UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

December 31, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-42

MEMORANDUM FOR Donna Kostanich
Chair, A.C.E. Revision II Planning Group

From: Mary H. Mulry
Chair, A.C.E. Revision II Assessment Subgroup

Prepared by: Mary H. Mulry
Statistical Research Division

Randal S. ZuWallack
Decennial Statistical Studies Division

Subject: A.C.E. Revision II - Confidence Intervals and Loss Functions

The attached document was prepared to assist in assessing the Accuracy and Coverage Evaluation Revision II estimates.

This report focuses on comparing the relative accuracy of the A.C.E. Revision II estimates with Census counts.

A.C.E. Revision II PP-42
December 31, 2002

Confidence Intervals and Loss Functions

Mary H. Mulry and
Randal S. ZuWallack

Statistical Research Division and
Decennial Statistical Studies
Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

CONTENTS

LIST OF TABLES AND FIGURES

EXECUTIVE SUMMARY

We use two methods to assess the relative accuracy of the estimates of population size from the Accuracy and Coverage Evaluation Revision II (A.C.E. Revision II) and Census 2000. One method examines the quality of the census and A.C.E. Revision II through the construction of confidence intervals for the census undercount rate corrected for bias as well as variance. The other method uses a loss function analysis to compare the relative accuracy of the census and the A.C.E. Revision II for states, counties, and places.

The construction of the confidence intervals incorporates both sampling and nonsampling error derived from evaluations of components of error in the A.C.E. Revision II.. Since most of the data available on the quality of the original A.C.E. are being incorporated in the A.C.E. Revision II, the estimation of the net bias uses the data that were not included. The bias combined the error due to inconsistent reporting of variables used in poststratification, the error due to using the inmovers to represent the movers in the PES-C formulation of the dual system estimator (DSE), and the error in the identification of duplicate enumerations in the census as measured by administrative records. The estimate of the variance in A.C.E. Revision II included three error components. These are the sampling error, the error due to the choice of the missing data model, and the error due to the choice of the model for correcting for P-sample cases with enumerations outside the search area.

The measure of accuracy used by the loss functions was weighted mean square error, with weights set inversely proportional to the census counts for levels and to census shares for shares. Mean square error equals the sum of variance and squared bias, and the bias and variance estimates account for both sampling and nonsampling errors. Of course, the bias and variance estimates will themselves have errors. The effect of omitting a variance component (if the corresponding error is uncorrelated with other random effects) would be to overstate the accuracy of the A.C.E. Revision II estimate and to understate the accuracy of the census, but we have not identified significant omitted variance components. The effects of neglecting bias components is more difficult to predict for two reasons: (1) positive biases may cancel with negative biases, and (2) omitting biases affects the estimates of accuracy of both the A.C.E. Revision II estimates and the census. Thus, in general, we cannot be certain whether omitted biases will tend to make any given loss function analysis overstate or understate the comparative accuracy of the A.C.E. Revision II estimates relative to the census. Further analysis could, in principle, be done to investigate this.

Though not fully included in the loss functions, the effects of synthetic error were investigated in another study (Griffin 2002). One source of synthetic error involves correcting the individual post-stratum estimates for errors estimated at more aggregate levels (such as the corrections for correlation bias and coding errors). Two of the variance components noted above (those related to choice of imputation models and to accounting for P-Sample cases matching to census enumerations outside the search area) were included in the loss functions, these components reflect the level of the errors, not the synthetic errors from such corrections. Errors from other

such corrections, such as the adjustments for correlation bias, were not reflected. Another source of synthetic error is variations of census coverage within post-strata (something not captured by synthetic application of post-stratum coverage correction factors for specific areas). Analyses based on artificial populations that simulated patterns of coverage variation within post-strata were done to assess whether omission of resulting synthetic biases from the loss function analysis tilted the comparisons in one direction or another. These analyses did not in general change the loss function results, though they had some limitations. It should be kept in mind that synthetic error is expected to be more important the smaller are the areas whose estimates are being compared, so that any limitations of the loss functions regarding synthetic error would be expected to be more important in comparisons for small places or counties than for large places or counties.

Although the loss function analysis incorporated all the components of error for which estimates were available, and there are no known potentially large errors excluded (with the possible exception of synthetic error discussed above), there may be other biases in the A.C.E. Revision II estimates that were not included. We cannot ascertain whether omitted biases cause the loss function for shares to favor the census or the A.C.E. Revision II, because the direction depends on the signs of the correlations between the omitted biases and the expected undercount rate for the areas considered. An alternative is a sensitivity analysis that examines the effect of different amounts and distributions of error which would lead to estimates of the amounts and distributions of error needed to change the indications from the loss function analysis.

Loss function analysis considered shares for five geographic groupings and levels for five geographic groupings, with some overlap of groupings. The loss function analysis indicated that the A.C.E. Revision II is more accurate than the census for every loss function considered with the exception of levels for places with population of at least 100,000. When the places with population of at least 100,000 are split into places with population between 100,000 and 1 million and places with population of at least 1 million, the loss function analysis indicates that the bulk of the error in the A.C.E. Revision II for places with population of at least 100,000 lies in the nine (9) places with population of at least 1 million. More research is needed to understand the one exceptional result. The loss function analyses did not take synthetic estimation error into account, but separate analyses (Griffin 2002) suggest that had synthetic error been included, the conclusions would have been the same. The validity of the loss function analysis depends on the quality of the estimates of components of error in the A.C.E. Revision II, and some of those components are not accurately quantified. The resulting limitations on the loss function analysis have been discussed above.

Considering the limitations, the bias-corrected estimate of the net undercount rate for the U. S. is -0.33 percent while the A.C.E. Revision II estimate is -0.49 percent. The explanation for the estimated bias appears to be due to error in the identification of duplicates since the effects of the error due to inconsistent post-stratification variables and the error due to using inmovers to estimate movers appear very small. Additional tabulations by enumeration and residency status by domain would indicate whether the increase in the undercount rate arises from the effect of

undetected duplicates in the P-Sample or the E-Sample. For example, if the evaluation detected duplications of erroneous enumerations in the E-Sample, the A.C.E. Revision II estimate would increase.

When we examine the 95 percent confidence intervals for the net undercount rates, we find that neither the census nor A.C.E. Revision II estimates for Non-Hispanic Blacks, Non-Hispanic Black Owners or Black Renters lie within their intervals. (Remember that both the interval and the A.C.E. Revision II DSE are adjusted for estimated correlation bias.) This indicates an undercount for Non-Hispanic Blacks, both for Owners and for Renters. Additional tabulations by enumeration and residency status by domain and tenure would indicate whether the increase in the undercount rate arises from the effect of undetected duplicates in the P-Sample or the E-Sample. For example, if the evaluation detected duplications of erroneous enumerations in the E-Sample, the A.C.E. Revision II estimate would increase. The estimate of a 2.78 percent undercount rate for Blacks based on Demographic Analysis (Robinson and Adlaka 2002) does lie within the 95 percent confidence interval.

The census estimate for Non-Hispanic Whites does not lie within the interval, indicating an overcount. The intervals for all the other domains cover both the census and the A.C.E. Revision II estimate.

When we consider groups defined by tenure and domain by tenure, the 95 percent confidence intervals for the undercount rate in all groups cover the A.C.E. Revision II estimate with the exception of the Black Owners and Black Renters as stated previously. However, the census numbers for all Owners, NonHispanic White Owners, and Hispanic Owners do not fall within the 95 percent confidence interval for their undercount rates, indicating overcounts in these groups. The intervals for the other groups do cover the census with the exception of Black Owners and Black Renters mentioned above.

