell Estimation

We first created an imputation cell estimation program to mimic the Census 2000 Dress Rehearsal procedure. We divided the E-Sample into mutually exclusive and exhaustive cells. We assigned people with unresolved enumeration status the correct enumeration probability was the weighted proportion of correct enumerations (among persons with resolved enumeration status) in that same cell. Thus, person j's probability of correct enumeration was:

$$\Pr_{ce,j} = \begin{cases} 1 \textit{ if person } j \textit{ is correctly enumerated} \\ 0 \textit{ if person } j \textit{ is NOT correctly enumerated} \\ \Pr^*_{ce,j} \textit{ if person is unresolved} \end{cases}$$

Where for each imputation cell, the estimated probability of correct enumeration was:

$$\Pr{}^*_{ce,j} = \frac{\sum_{\textit{resolved units}} w_i \Pr_{ce,i}}{\sum_{\textit{resolved units}} w_i}$$

Where w_i is the weight of person i and $Pr_{ce,i}$ is the correct enumeration probability for person i.

Similarly for the P Sample, we calculated the weighted average match probability for each cell. We then gave each unresolved person the weighted average for that cell. Thus person j's probability of match was:

$$\Pr_{m,j} = \begin{cases} 1 \text{ if person j is a match on Census Day} \\ 0 \text{ if person j is NOT a match on Census Day} \\ \Pr^*_{m,j} \text{ if person j is unresolved} \end{cases}$$

Where for each imputation cell, the estimated match probability is:

$$Pr^*_{m,j} = \frac{\sum_{\textit{resolved units}} w_i \Pr_{m,i}}{\sum_{\textit{resolved units}} w_i}$$

Where w_i is the weight of person i and $Pr_{m,i}$ is the match probability for person i.

We created two ICE programs. The one we call ICE1 is the most similar to the Census 2000 Dress Rehearsal and is the most efficient procedure as far as programming and computing time is concerned. We ran ICE1 on both the P and E samples using only one variable to determine the cells. We defined the E-Sample cells by the before-followup group and the P-Sample cells by mover status. The second ICE program, ICE2, adds two more variables in defining the cells. We defined the E-Sample cells by BFUGP2, AGE, and REL. The P-Sample cells were defined by MOVERS, AGE, and IMPUTE. These effects were chosen because they were the three most significant in the logistic regression model, either alone or in a two-way interaction and because we wanted to maintain a cell size of at least 100 resolved persons in all cells. Additional variables would have caused the cells to be too small, thus increasing variability and decreasing reliability.

3. Comparisons of ICE and LRM

3.1 Comparing Percent Differences in the DSEs

We now consider the percent differences in the DSEs between these two methodologies. We looked at the average percent difference, the standard deviation of the percent difference, the upper and lower percentiles, and histograms for analysis. We show these analyses in the next two figures.

Figure 5. Comparing Logistic Regression Modeling to Imputation Cell Estimation.

	ICE1	ICE2
Average Percent Difference between ICE and LRM	0.0680	0.0747
Standard Deviation of Percent Differences	0.3306	0.3100
Number of Post-Strata with More than 1% Difference	2 out of 357	9 out of 357

Figure 6. Percentiles of the Percent Differences Between DSEs obtained by LRM and ICE

	ICE1	ICE2
5th Percentile	-0.2655	-0.2510
25th Percentile	-0.0696	-0.0578
75th Percentile	0.1014	0.1304
95th Percentile	0.7620	0.6919

In comparing the two methods, ICE and logistic regression, we found very little difference on the resulting dual system estimates for the 357 post-strata used in 1990. With both of the averages being less than one percent and less than ten of the post-strata having greater than a one percent difference, we concluded that the change in methodology had little affect on the DSEs. The percent differences were generally scattered about zero and did not appear to be systematic as can be seen in the following two figures.

Figure 7. Histogram of Percent Differences Between DSEs Obtained Using ICE1 and Logistic Regression Modeling

```
 1.75.*
     .
     .*
     .*
     .*
     .*
     .*
     .*
     .*
     .*
     .**
     .*
 0.55.**
     .***
     .***
     .*****
     .***********
 0.05.****************************
-0.05.*****************************************
     .************
     .****
     .***
     .*
     .**
-0.65.*
      ----+----+----+----+----+----+----+----+-
      * may represent up to 3 counts
```

Figure 8. Histogram of Percent Differences Between DSEs Obtained Using ICE2 and Logistic Regression Modeling

```
 1.85.*
     .
     .*
     .*
     .
     .*
     .*
     .*
     .*
     .**
     .*
     .*
 0.65.**
     .**
     .***
     .**
     .********
     .*************
 0.05.*****************************
-0.05.*************************************
     .***********
     .****
     .***
     .**
-0.55.*
     ----+----+----+----+----+----+----+--
      * may represent up to 3 counts
```

3.2 Comparison of Total Population Estimates

In addition to comparing the DSEs by post-strata, we compared the total population DSEs, total number of matches, and total number of correct enumerations resulting from each method. The rounded national estimates and the raw differences are given in Figures 9 and 10 below.

Figure 9. Rounded National Totals

	ICE1	ICE2	LRM
Matches	223,428,291	223,406,197	223,365,605
Correct Enumerations	233,944,094	233,893,236	233,828,550
US Population	250,355,802	250,327,109	250,351,983

Figure 10. Raw Differences in Rounded Totals

	ICE1	ICE2
Matches	-62,686	-40,592
Correct Enumerations	-115,544	-64,686
US Population	-3,819	24,874

Figure 11. Percentage Differences in Number of Matches and Correct Enumerations and Total Population.

