AN ANALYSIS OF THE EFFECT OF REASONABLE IMPUTATION ALERNATIVES ON ESTIMATES OF COVERAGE ERROR IN THE 1990 DECENNIAL CENSUS

Stephen Mack[1]
Bureau of the Census, Washington D.C. 20233

KEY WORDS: Census, Post Enumeration Survey, Coverage Error, Imputation

1. INTRODUCTION

The U.S. Census Bureau has produced estimates of census coverage error for each decennial census since 1950. These estimates have been helpful in efforts to improve coverage in subsequent censuses and will form the basis of any adjustment of the 1990 Decennial Census that may occur. One of the primary methods used to estimate coverage error is the post-enumeration survey (PES). Three months after the 1990 census, the Census Bureau conducted a post-enumeration survey of approximately 5,400 block clusters. A block cluster may contain part of a block, a whole block, or several blocks. Hogan (1990) gives an overview of many aspects of the PES. Two samples of people were defined by the sampled block clusters. The E-Sample, or enumeration sample, were people enumerated during the census as living in the sample block. The P-Sample, or population sample, consisted of people determined to be living in the sample block during the PES. The PES attempts to determine the match status of each P-Sample person, i.e.:

Was this person enumerated during the census?

The PES also tries to ascertain the enumeration status of each E-Sample enumeration, i.e.:

Was the census enumeration correct?

Even though only three months had elapsed, almost 8% of the P-Sample were movers (people moving into a sample block after the census). Likewise, many E-Sample enumerations were of people who had moved away after the census and thus were not in the P-Sample.

Missing data in the P and E Samples was imputed by three methods. Noninterviewed households were taken care of by a standard weighting adjustment procedure. Missing characteristics of P and E sample persons were imputed by hot deck. A logistic regression approach was taken to impute missing match and enumeration statuses.

Dual system estimates (see Wolter 1986) were used in the PES to estimate total population for a given poststratum. The PES poststrata, 1392 in all, were based on race, age, sex, tenure, and geography.

In order to assess the accuracy of the coverage error estimates resulting from the 1990 PES, the Census Bureau conducted eighteen evaluation studies. This was one of three studies to assess the effects of missing data on PES estimates of census undercount.

This study focused on the missing data most directly affecting the PES undercount estimates: P-sample match status and E-sample enumeration status. Reasonable alternative methods of imputing for missing match and enumeration status were explored to determine the overall effect on the dual system estimates.

Some alternative treatments were suggested by possible difficulties arising during data collection, i.e. data from proxies and difficulties in determining match status for movers. Other treatments, i.e. the bootstraps, were motivated by the need to measure uncertainty in the dual system estimates due to the variance of the parameter estimates in the production imputation model. Another treatment replaces the production imputation model with an alternative model.

Unsmoothed estimates of census undercount were obtained for a number of reasonable alternative methods of imputing for missing match and enumeration status. The range of alternative estimates indicates the sensitivity of the dual system estimates to the treatment of missing data.

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

For example, a narrow range implies that the estimates are robust, and the missing data cause little uncertainty in the estimates.

2. IMPUTATION ALTERNATIVES

2.1 Production

The 1990 PES used a logistic regression strategy to impute for unresolved P-Sample matches and E-Sample enumerations. A description of the production imputation model is given in Belin, et al. (1991). The number of unresolved matches and enumerations is given in table 1. Given the observed percentage of unresolved cases and assuming equal weights, the range of unsmoothed national undercount estimates is only about 3% over all possible imputation schemes. The numbers in table 1 also hold for the other imputation treatments considered in this study except for proxies.

Table 1 Production Imputation Final Match Status.