The major source of estimated bias in the A.C.E. Revision II concerns the estimation of census duplicates. There are two evaluations of those estimates, Census and Administrative Records Study (CARDS) (Bean and Bauder 2002) and Clerical Review of Census Duplicates (Byrne, Beaghen, and Mulry 2002). The estimation of the bias in the loss function analysis is based on CARDS. There are some discrepancies in findings from CARDS and CRCD. If these differences were resolved, one or more of the conclusions from the outcome of the loss function analysis could change. However, under the assumption that the A.C.E. Revision II estimates are unbiased and the only error components are the estimated sampling and nonsampling variance components, the loss function analysis finds that the A.C.E. Revision II estimates are more accurate than the census for all groupings considered, even for levels for places with population of at least 100,000. Further analyses assuming larger amounts of bias or a different distribution of the bias would increase the knowledge of the limitations of the data.

In summary, when viewing the results of the loss function analysis, one must keep the assumptions and limitations in mind, as well as realize that the effect of any omitted biases could

be in either direction (increasing or decreasing the estimate of the relative accuracy of the census versus the A.C.E. Revision II estimates). While the loss function evaluations suggest the superiority of the A.C.E. Revision II estimates, concerns do remain about whether the bias estimates used in the loss function analysis are of sufficient quality to assure the correctness of the results.

1. BACKGROUND

Two methods of assessing the relative accuracy of the Census and the A.C.E. Revision II are using confidence intervals for the net undercount rate and a loss function analysis. We form the confidence intervals for net undercount rate using estimates of net bias and variance for the census coverage correction factors. Since most of the data available on the quality of the original A.C.E. are being incorporated in the A.C.E. Revision II, the estimation of the net bias will use the data that were not included. In the loss function analysis, we estimate loss by the weighted Mean Squared Error (MSE), with the weight of the reciprocal of the census count for levels and the reciprocal of the census share for shares. We estimate the aggregate loss for levels and shares for states, counties, and places across the nation and for counties and places within state.

These methods for evaluating the accuracy of the census and an adjustment of the census have been used previously (Mulry and Spencer 1993, 2001; CAPE 1992).

2. METHODOLOGY

Appendix A contains a detailed description of the methodology for the construction of the confidence intervals for the undercount rate while Appendix B describes the loss function methodology. Appendix C contains the specifications underlying the computer programs used to calculate expected loss.

The construction of the confidence intervals incorporates both sampling and nonsampling error. Since most of the data available on the quality of the original A.C.E. is being incorporated in the A.C.E. Revision II, the estimation of the net bias will use the data that was not included. The bias combined the error due to inconsistent reporting of variables used in poststratification (Bench 2002), the error due to using the inmovers to represent the movers in the PES-C formulation of the dual system estimator (DSE) (Keathley 2002), and the error in the identification of duplicate enumerations in the census as measured by administrative records (Bean and Bauder 2002). The estimate of the variance in A.C.E. Revision II included three error components. These are the sampling error, the error due to the choice of the missing data model (Kearney 2002), and the error due to the choice of model for correcting for duplicate enumerations (Davis 2002).

Confidence intervals that incorporate the net bias as well as the variance for the net undercount rate $\hat{U}$ provide a method for comparing the relative accuracy of the census and the A.C.E. Revision II estimates. We construct the intervals by estimating the net bias and variance in the census coverage correction factor for each poststratum. Then we can estimate the bias $\hat{B}(\hat{U})$ and variance V in the net undercount rate $\hat{U}$ and form the 95% confidence interval for the net undercount rate for a poststratum or a group of poststrata by

$$(\hat{U}-\hat{B}(\hat{U})-2\hat{V}^{1/2},\hat{U}-\hat{B}(\hat{U})+2\hat{V}^{1/2}).$$

Since $\hat{U}=0$ corresponds to no adjustment of the census, one comparison of the relative accuracy of the census and the A.C.E. Revision II estimates is based on an assessment of whether the confidence intervals for the evaluation poststrata cover 0 and $\hat{U}$.

The loss function analysis uses the estimated bias and variance to estimate an aggregate expected loss for the census and the A.C.E. Revision II for levels and shares for counties and places across the nation and within state. The loss function is the weighted squared error, which also may be described as the weighted Mean Squared Error (MSE). The weight for both the census loss and the A.C.E. Revision II loss calculation is the reciprocal of the census count for levels and the reciprocal of the census share for shares. The motivation for the selection of the groupings of areas for the loss functions is the potential use of the A.C.E. Revision II estimates in the postcensal estimates program.

Loss function analyses were carried for the following groups:
Levels
1. All Counties with population of 100,000 or less
2. All Counties with population greater than 100,000
3. All places with population at least 25,000 but less than 50,000
4. All places with population at least 50,000 but less than 100,000
5. All places with population greater than 100,000

State Shares
1. All Counties
2. All places

US Shares
1. All places with population at least 25,000 but less than 50,000
2. All places with population at least 50,000 but less than 100,000
3. All places with population greater than 100,000
4. All states

3. LIMITATIONS

- The estimated bias in the A.C.E. Revision II estimates may not account for all the sources of bias or may not account for the included nonsampling error components well. This could be a problem for some of the estimates derived from Census and Administrative Records Study (Bean and Bauder 2002), for example (based on Clerical Review of Census Duplicates (Byrne, Beaghen, and Mulry 2002)). Due to time limitations, estimates of ratio-estimator bias are not included. Estimates of correlation bias used in the A.C.E. Revision II are assumed to be without error.

- The estimated variance in the A.C.E. Revision II estimates may not account for all the sources of variance or may not account for the included nonsampling error components well, especially for error from choice of model for accounting for duplicates.

- Synthetic error, which is not included in the loss function analysis, may arise from two sources. One source of synthetic error involves correcting the individual post-stratum estimates for errors estimated at more aggregate levels (such as the corrections for correlation bias and coding errors). Another source of synthetic error is variations of census coverage within post-strata (something not captured by synthetic application of post-stratum coverage correction factors for specific areas). Analyses based on artificial populations that simulated patterns of coverage variation within post-strata were done to assess whether omission of resulting synthetic biases from the loss function analysis tilted the comparisons in one direction or another. These analyses did not in general change the loss function results, though they had some limitations (Griffin 2002). It should be kept in mind that synthetic error is expected to be more important the smaller are the areas whose estimates are being compared, so that any limitations of the loss functions regarding synthetic error would be expected to be more important in comparisons for small places or counties than for large places or counties.

- The construction of the bias-corrected confidence intervals and the loss function analysis excludes consideration of the following errors:
  - synthetic estimation error
  - response error and coding error in A.C.E. Revision II P-Sample residency and match status and E-Sample correct enumeration status (e.g., conflicting cases)
  - response error and coding error in A.C.E. Revision II P-Sample mover status
  - error in Demographic Analysis sex ratios for correlation bias estimation
  - error due to the model used to estimate correlation bias from Demographic Analysis sex ratios
  - error due to the model for estimating the effect of E-Sample cases with

- The expected loss could instead have been measured by a loss function other than squared error weighted by the reciprocal of the census count.

- The effect of omitting a variance component (if the corresponding error is uncorrelated with other random effects) would be to overstate the accuracy of the A.C.E. Revision II estimate and to understate the accuracy of the census.

- The effects of neglecting bias components are more difficult to predict for two reasons: (1) positive biases may cancel with negative biases, and (2) omitting biases affects the estimates of accuracy of both the A.C.E. Revision II estimates and the census. The direction of the effect of omitted biases on the comparison of accuracy depends on the sign of a weighted sum of products of neglected biases and expected values of the undercount estimates (Mulry and Spencer 2001, p.6). The limitation of omitted biases

does not predictably tilt the loss function analysis to "favor" either the A.C.E. Revision II estimates or the census estimates in the comparisons of accuracy.