	ICE1	ICE2
Matches	-0.028	-0.018
Correct Enumerations	-0.049	-0.028
US Population	-0.002	0.010

As can be seen in Figure 11, the difference in the total number of matches and correct enumerations was very small at the national level. Logistic regression tended to produce smaller estimates of the total number of matches and correct enumerations, although the estimate of the total population was between the estimates obtained using ICE1 and ICE2.

3.3 Cross-Validation Comparison

We also compared the effects of ICE2 and Logistic Regression Modeling on dual system estimates using cross-validation. We applied the Logistic Regression Model, which we developed using a 20% sample of the entire PES data, to another 20% sample. Initially we drew 5 mutually exclusive samples from the PES data. The one used to develop the Logistic Regression Model we called Sample 1 and the one used for cross-validation we called Sample 2.

We used Sample 1 to estimate the probabilities using both ICE and logistic regression modeling (as described in section 2). To compare the results of ICE2 and LRM using Sample 2, we eliminated all unresolved cases and imputed probabilities for the resolved cases. Next, we calculated DSEs based on the results from ICE2, logistic regression, and the actual outcome for these resolved cases. Finally, we calculated the absolute relative difference between the DSEs resulting from the actual outcome and those from the imputed probabilities. We did so to see how far each method was from the actual results, regardless of direction. We compared the methods by looking at the difference, referred to as DIFF, between these absolute relative differences to see which method came closer to predicting the DSEs obtained using the actual outcome. The formula is given by

$$DIFF = \left|\frac{DSE_A - DSE_{ICE}}{DSE_A}\right| - \left|\frac{DSE_A - DSE_{LRM}}{DSE_A}\right|$$

Where DSE_A is the DSE using the final enumeration and match status assigned as in section 2.1, DSE_{ICE} is the DSE using the final enumeration and match status assigned by ICE2, and DSE_{LRM} is the DSE using the final enumeration and match status assigned by LRM. Figure 12 below gives the statistics that we looked at.

Figure 12. Indications of Cross-Validation Results

Average Difference between Absolute Relative Differences	0.0063
Standard Deviation of Differences	0.0635
5th Percentile	-0.0912
25th Percentile	-0.0148
75th Percentile	0.0309
95th Percentile	0.1968

The results show that the DSEs computed using the LRM and ICE2 methodology have, on average, very close absolute relative differences from the DSEs computed using the actual data for match rate and correct enumeration. While the LRM method is closer on average, the average difference between absolute relative differences is very small, as illustrated by figure 13 below.

Figure 13. Histogram of DIFF

```
 0.225.*
      .**
      .**
      .**********
 0.025.*******************************************
-0.025.*************************
      .******
      .**
-0.175.*
      .*
      .*
      .
      .
      .
      .
      .
-0.575.*
    ----+----+----+----+----+----+----+----+---
    * may represent up to 4 counts
```

4. Summary of Comparisons

The comparisons of the DSEs for ICE1, ICE2, and LRM show that the resulting DSEs are similar between methods. The mean differences are below 1% and less than 10 of the 357 post-strata have more than a 1% difference in DSEs for each comparison.

The cross-validation analysis shows that although the LRM methodology tends to give DSEs slightly closer to the actual DSEs, the difference is very small. The mean difference between absolute relative differences of the two methods is only 0.0063. Again, these results may not be directly applicable to unresolved cases if they differ statistically from the resolved cases used in the analysis for the given cells.

References:

T. Belin, G. Diffendal, S. Mack, D. Rubin, J. Schafer, and A. Zaslavky (1993). "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," Journal of the American Statistical Association, 88, No. 423, 1149-1159.

G. Diffendal and T. Belin. "Documentation of the Handling of Unresolved Enumeration Status in 1990 Census/Post-Enumeration Survey", STSD Decennial Census Memorandum Series #V-98. Internal Census Bureau Memorandum, January 15, 1991.

S. Dorinski, R. Petroni, M. Ikeda, and R. Singh. "Comparison and Evaluation of Alternative ICM Imputation Methods", 1996 Proceedings of the Section on Survey Research Methods, American Statistical Association, 299-304.

H. Hogan (1992), "The 1990 Post-Enumeration Survey: An Overview", The American Statistician, Vol. 46, No. 4, 261 – 268.

H. Hogan (1993). "The 1990 Post-Enumeration Survey: Operations and Results", Journal of the American Statistical Association, 88, No. 423, 1047–1059.

M. Ikeda, A. Kearney, and R. Petroni. "Missing Data Procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample", 1999 Proceedings of the Section on Survey Research Methods, American Statistical Association,

M. Ikeda. "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. Internal Census Bureau memorandum, December 11, 1997.

M. Ikeda. "Effect of Different Methods for Calculating Match and Residence Probabilities for the 1995 P-Sample Data", DSSD 2000 Census Dress Rehearsal Memorandum Series #A-23. Internal Census Bureau memorandum, January 5, 1998.

M. Ikeda. "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. Internal Census Bureau memorandum, December 11, 1997.

M. Ikeda. "Handling of Missing Data in the 1996 Integrated Coverage Measurement Sample", DSSD Census 2000 Dress Rehearsal Memorandum Series #A-26. Internal Census Bureau memorandum, January 5, 1998.

M. Ikeda. "Effect of Different Methods for Calculating Correct Enumeration Probabilities for the 1995 E-Sample Data", DSSD Census 2000 Dress Rehearsal Memorandum Series #A-28. Internal Census Bureau memorandum, January 5, 1998.

M. Ikeda. "Effect of Using Simple Ratio Methods to Calculate P-Sample Residence Probabilities and E-Sample Correct Enumeration Probabilities for the 1995 Data", DSSD Census 2000 Dress Rehearsal Memorandum Series #A-30. Internal Census Bureau memorandum, January 28, 1998.

[1] A table of variables with associated p-values at the time of removal can be obtained from the authors