Final Match /Enumeration Status	P-Sample	E-Sample
Match/Correct Enumeration	338597 89.7%	366743 93.4%
Nonmatch /Erroneous Enumeration	31628 8.4%	20485 5.2%
Unresolved	7156 1.9%	5359 1.4%
Total	377381 100%	392587 100%

2.2 P-Sample Followup Proxy Alternative

Information obtained from nonhousehold members, i.e. proxies, may be less reliable than information obtained from household members. P-Sample households may have had initial and/or followup interviews that were completed by proxy. The P-Sample Followup Proxy Alternative focuses on the followup interview. P-Sample followup interviews completed by proxies were disregarded in this alternative. The final match and mover status codes of such interviews were recoded to codes consistent with a noninterview during followup. Table 2 gives a breakdown of the changes in match status resulting from the recode. The number of unresolved matches rose from 7156 in production to 11486 after the recode. This number represents 3% of the total.

Table 2 Final Match Status of P-Sample Followup Proxies in Production Imputation vs P-Sample Proxy Alternative.

Production Match Status	P-Sample Proxy Match Status Matched	 Not Matched	 Not Resolved	Total
Matched	54	11	753	818
Not Matched	0	15	3578	3593
Unresolved	0	1	1056	1057
Total	54	27	5387	5580

2.3 E-Sample Followup Proxy Alternative

The number of E-Sample followup interviews that were completed by proxy was higher than for P-Sample followup. The higher rate may be due to outmovers who presumably were not available during followup. As in the P-Sample, the amount of missing and erroneous information may be higher for proxy interviews than for interviews with household members. The procedure for the E-Sample followup proxy alternative is to disregard followup interviews completed by proxies. The final match code for such persons was recoded to indicate that no interview was obtained during followup. Table 3 shows the resulting change in enumeration status from production. The number of unresolved cases rose by 10450 persons. The percent of unresolved enumerations rose to four percent for this alternative.

Table 3 Final Enumeration Status of E-Sample Followup Proxies in Production Imputation vs E-Sample Proxy Alternative

Production Enumeration Status	E-Sample Followup Proxy Enumeration Status Correct	 Incorrect	 Unresolved
Correct	0	0	8983
Incorrect	0	0	1467
Unresolved	0	0	832
Total	0	0	11282

2.4 Movers Alternative

Movers account for over 50% of unresolved P-Sample matches even though they represent only 8% of the sample. The higher rate of unresolved matches for movers may be a result of uncertainty about the correct mover address. The movers alternative imputes unresolved movers as if they were nonmovers with insufficient information to attempt matching. The original mover status was retained for dual system estimation. This alternative was regarded from the beginning as extreme and was selected primarily to observe its effect on the dual system estimates. The average match probability for unresolved movers was expected to rise considerably, however the effect on the dual system estimates might be negligible if the number of unresolved matches was small. Figure 1 compares the movers alternative with production. Note that unresolved movers were matched at about the same rate as resolved nonmovers under the movers alternative.

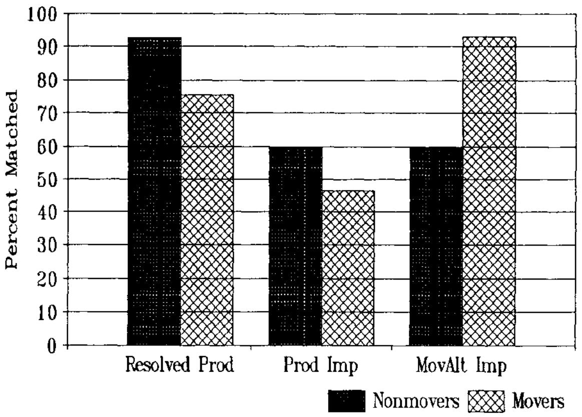

Figure 1 Comparison of resolved cases with production and mover alternative imputations.

2.5 1988 Style Logistic Regression Alternative

The 1990 PES imputation model is quite different than the model that was used in the 1988 Dress Rehearsal. The 1988 Style Logistic Regression Alternative implements an imputation model that is more similar to the 1988 Dress Rehearsal model. A number of standard logistic regression models are used to impute for match/enumeration status as was the case in the dress rehearsal. Other details of the models, such as coding of variables, etc., differ. A summary of the 1988 Style imputation models is given below:

P-Sample Model

The P-Sample imputation model is composed of four logistic regression models. Models 1,2, and 3 are used to impute the probability of match given mover status (mover or nonmover). If mover status is unresolved, the probability of being a nonmover is estimated by model 5.