4. RESULTS

4.1 Confidence Intervals

Table 1 shows the variance components for the A.C.E. Revision II estimates of net undercount rate for groups defined by race/Hispanic origin domain and tenure. The sampling variance was estimated using an alternative variance estimator that treats the correlation bias correction factor as a scalar. Table 2 displays 95 percent confidence intervals for the net undercount rate, tabulated by evaluation poststratum. Figures 1 and 2 show graphs of the bias-corrected confidence intervals and by domain and the cross-classification of domain by tenure, respectively.

Table 1. Variance components for the A.C.E. Revision II estimates of undercount rate (percent)

Domain & tenure group	Census	A.C.E. Revision II UC Rate	SE sampling*	SE imputation model	SE duplication model
Total	273586997	-0.49	0.19	0.10	0.18
Owner	187924850	-1.25	0.19	0.07	0.16
Renter	85662147	1.14	0.35	0.18	0.21
AIAN on Reservations	540158	-0.88	1.53	1.48	0.54
AIAN on Res. Owner	366462	-0.74	1.62	1.79	0.62
AIAN on Res. Renter	173696	-1.17	1.94	0.98	0.37
AIAN off Reservations	1564953	0.62	1.30	0.14	0.28
AIAN off Res Owner	921447	-1.53	1.81	0.11	0.32
AIAN off Res Renter	643506	3.54	2.04	0.21	0.23
Hispanic	34538121	0.71	0.42	0.14	0.19
Hispanic Owner	16793484	-1.08	0.48	0.10	0.17
Hispanic Renter	17744637	2.35	0.58	0.18	0.21
Non-Hispanic Black	33469965	1.84	0.37	0.12	0.20
Black Owner	16547598	0.56	0.48	0.08	0.18
Black Renter	16922367	3.06	0.52	0.17	0.22
Hawaiian / PI	590208	2.12	2.61	0.28	0.41
NHPI Owner	306450	0.67	3.89	0.16	0.10
NHPI Renter	283758	3.64	3.41	0.53	0.72
Asian	9959604	-0.75	0.67	0.14	0.11
Asian Owner	6032323	-1.71	0.86	0.12	0.07
Asian Renter	3927281	0.68	0.94	0.17	0.18
Non Hispanic White	192923988	-1.13	0.19	0.09	0.17
White Owner	146957086	-1.46	0.19	0.06	0.15
White Renter	45966902	-0.07	0.40	0.19	0.21

*The standard error for sampling uses an alternative variance estimator.

Table 2. 95 % Confidence Intervals for Undercount Rate (percent)

Domain & tenure group	Census	UC rate	bias-corrected UC rate	SE(UC)	Lower bound	Upper bound
Total US	273,586,997	-0.49	-0.33	0.28	-0.89	0.22
Owner	187,924,850	-1.25	-0.89	0.26	-1.41	-0.37
Renter	85,662,147	1.14	0.86	0.44	-0.02	1.74
AIAN on Res	540,158	-0.88	-1.41	2.29	-6.00	3.18
AIAN on Res Owner	366,462	-0.74	-1.49	2.54	-6.58	3.60
AIAN on Res Renter	173,696	-1.17	-1.23	2.48	-6.19	3.73
AIAN off Res	1,564,953	0.62	-0.49	1.43	-3.35	2.37
AIAN off Res Owner	921,447	-1.53	-2.91	1.95	-6.81	0.98
AIAN off Res Renter	643,506	3.54	2.79	2.21	-1.62	7.20
Hispanic	34,538,121	0.71	-0.18	0.49	-1.16	0.80
Hispanic Owner	16,793,484	-1.08	-1.63	0.54	-2.70	-0.55
Hispanic Renter	17,744,637	2.35	1.15	0.66	-0.17	2.47
Non-Hispanic Black	33,469,965	1.84	3.56	0.42	2.72	4.40
Black Owner	16,547,598	0.56	2.83	0.49	1.84	3.81
Black Renter	16,922,367	3.06	4.27	0.56	3.16	5.39
Hawaiian/PI	590,208	2.12	0.30	2.23	-4.16	4.76
NHPI Owner	306,450	0.67	-1.82	3.33	-8.47	4.84
NHPI Renter	283,758	3.64	2.49	2.91	-3.33	8.32
Asian	9,959,604	-0.75	-0.83	0.68	-2.19	0.54
Asian Owner	6,032,323	-1.71	-1.73	0.85	-3.44	-0.02
Asian Renter	3,927,281	0.68	0.53	0.96	-1.40	2.45
Non Hispanic White	192,923,988	-1.13	-1.04	0.27	-1.58	-0.50
NH White Owner	146,957,086	-1.46	-1.19	0.26	-1.70	-0.68
NH White Renter	45,966,902	-0.07	-0.57	0.48	-1.53	0.38

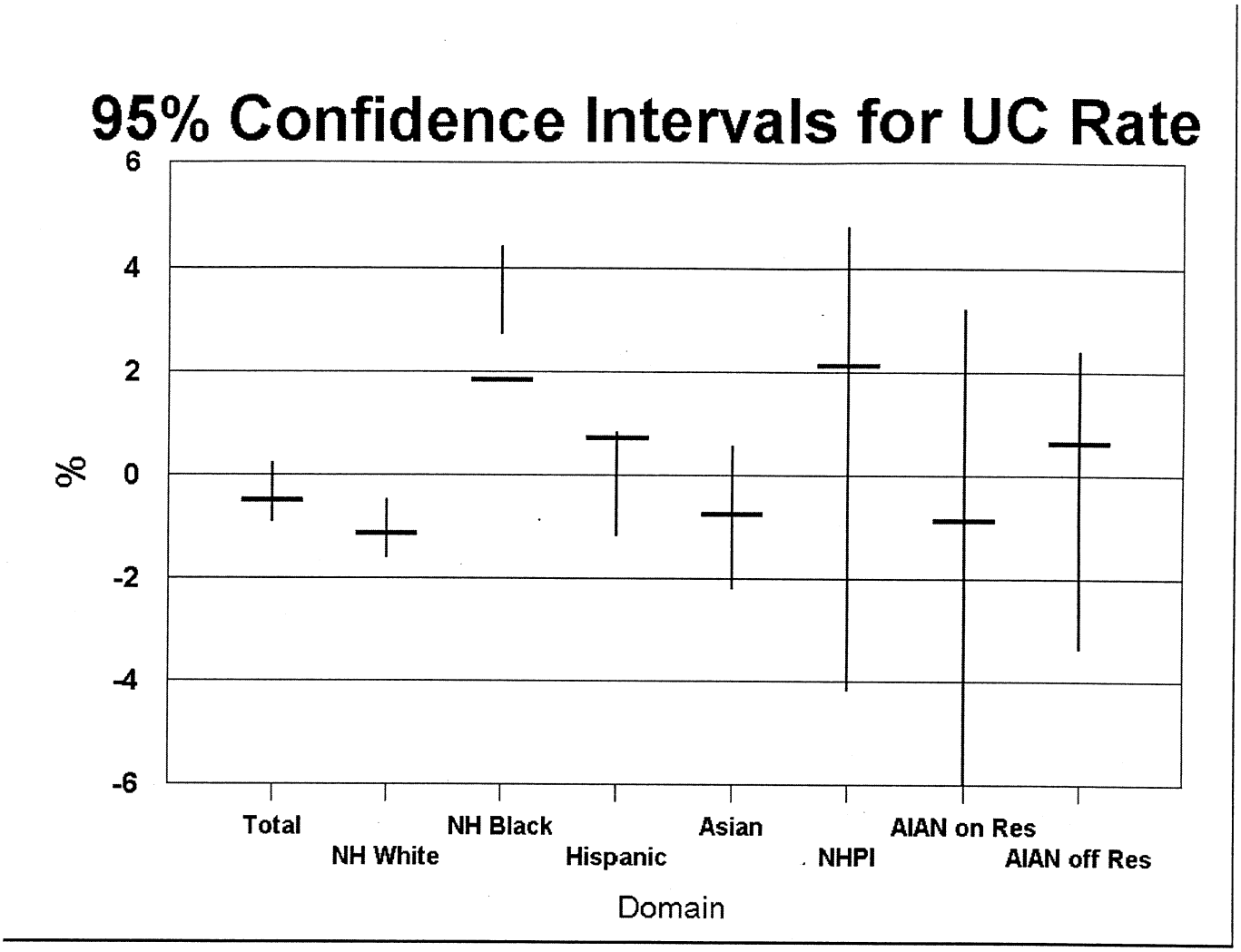

Figure 1. 95% confidence intervals for census undercount rate by domain.
Note. The intervals, denoted by vertical lines, are adjusted for estimated biases from inconsistent reporting of poststratification variables, estimation of movers with inmovers, and identification of duplicates by use of administrative records. The horizontal lines denote the A.C.E. Revision II DSEs.

Neither the census nor A.C.E. Revision II estimate for Non-Hispanic Blacks lies within the interval. (Remember that both the interval and the A.C.E. Revision II DSE are adjusted for estimated correlation bias.). Additional tabulations by enumeration and residency status by domain would indicate whether the increase in the undercount rate arises from the effect of undetected duplicates in the P-Sample or the E-Sample. The estimate of a 2.78 percent undercount rate for Blacks based on Demographic Analysis (Robinson and Adlaka 2002) does lie within the 95 percent confidence interval. The census estimate for Non-Hispanic Whites does not lie within the interval although the A.C.E. Revision II estimate does. The intervals for all the other domains cover both the census and the A.C.E. Revision II estimate.

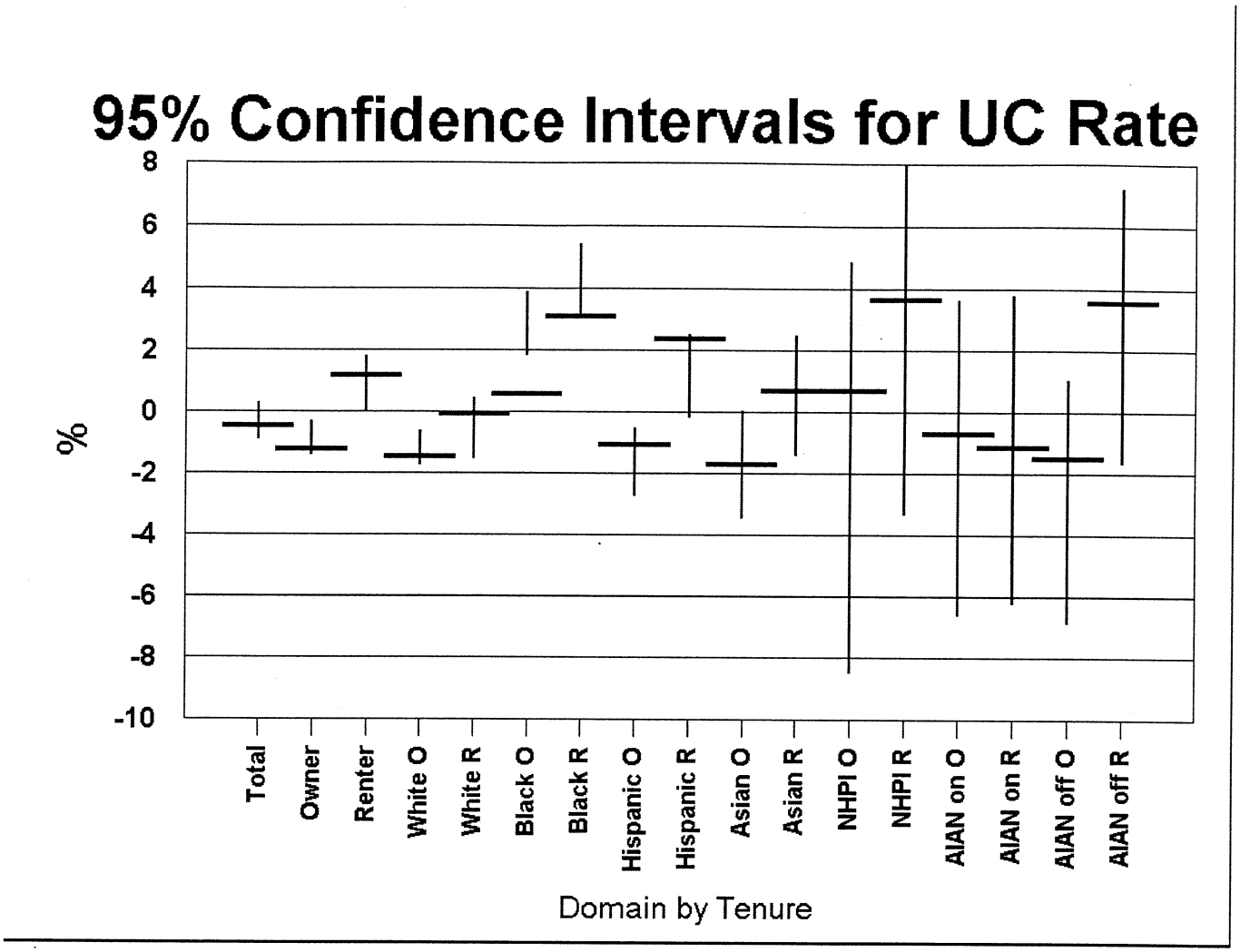

Figure 2. 95% confidence intervals for census undercount rate by tenure and domain.
Note. The intervals, denoted by vertical lines, are adjusted for estimated biases from inconsistent reporting of poststratification variables, estimation of movers with inmovers, and identification of duplicates by use of administrative records. The horizontal lines denote the A.C.E. Revision II DSEs.

Neither the census nor the A.C.E. Revision II estimates for Black Owners and for Black Renters lie within the 95 percent confidence intervals. (Remember that both the interval and the A.C.E. Revision II DSE are adjusted for estimated correlation bias.) Additional tabulations by enumeration and residency status by domain and tenure would indicate whether the increase in the undercount rate for these groups arises from the effect of undetected duplicates in the P-Sample or the E-Sample. For example, if the evaluation detected duplications of erroneous enumerations in the E-Sample, the A.C.E. Revision II estimate would increase. The intervals all other groups cover the A.C.E. Revision II estimate.

The census numbers for all Owners, NonHispanic White Owners, and Hispanic Owners do not fall within their intervals. The intervals for the other groups do cover the census with the exception of Black Owners and Black Renters mentioned in the previous paragraph.

4.2 Loss Functions

The loss function analyses are available for all groups except the within state shares for all places, which was planned but not completed. The bias estimate combines three error components:

- error from inconsistent reporting of poststratification variables
- error from estimating the movers with inmovers
- error in the identification of E-sample cases with duplicates and P-sample cases that link to enumerations outside the search area as measured by administrative records.

The variance includes components for the following:

- sample error
- error due to choice of imputation model
- error due to choice of the model for correcting for P-sample duplicates

The results indicate smaller expected loss for the DSE than the census for all of the shares considered, and smaller expected loss for all of the levels except for all places with population greater than 100,000. For insight, consider the following totals for all places with population of at least 100,000 as estimated by the census, the A.C.E. Revision II DSE, and the Target, which is equal to the DSE minus its estimated bias:

Census	71,829,465
A.C.E. Revision II	71,967,488
Target	71,512,212.

Comparison of the census and the target shows a *net overcount* in the census for these areas, but the excess of the DSE over the target and the census indicates that the DSE estimated a *net undercount*. Thus, the analysis indicates that DSE has either overestimated the number of census misses or underestimated the number of duplicates or both. A tabulation of the CARDS results would determine if it suggests that the A.C.E. Revision II missed large numbers of duplicates in these areas. When CARDS finds duplicates for erroneous enumerations, the effect in the estimation is to increase the correct enumeration rate.