Model 1: Imputes probability of match for most unresolved cases.
Model 2: Imputes probability of match for persons with insufficient information to attempt matching and for persons with unexpected match codes.
Model 3: Imputes probability of match for unresolved movers for whom a census questionnaire was unavailable at the mover address.
Model 5: Imputes probability of being a nonmover for persons with unresolved mover status (status 5).

The imputed match probability under the 1988 Style P-Sample model is

$$P_{match} = P_{match|nonmover} P_{nonmover} + P_{match|mover} P_{mover}$$

Variables that were used in the 1988 style P-Sample model are listed in table 4.

E-Sample Model

A correct census enumeration must satisfy three criteria:

(1) Correct Enumeration Status - The enumeration address must be correct.
(2) Correct Geocode - The address must be placed in the correct census geography.
(3) No Duplications - A person should be counted only once.

The probability of correct enumeration is defined as

$$P(CE) = P(CES)P(CG)/(1+N_d)$$

where P(CES) is the probability of having a correct enumeration status, P(CG) is the probability of being correctly geocoded, and N_d is

Table 4 Predictor variables used in 1988 Style P-Sample Models

Variable	Used in Model(s)	No. Levels
Constant	1,2,3,5	1
Census Division	1,2,3,5	10
Place/Type	1,2,3,5	10
Type of Enumeration Area	1,2,3	4
Small Block (yes/no)	1,2,3	2
Tenure (renter/owner)	1,2,3,5	2
Structure (single family/other)	1,2,3,5	2
Sex	1,2,3,5	2
Age Group	1,2,3,5	6
Race/Origin	1,2,3,5	5
Match Code Group (model 1)	1	14
Mover/Nonmover	2	2
Match Code Group (model 3)	3	2
Before Followup Mover Status (resolved/unresolved)	5	2
Sex*Age	1,2,3	
Sex*Race/Origin	1,2,3	
Age*Race/Origin	1,2,3	
Sex*Age*Race/Origin	1,2,3	

the number of duplications found within the surrounding block search area.

The E-Sample imputation model is composed of three logistic regression models. Models 1 and 2 are used to impute the probability of correct enumeration status and model "G" is used to impute the probability of that a household was correctly geocoded. It was considered unlikely that households with unresolved geocodes would be located within a sample block, so only households located outside of the sample blocks were used in model "G".

Model 1: Imputes probability of correct enumeration status for most unresolved cases. This model was used for all but thirty four imputations.
Model 2: Imputes probability of correct enumeration status for persons with insufficient information to attempt matching and for persons with unexpected match codes.
Model G: Imputes probability of correct geocode for households with unresolved geocodes.

Variables used in the 1998 Style E-Sample models in listed in table 5.

Table 5 Predictor variables used in 1988 Style E-Sample models.

Variable	Used in Model(s)	No. Levels
Constant	1,2,G	1
Census Division	1,2	10
Place/Type	1,2	10
Place/Type	G	4
Type of Enumeration Area	1,2	4
Type of Enumeration Area	G	3
Small Block (yes/no)	1,2,G	2
Source of Unit	1,2	8
Source of Unit	G	2
Mail Return (yes/other)	1,2	2
Tenure (renter/owner)	1,2	2
Structure (single family/other)	1,2	2
Sex	1,2	2
Age Group	1,2	6
Race/Origin	1,2	5
Match Code Group	1	8
Late Census	1,2	6
Mail Return*Match Code	1	
Sex*Age	1,2	
Sex*Race/Origin	1,2	
Age*Race/Origin	1,2	
Sex*Age*Race/Origin	1,2	

2.6 Bootstrap Samples

Three E-Sample and three P-Sample bootstrap samples were drawn in order to measure variation in the production dual system estimates arising from uncertainty in imputation model parameter estimates holding the PES sample of blocks fixed. Each bootstrap consisted of selecting households with replacement within blocks. Production model parameters were obtained for each bootstrap sample. The bootstrap sample was not used after this point. Model parameters obtained from fitting the production model to the bootstrap sample were used to impute for missing match and enumeration statuses in the original sample. Dual system estimates were computed for all combinations of P&E Sample bootstraps and the production sample. Fifteen sets of dual system estimates were obtained in addition to production.