Table 3. Loss function results shares

Geo Group	No. of Areas	Census Loss	DSE Loss	Census Loss / DSE Loss	Census Loss - DSE Loss
St Share All counties	3141	.001716411	.000590178	2.90829	.001126233
US Share Places > 25,000 and < 50,000	595	.000060290	.000016450	3.66513	.000043840
US Share Places > 50,000 and <100,000	322	.000054805	.000014229	3.85154	.000040575
US Share Places > = 100,000	223	.000035764	.000009452	3.78357	.000026311
US Share All states	51	.000023505	.000005291	4.44263	.000018214

Table 4. Loss function results levels

Geo Group	No. of Areas	Census Loss	DSE Loss	Census Loss / DSE Loss	Census Loss - DSE Loss
Counties < =100,000	2617	15514.05	3730.64	4.15855	11783.41
Counties > 100,000	524	21810.87	9258.60	2.35574	12552.27
Places > 25,000 and < 50,000	595	2785.92	966.25	2.88323	1819.67
Places > 50,000 and <100,000	322	2537.46	1070.09	2.37127	1467.37
Places > = 100,000	223	3251.54	4271.10	0.76129	-1019.56

To gain further insight into the loss function results for levels for places with population of at least 100,000, we examined loss function results when these places were separated into two groups, those with population between 100,00 and 1 million and those with population of at least 1 million. We found that the loss functions still indicated less expected loss for the DSE for levels for both groups of counties. For levels for places with population between 100,000 and 1 million, the ratio of census loss to DSE increased almost to 1. However, for levels for places with population over 1 million, the ratio of census loss to DSE loss decreased substantially compared to the ratio for levels for places with population of at least 100,000. This results indicates that the bulk of the error in the A.C.E. Revision II for places with population of at least 100,000 appears to lie in the nine(9) places with population of at least 1 million. Additional tabulations would aid in explaining this result.

Table 5. Loss function results levels

Geo Group	No. of Areas	Census Loss	DSE Loss	Census Loss / DSE Loss	Census Loss - DSE Loss
Counties > 100,000 and < 1 million	490	16779.25	5726.92	2.92989	11052.33
Places > 100,000 and < 1 million	214	2573.07	2671.02	0.96333	-97.95
Counties >1 million	34	5031.62	3531.68	1.42471	1499.94
Places > 1 million	9	678.47	1600.08	0.42402	-921.61

Loss function analyses were also carried out under the assumption that the modeling of duplicates was without error. The latter assumption served to increase the estimate of census loss and decrease the estimate of DSE loss, but the findings were not qualitatively different than the results discussed above.

Table 6. Loss function results shares (without variance from duplication modeling)

Geo Group	No. of Areas	Census Loss	DSE Loss	Census Loss / DSE Loss	Census Loss - DSE Loss
St Share All counties	3141	.001726743	.000579856	2.97788	.001146886
US Share Places > 25,000 and < 50,000	595	.000060451	.000016288	3.71130	.000044163
US Share Places > 50,000 and <100,000	322	.000054911	.000014123	3.88803	.000040788
US Share Places > = 100,000	223	.000035813	.000009403	3.80848	.000026409
US Share All states	51	.000023566	.000005229	4.50659	.000018337

Table 7. Loss function results levels (without variance from duplication modeling)

Geo Group	No. of Areas	Census Loss	DSE Loss	Census Loss / DSE Loss	Census Loss - DSE Loss
Counties < = 100,000	2618	15814.11	3430.60	4.60973	12383.52
Counties > 100,000	524	22370.33	8699.14	2.57156	13671.19
Places > 25,000 and < 50,000	595	2841.12	911.05	3.11851	1930.07
Places > 50,000 and <100,000	322	2593.61	1013.94	2.55795	1579.67
Places > = 100,000	223	3440.26	4082.37	0.84271	-642.11

Tables 8 and 9 show the loss function results when the Targets include only the bias due to inconsistent reporting of poststratification variables, which is very near zero. Under the assumption that the remaining errors are the only errors, this loss function analysis shows that the A.C.E. Revision II estimate has less error than the census for levels and shares for all groups considered, even for levels for places with population of at least 100,000.

Table 8. Loss function results shares (assumes only bias is due to inconsistent poststratification variables)

Geo Group	Areas	Census Loss	DSE Loss	Census Loss / DSE Loss	Census Loss - DSE Loss
St Share All counties	3141	0.001834008	0.000554593	3.30694	0.001279414
US Share Places > 25,000 and < 50,000	595	0.000070066	0.000013295	5.27013	0.000056771
US Share Places > 50,000 and <100,000	322	0.000064339	0.000011150	5.77039	0.000053189
US Share Places > = 100,000	223	0.000040845	0.000007713	5.29556	0.000033132
US Share All states	51	0.000028659	0.000004483	6.39283	0.000024176

Table 9. Loss function results levels (assumes only bias is due to inconsistent poststratification variables)

Geo Group	Areas	Census Loss	DSE Loss	Census Loss / DSE Loss	Census Loss - DSE Loss
Counties < 100,000	2617	9420.74	2266.65	4.15624	7154.09
Counties > 100,000	524	10508.6	3622.92	2.90058	6885.63
Places > 25,000 and < 50,000	595	1623.26	430.03	3.77475	1193.23
Places > 50,000 and <100,000	322	1419.14	416.66	3.40598	1002.48
Places > = 100,000	223	2596.08	1257.78	2.06402	1338.3

5. CONCLUSIONS

Evaluations were performed on the A.C.E. Revision II estimates to estimate bias (systematic error) and variance (random error) for use in constructing bias-corrected confidence intervals and in a loss function analysis. The evaluations of bias were relatively limited because data that previously were used to estimate bias were incorporated into the A.C.E. Revision II estimates in order to correct for major errors discovered in the March 2001 A.C.E. estimates. The limited data available for evaluation of bias does not itself reflect negatively on the A.C.E. Revision II estimates; in fact, it is because of the corrections for major errors that we believe the A.C.E. Revision II estimates to be of much higher quality than the March 2001 A.C.E. estimates. Nevertheless, although the evaluations do account for the variance arising from the corrections for bias, the corrections for bias in the A.C.E. Revision II estimates may themselves be subject to bias, the magnitude of which has not been quantified. This is particularly true for the corrections for correlation bias and for P-Sample cases that matched census enumerations outside the A.C.E. search area.

The evaluations detected a small amount of bias in the A.C.E. Revision II estimate of the net undercount rate at the national level, only 0.16 percent. The explanation for the estimated bias appears to be due to error in the identification of duplicates since the effects of the error due to inconsistent post-stratification variables and the error due to using inmovers to estimate movers appear very small. Additional tabulations by enumeration and residency status by domain would

indicate whether the increase in the undercount rate arises from the effect of undetected duplicates in the P-Sample or the E-Sample. For example, if the evaluation detected duplications of erroneous enumerations in the E-Sample, the A.C.E. Revision II estimate would increase.

Based on the bias-corrected 95-percent confidence intervals, both the census and the A.C.E. Revision II estimates are too low for Non-Hispanic Blacks and both Non-Hispanic Black Owners and Renters. The intervals show the census is too high for Non-Hispanic Whites, Owners, White Owners, and Hispanic Owners. All other census and A.C.E. Revision II estimates are covered by their bias-corrected 95-percent confidence intervals. The source of most of the bias estimate is the CARDS evaluation of the identification of duplicates. Tabulations of the CARDS E-Sample and P-Sample cases by race/ethnicity domain and enumeration (or residency) status would explain how the bias arises.

The loss function analysis examines the relative accuracy by using the estimates of sampling variance and nonsampling bias and variance to estimate the aggregate expected loss for the census and the A.C.E. Revision II for levels and shares for counties and places across the nation and within state. The analyses indicated that the A.C.E. Revision II is more accurate than the census for every loss function considered with the exception of levels for places with population of at least 100,000. The bulk of the error in the A.C.E. Revision II for places with population of at least 100,000 appears to lie in the nine (9) places with population of at least 1 million. More research is needed to understand the one exceptional result. The validity of the loss function analysis depends on the quality of the estimates of components of error in the A.C.E. Revision II, and some of those components are not accurately quantified. The resulting limitations on the loss function analysis are discussed in Section 3.

The major source of estimated bias in the A.C.E. Revision II concerns the estimation of census duplicates. There are two evaluations of those estimates, Census and Administrative Records Study (CARDS) (Bean and Bauder 2002) and Clerical Review of Census Duplicates (Byrne, Beaghen, and Mulry 2002). The estimation of the bias in the loss function analysis is based on CARDS. There are some discrepancies in findings from CARDS and CRCD. If these differences were resolved, one or more of the conclusions from the outcome of the loss function analysis could change. However, under the assumption that the A.C.E. Revision II estimates are unbiased and the only error components are the estimated sampling and nonsampling variance components, the loss function analysis finds that the A.C.E. Revision II estimates are more accurate than the census for all groupings considered, even for levels for places with population of at least 100,000. Further analyses assuming larger amounts of bias or a different distribution of the bias would increase the knowledge of the limitations of the data.

In summary, when viewing the results of the loss function analysis, one must keep the assumptions and limitations in mind, as well as realize that the effect of any omitted biases could be in either direction (increasing or decreasing the estimate of the relative accuracy of the census versus the A.C.E. Revision II estimates). While the loss function evaluations suggest the superiority of the A.C.E. Revision II estimates, concerns do remain about whether the bias estimates used in the loss function analysis are of sufficient quality to assure the correctness of the results.

6. REFERENCES

Byrne, Rosemary, Beaghen, Michael, and Mulry, Mary H. (2002) "Clerical Review of Census Duplicates". DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 43. Census Bureau, Washington, DC.

Bean, Susanne L. and Bauder, D. Mark (2002) "Census and Administrative Records Duplication Study," DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-44. Census Bureau, Washington, DC.

Bench, Katie (2002) "P-sample Match Rate Corrected for Error Due to Inconsistent Poststratification Variables." DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 46. Census Bureau, Washington, DC.

CAPE (1992) "Additional Research on Accuracy of Adjusted Versus Unadjusted 1990 Census Base for Use in Intercensal Estimates". Report of the Committee on Adjustment of Postcensal Estimates, Census Bureau, November 25, 1992.

Davis, Pete (2002) "Integration of Duplicate Links into the Full Sample Estimation Files". DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 25. Census Bureau, Washington, DC.

Griffin, Richard (2002) "Assessment of Synthetic Assumption". DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-49. Census Bureau, Washington, DC.

Kearney, Anne (2002) " Evaluation of Missing Data Model." DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 48. Census Bureau, Washington, DC.

Keathley, Don (2002) "Error Due to Estimating Outmovers Using Inmovers in PES-C". DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 47. Census Bureau, Washington, DC.

Mulry, Mary (2002). "Chapter 7: Assessing the Estimates," in "A.C.E. Revision II: Design and Methodology." A.C.E. REVISION II MEMORANDUM SERIES #PP- 30. Census Bureau, Washington, DC.

Mulry, Mary H. and Spencer, Bruce D. (2001) "Overview of Total Error Modeling and Loss Function Analysis". DSSD Census 2000 Procedures and Operations Memorandum Series B-19* Census Bureau, Washington, DC.

Mulry, Mary H. and Spencer, Bruce D. (1993) "Accuracy of the of the 1990 Census and Undercount Adjustments". Journal of the American Statistical Association, 88, 1080-1091.

Robinson, J. Gregory and Adlaka, Arjun (2002) "Comparison of A.C.E. Revision II Results with Demographic Analysis".DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 41. Census Bureau, Washington, DC.

APPENDIX A

Estimating Bias in the A.C.E. Revision II Estimates and Forming Confidence Intervals

Mary H. Mulry

This appendix describes a method for estimating the bias in the A.C.E. Revision II from four sources of error under the assumption that all other errors are zero, or at least negligible. The four sources of error are the error due inconsistency in the E-sample and P-sample reporting of the characteristics used in defining the poststrata, the error in identifying cases with census duplicates in both the E- and P-samples, the error due to using inmovers for outmovers in PES-C, and ratio estimator bias.

In addition, we describe the construction of confidence intervals for adjustment factors for estimation cells or aggregates of estimation cells, such as evaluation estimation cells.

We examine the errors with the current formulation of the match rate for the calculation of the A.C.E. Revision II presented in "Summary of A.C.E. Revision II Methodology" (Kostanich 2002). We will use the same definitions of variables as found in the draft of Summary of A.C.E. Revision II Methodology.

First we discuss the correct enumeration rate and the match rate defined in "Summary of A.C.E. Revision II Methodology". Then we discuss each of the four error components and develop how to estimate the bias from their combined effect.

Correct enumeration rate for A.C.E. Revision II

The correct enumeration rate for poststratum i for the calculation of the A.C.E. Revision II from Equation (5) of the draft of Chapter 6 is the following:

$$r_{CE,i} = \frac{CE_i^{ND} f_{i'}^{E} + \sum_{t \in i} W_{P,t} p_t z_t PR(CE)}{E_i}$$

where

E_i = total weighted E-sample in poststratum i.

CE_i^{ND} = correct enumerations without a census duplicate in poststratum i

$f_{i'}^{E}$ = double sampling adjustment for E-sample in Revision Sample poststratum i'. The Revision Sample poststrata are collapsed A.C.E. sample poststrata.

$PR(CE)$ = probability of that t is a correct enumeration (CEPROBF)

p_t = probability that enumeration t has a census duplicate outside the search area

z_t = probability that enumeration t with a census duplicate outside the search area is retained after unduplication (see draft of "Summary of A.C.E. Revision II Methodology")

$W_{P,t}$ = E-sample weight for person t.

For ease of discussion, we rewrite the correct enumeration rate for poststratum i as

$$r_{CE,i} = \frac{CE_i^{ND} f_{i'}^{E} + CE_i^{D}}{E_i}$$

where

$$CE_i^{D} = \sum_{t \in i} W_{P,t}\, p_t\, z_t\, PR(CE)$$

= correct enumerations with census duplicates in poststratum i .

Match rate for A.C.E. Revision II

The match rate for poststratum j for the calculation of the A.C.E. Revision II DSE from the draft of Summary of A.C.E. Revision II Methodology is the following:

$$r_{M,j} = \frac{M_{nm,j}^{ND} f_{j'}^{Mnm} + \frac{M_{om,j} f_{j'}^{Mom}}{P_{om,j} f_{j'}^{Pom}} [\, P_{im,j} f_{j'}^{Pim} + g_j \sum_{s \in j} W_{P,s}\, p_s\, (1-h_s)\, PR(res)\,] + \sum_{s \in j} W_{P,s}\, p_s\, h_s\, PR(res)\, PRm_{P,s}}{P_{nm,j}^{ND} f_{j'}^{Pnm} + [\, P_{im,j} f_{j'}^{Pim} + g_j \sum_{s \in j} W_{P,s}\, p_s\, (1-h_s)\, PR(res)\,] + \sum_{s \in j} W_{P,s}\, p_s\, h_s\, PR(res)}$$

where

$P_{nm,j}^{ND}$ = P-sample nonmovers without a census duplicate in poststratum j

$M_{nm,j}^{ND}$ = P-sample nonmover matches without a census duplicate in poststratum j

$P_{om,j}$ = P-sample outmovers in poststratum j

$M_{om,j}$ = P-sample outmover matches in poststratum j

$P_{im,j}$ = P-sample inmovers in poststratum j

$f_{j'}^{PG}$ = double sampling adjustment for P-sample group G, where G = nm, om, or im, in Revision Sample poststratum j' . The Revision Sample poststrata are collapsed A.C.E. sample poststrata.