2.7 Treatment Combinations

Dual system estimates were computed for combinations of the P&E Sample Followup Proxy Alternatives and the 1988 Style Imputation Model. All possible treatment combinations as listed below were run. This permitted analysis to be done on the treatments as a 2x2x2 factorial design.

Treatment Variable	Values
Imputation Model	Production Model 1988 Style Model
P-Sample Proxies	Production Treatment P-Sample Proxy Treatment
E-Sample Proxies	Production Treatment E-Sample Proxy Treatment

3. ANALYSIS

Percent undercounts were estimated for PES poststrata and other levels for each imputation alternative. Unsmoothed rather than smoothed adjustment factors were used in the calculations. Table 6 lists the percent undercounts for each imputation alternative at the national level. Except for the movers alternative, the differences between the production PES and the alternative imputation estimates are small. The narrow range of undercounts is due in part to the low level of missing match status (1.9%) and enumeration status (1.4%) in the PES. The movers alternative, which may be characterized as "extreme", lowered the undercount by about a half of a percent. The range of undercounts, .09% excluding the movers alternative, is small compared to the standard deviation (.19) of the unsmoothed undercount at the national level.

Table 6 Percent Undercount at National Level

Treatment	Undercount	Treatment	Undercount
Production	2.11%	Bootstrap 4	2.11%
P-Proxy	2.06%	Bootstrap 5	2.15%
E-Proxy	2.16%	Bootstrap 6	2.12%
Movers	1.65%	Bootstrap 7	2.14%
1988 Style	2.07%	Bootstrap 8	2.10%
P/E Proxy	2.11%	Bootstrap 9	2.14%
P-Proxy*1988	2.03%	Bootstrap 10	2.11%
E-Proxy*1988	2.07%	Bootstrap 11	2.13%
P/E Proxy*1988	2.02%	Bootstrap 12	2.10%
Bootstrap 1	2.15%	Bootstrap 13	2.14%
Bootstrap 2	2.12%	Bootstrap 14	2.11%
Bootstrap 3	2.14%	Bootstrap 15	2.13%

Separate analyses were done on the bootstrap treatments versus the alternative P-Proxy, E-Proxy, and 88-style imputation treatments. The bootstraps were not considered as alternatives to the production imputation, but as a means of measuring uncertainty in the production model parameter estimates.

Analysis of Reasonable Alternatives

The purpose of this analysis was to measure uncertainty in the undercounts arising from the selection of a reasonable imputation model. The P-Proxy, E-Proxy, and 88-Style alternatives may be viewed as coming from an infinite population of reasonable alternatives to the production model.

Let $\alpha, \beta,$ *and* γ denote the P-Proxy, E-Proxy, and 88-Style treatments respectively. We assume a no interaction random effects model for percent undercount,

$$u_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+e_{ijk} \qquad i=1,2;j=1,2;k=1,2$$

where

$$\alpha_1=-\alpha_2;\ \beta_1=-\beta_2;\ \gamma_1=-\gamma_2,$$

$$\alpha_1,\beta_1,\gamma_1 \ \textit{iid}\ N(0,\sigma^2),\ \textit{and}\ e_{ijk}\sim N(0,\sigma_e^2)$$

The standard deviation due to selection of the imputation model is $\sqrt{3}\sigma$. Maximum likelihood estimates of μ, σ^2, *and* σ_e^2 where obtained for the 116 PES poststrata collapsed on age and sex. The model selection standard deviations, $\sqrt{3}\sigma$, are compared with sampling standard deviations in figure 2. The model selection standard deviation of approximately 3.5 in the figure comes from a poststratum in which the enumeration status of working age males was determined