$f_{j'}^{MG}$ = double sampling adjustment for matches in group G, where G = nm or om, in Revision Sample poststratum j'.

p_s = probability that person s has a census duplicate outside the search area

h_s = probability that person s with a census duplicate outside the search area is retained after unduplication (see draft of "Summary of A.C.E. Revision II Methodology")

$W_{P,s}$ = P-sample weight for person s. The weight is assumed to include the probability of residence in draft Chapter 6, but that formulation needs to be reconsidered.

$PR(Res)$ = probability of being a resident of the sample block on Census Day (RPROB).

$PRm_{P,s}$ = probability person s with a census duplicate was matched in production

g_j = estimated proportion of P-sample persons in poststratum j with census duplicates outside the search area who are not retained as resident nonmovers by the duplicate study because they should have been coded as inmovers.

For ease of discussion, we rewrite the match rate for poststratum j as

$$r_{M,j} = \frac{M_{nm,j}^{ND} f_{j'}^{Mnm} + \frac{M_{om,j} f_{j'}^{Mom}}{P_{om,j} f_{j'}^{Pom}} [P_{im,j} f_{j'}^{Pim} + P_{nm-im,j}^{D}] + M_{nm,j}^{D}}{P_{nm,j}^{ND} f_{j'}^{Pnm} + P_{im,j} f_{j'}^{Pim} + P_{nm-im,j}^{D} + P_{nm,j}^{D}}$$

where

$P_{nm,j}^{D} = \sum_{s \in j} W_{P,s} p_s h_s Pr(Res)$ = P-sample nonmovers with census duplicates in poststratum j

$P_{nm-im,j}^{D} = g_j \sum_{s \in j} W_{P,s} p_s (1-h_s) Pr(Res)$ = P-sample nonmovers with census duplicates in poststratum j who are not retained as nonmovers by the duplicate study because they should have been coded as inmovers.

$M_{nm,j}^{D} = \sum_{s \in j} W_{P,s} p_s h_s PRm_{P,s} Pr(Res)$ = P-sample nonmover matches with census duplicates in poststratum j

Pr(Res) = probability of being a resident (RBROB)

Corrections based on results of CARDS

Errors in the identification of census duplicates outside the search area may create a bias in the dual system estimate designed for the A.C.E. Revision II This bias affects the estimation of the E-sample correct enumeration rate and the P-sample match rate.

The Census and Administrative Records Duplication Study (CARDS) uses files created with administrative records to examine the effectiveness of the Further Study of Person Duplication in Census 2000 (FSPD) methodology. The FSPD refines the methodology for identifying and estimating the number of census duplicates. Using a computer matching algorithm, the study performs a match of the cases in the E-sample and the A.C.E. population sample, called the P-sample, to the census records for the entire nation. Links between the E-sample or the P-sample and the census enumerations are referred to as duplicates. CARDS first links the E and P samples to the administrative records and then attempts to confirm or deny duplicates identified by the FSPD. In addition, CARDS also identifies duplicates missed by the computer study, evaluates the FSPD process rules, and examines patterns of duplication.

The approach we are taking uses the results of the CARDS to correct the identification of cases with duplicates in the E- and P-sample for the A.C.E. Sample and the Revision Sample, a subsample of the A.C.E. Sample. We will use a research files for the E- and P-samples.
For each case in the E- and P-sample with a FSPD census duplicate, CARDS will report it as correct, denied, or undetermined. CARDS also will identify cases in the E- and P-samples that have census duplicates not identified by the FSPD. With these results we will create new designations of cases with census duplicates, new values of the probabilities of having a census duplicate, and new values of the probabilities of cases with census duplicates being retained after unduplication.

First define new probabilities of having a census duplicate,
p_t^A = probability of census enumeration t has a census duplicate, corrected with data from administrative records,
p_s^A = probability of P-sample person s has a census duplicate, corrected with data from administrative records.

The steps for defining the corrected probabilities of having a census duplicate are as follows:

1. For census duplicates that CARDS denies:

- If in the E-sample, set $p_t^A = 0$, and if t matches s in the P-sample, set the corresponding $p_s^A = 0$.
- If in the P-sample, set $p_s^A = 0$, and if s matches t in the E-sample, set the corresponding $p_t^A = 0$.

2. For census duplicates that CARDS identifies, whether or not FSPD did:

- If in the E-sample, set p_t^A = 1, and if t matches s in the P-sample, set the corresponding p_s^A = 1.
- If in the P-sample, set p_s^A = 1, and if s matches t in the E-sample, set the corresponding p_t^A = 1.

3. For census duplicates CARDS can not determine and cases without duplicates:

- If in the E-sample, set $p_t^A = p_t$
- If in the P-sample, set $p_s^A = p_s$.

Next recalculate the new probabilities of P-sample person with a census duplicate being retained after unduplication, h_s^A, using the method described in Appendix 6.1 in draft Chapter 6 and the new set of cases with census duplicates and the new duplication probabilties. Also recalculate the new probabilities of E-sample enumeration with a census duplicate being retained after unduplication, z_t^A .

With the new p_t^A, p_s^A, z_t^A, and h_s^A, calculate a new correct enumeration rate and a new match rate. Let the superscript A denote a quantity calculated with the new p_t^A, p_s^A, z_t^A, and h_s^A .

$$r_{CE,i}^A = \frac{CE_i^{NDA} f_{i'}^{EA} + CE_i^{DA}}{E_i}$$

$$r_{M,j}^A = \frac{M_{nm,j}^{NDA} f_{j'}^{MnmA} + \frac{M_{om,j} f_{j'}^{Mom}}{P_{om,j} f_{j'}^{Pom}} [P_{im,j} f_{j'}^{Pim} + P_{nm-im,j}^{DA}] + M_{nm,j}^{DA}}{P_{nm,j}^{NDA} f_{j'}^{PnmA} + P_{im,j} f_{j'}^{Pim} + P_{nm-im,j}^{DA} + P_{nm,j}^{DA}} .$$

Correcting match rate for error due to using inmovers for outmovers

PES-C uses the number of inmovers to estimate the number of outmovers to avoid a bias caused by an underestimate of the number of movers. To examine the error caused by using the inmovers to represent outmovers, we rake the number of outmovers to total inmovers. The distribution of the raked outmovers may better describe the outmovers than the distribution of the inmovers. Incorporating a correction in the match rate for using inmovers for outmovers requires defining:

$P_{im,j}^O$ = estimate of inmovers in poststratum j after raking the outmovers to the inmovers.

Correcting match rate for inconsistent poststratification variables

As discussed in "P-Sample match rate corrected for error due to inconsistent poststratification variables", inconsistency in the E-sample and P-sample reporting of the characteristics used in defining the poststrata may create a bias in the dual system estimate (DSE). This bias affects the estimation of the P-sample match rate.

The analysis for the A.C.E. Revision II will follow a similar investigation as for the original A.C.E. The basic approach is to estimate the inconsistency in the poststratification variables using the matches and then assume that the rates also held for the nonmatches. The models used for the inconsistency of the original A.C.E. poststrata ("Estimation of Inconsistent Poststratification in the 2000 A. C. E.", by Shelby J. Haberman and Bruce D. Spencer, 12/17/01) were fitted in two steps, first (i) models for inconsistency of basic variables, and then (ii) derivation of inconsistency probabilities for poststratification given the inconsistency probabilities of the basic variables. The inconsistency probabilities led to an estimate of the bias in the P-sample match rate that was used to estimate the bias in the DSE.