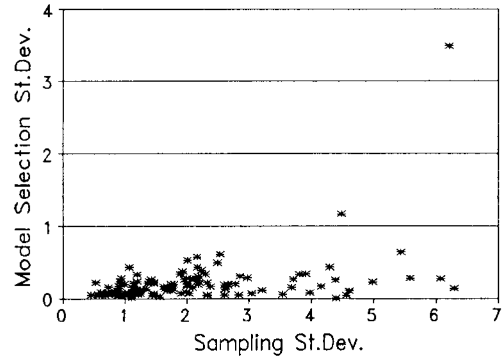

Figure 2 Standard deviations for percent undercounts of the 116 collapsed PES poststrata due to selection of reasonable imputation model versus sampling standard deviations.

largely through proxy interviews during followup. The standard deviations for model selection were always smaller than the sampling standard deviations and usually much smaller.

Given the level of missing data, the production undercount estimates were not greatly affected by the selection of a reasonable imputation model.

Bootstrap Analysis

The undercount estimates for the 116 PES poststrata were expressed as a random effects model by Joe Schafer in Mack, Schindler, and Schafer (1991). A brief summary of the model will be given here.

Undercounts were obtained for fifteen bootstrap alternatives as well as production. These alternatives represent all possible pairings of four E-Sample and four P-Sample samples (the production sample together with three P-Sample and three E-Sample bootstrap samples). The percent undercounts for the bootstrap alternatives may be expressed by the model,

$$U_{ij} = \mu + P_i + E_j + PE_{ij} \qquad i=1,2,3,4; j=1,2,3,4$$

where P_i, E_j, and PE_{ij} are normally distributed with zero means and variances σ_P^2, σ_E^2, *and* σ_{PE}^2 respectively. The total variance of the undercounts, U_{ij}, is $\sigma_T^2 = \sigma_P^2 + \sigma_E^2 + \sigma_{PE}^2$.

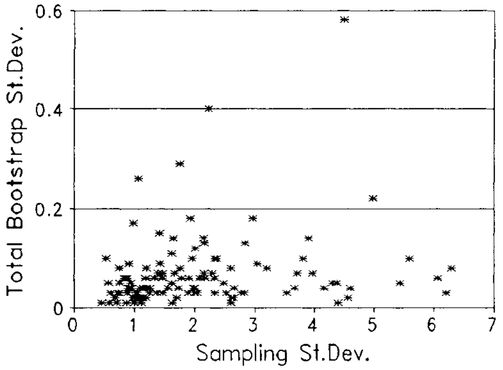

Figure 3 Total standard deviations of bootstrap procedure for percent undercounts of the 116 collapsed PES poststrata versus sampling standard deviations.

A scatter plot of σ_T versus the sampling standard deviation is given in figure 3. The uncertainty in the undercounts due to the bootstrap procedure was small compared with the uncertainty due to sampling.

ACKNOWLEDGEMENTS

This study represents the work of Joe Schafer, Eric Schindler, Roy Smith, Doug Hallam, and many others in addition to the author.

REFERENCES

Belin, T.R., Diffendal, G.J., Rubin, D.B., Schafer, J.L., and Zaslavsky, A.M. (1991). Strategy for Handling Cases with Unresolved After-Follow-up Enumeration Status in Production of Adjusted Census Counts from 1990 Census and Post-Enumeration Survey, STSD Decennial Census Memorandum Series #V-98, United States Bureau of the Census, Washington, D.C.

Hogan, Howard R. (1990). "The 1990 Post-Enumeration Survey: An Overview," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 518-523.

Mack, S.P., Schindler, E., Schafer, J.L. (1991). PES Evaluation Project P1: Analysis of Reasonable Alternatives, Final Report, 1990 Coverage Studies and Evaluation Memorandum Series #A-9, U.S. Bureau of the Census, Washington D.C.

Marks, E.S. (1978). "The Use of Dual System Estimation in Census Evaluation," In *Developments in Dual System Estimation*, (Ed. K. Krotki), Edmonton: University of Alberta Press.

Schafer, Joe (1991). Memorandum for Dave Bateman and Mary Mulry, Subject: Calculation of Bootstrap Variance for PES Evaluation Project P16, Statistical Support Division, U.S. Bureau of the Census.

Wolter, K.M. (1986). "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association*, 81, 338-346.