The approach we are taking for the Revised DSE is to calculate the proportions for the poststrata for the A.C.E. Sample. The proportions will not be applied in calculations of the double sampling adjustments based on the Revision Sample, a subsample of the A.C.E. Sample. We assume the models in (i) and (ii) have been revised to reflect revisions to the variables used in the P-sample poststratification. Incorporating a correction in the match rate for inconsistent poststratification variables requires defining:

$\hat{f}_G(j,k)$ = the proportion of group G persons enumerated in P-sample poststratum k who belong to P-sample poststratum j, based on their E-sample poststratification variables. The estimation of this proportion is based on the matched P-sample persons in group G. In this application, group G may be nonmovers, outmovers, or inmovers.

Next we need to define the following quantities:

$$P^{NDA}_{nm,j,I} = \sum_k \hat{f}_{nm}(j,k)\ P^{NDA}_{nm,k}$$

$$M^{NDA}_{nm,j,I} = \sum_k \hat{f}_{nm}(j,k)\ M^{NDA}_{nm,k}$$

$$P_{om,j,I} = \sum_k \hat{f}_{om}(j,k)\ P_{om,k}$$

$$P^{O}_{im,j,I} = \sum_k \hat{f}_{im}(j,k)\ P^{O}_{im,k}$$

$$M_{om,j,I} = \sum_k \hat{f}_{om}(j,k)\ M_{om,k}$$

$$P^{DA}_{G,j,I} = \sum_k \hat{f}_{nm}(j,k)\ P^{DA}_{G,j}\ \text{, for G = nm or nm-im}$$

$$M_{nm,j,I}^{NDA} = \sum_{k} \hat{f}_{nm}(j,k)\ M_{nm,k}^{NDA}$$

Then we define the match rate corrected for the combination of error due to inconsistent poststratification variables, errors in identifying census duplicates, and error from using inmovers for outmovers, assuming no other errors are present, by the following:

$$r_{M,j,I}^{OA} = \frac{M_{nm,j,I}^{NDA} f_{j'}^{MnmA} + \frac{M_{om,j,I} f_{j'}^{Mom}}{P_{om,j,I} f_{j'}^{Pom}} [\ P_{im,j,I}^{O} f_{j'}^{Pim} + P_{nm-im,j,I}^{DA}\ ] + M_{nm,j,I}^{DA}}{P_{nm,j,I}^{NDA} f_{j'}^{PnmA} + P_{im,j,I}^{O} f_{j'}^{Pim} + P_{nm-im,j,I}^{DA} + P_{nm,j,I}^{DA}} .$$

Calculation of Bias in the A.C.E. Revision II

From the draft of the "Summary of A.C.E. Revision II Methodology", the A.C.E. Revision II estimate for estimation cell ij formed by the intersection of E-sample poststratum i and P-sample poststratum j is

$$DSE_{ij} = Cen_{ij}\ (1 - r_{II,ij})\ \frac{r_{CE,i}}{r_{M,j}} CB_{ij}$$

where

$r_{II,ij} = (\ II_{ij} + LA_{ij}\ ) / Cen_{ij}$, with II_{ij} as the census imputations, LA_{ij} as the late adds, Cen_{ij} as the census count including the late adds, and CB_{ij} the correlation-bias adjustment factor.

Then the bias in the A.C.E. Revision II estimate due to the combination of error due to inconsistent poststratification variables, errors in identifying census duplicates, and error from using inmovers for outmovers, assuming no other errors are present, for the estimation cell ij is given by

$$b_{ij,I}^{OA} = Cen_{ij}\ (1 - r_{II,ij})\ (\ \frac{r_{CE,i}}{r_{M,j}} - \frac{r_{CE,i}^{A}}{r_{M,j,I}^{OA}}\ ) CB_{ij} .$$

When we add the correlation bias and the ratio estimator bias, we have the following bias estimate for the A.C.E. Revision II in estimation cell ij.

$$b_{ij} = b_{ij,I}^{OA} + b_{ij}^{CB} + b_{ij}^{R} .$$

[note: delete the middle term from the preceding equation]
where

b_{ij}^{R} = the ratio estimator bias in the A.C.E. Revision II in estimation cell ij .

The calculation of the ratio estimator bias makes use of the replicates formed for calculating the variance of the adjustment factors from the A.C.E. Revision II The calculation of b_{ij}^{R} is independent of the calculation of $b_{ij,I}^{OA}$ and hopefully will be a by-product of the variance calculations.

The bias in the adjustment factor for estimation cell ij is calculated by dividing the bias by the census count

$$b_{ij}^{*} = \frac{b_{ij}}{Cen_{ij}}$$

since the definition of the adjustment factor for estimation cell ij is

$$f_{ij}^{*} = \frac{DSE_{ij}}{Cen_{ij}}$$

Calculation of variance of bias in A.C.E. Revision II

The calculation of the variance of b_{ij} considers the variance of each of the three terms separately. For the loss function analysis we will assume that the variances of the correlation bias and the ratio estimator bias are zero. When constructing confidence intervals, we will assume that the multiplicative factor used to estimate correlation bias is a scalar and multiply the sampling variance by the square of the multiplicative factor. The calculation of the variance of $b_{ij,I}^{OA}$ will use the 32 replicates of the E- and P-samples. These replicates were constructed by first partitioning the E- and P-samples into 32 groups and then removing a group. The replicate n is the whole sample with the nth group removed. For each replicate n, we will calculate $b_{ij,I,n}^{OA}$, n= 1,...,32, for each estimation cell ij. Then we will estimate the variance using a random group estimator as follows:

$$Var\ (\ b_{ij,I}^{OA}\ ) = \sum_{n}\ (\ b_{ij,I,n}^{OA} - b_{ij,I}^{OA}\ )^{2}\ .$$

We estimate the variance of the bias of the adjustment factor for estimation cell ij as follows:

$$Var\ (\ b_{ij,I}^{OA*}\ )\ =\ \frac{Var\ (\ b_{ij,I}^{OA}\ )}{Cen_{ij}^{2}}$$

Forming confidence intervals

The confidence interval for the adjustment factor for the estimation cell ij for the A.C.E. Revision II estimate uses both the bias b_{ij}^{*} and the variance V_{ij}^{*} as follows:

$$(\ f_{ij}^{*}\ -\ b_{ij}^{*}\ -\ 2\ \sqrt{V_{ij}^{*}}\ ,\ f_{ij}^{*}\ -\ b_{ij}^{*}\ +\ 2\ \sqrt{V_{ij}^{*}}\ )$$

where

$$V_{ij}^{*}\ =\ (\ f_{ij}^{CB}\ )^{\ 2}\ S_{ij}^{2}\ +\ V_{ij,M}\ +\ Var\ (\ b_{ij,I}^{OA*}\ )$$

S_{ij}^{2} = the sampling variance for the adjustment factor

f_{ij}^{CB} = the correlation bias correction factor

$V_{ij,M}$ = the variance due to missing data.

Confidence intervals for adjustment factors for aggregates of estimation cells, such as evaluation estimation cells, are defined using the same methodology. Estimates of the bias and variance as well as confidence intervals may be formed analogously for the undercount rate.

Reference

Kostanich, Donna (2002) “Summary of A.C.E. Revision II Methodology”.DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-35. Census Bureau, Washington, DC.

Total Error Model and Loss Function Analysis

for the

A.C.E. Revision II Estimates of Population

Bruce D. Spencer

Draft 2.0

October 21, 2002

1. Overview

We consider estimation of expected loss when the loss functions are weighted sums of mean squared errors (MSEs) of the form $E(X_i - \theta_i)^2$, with i referring to the population of an area or other subgroup and X_i and θ_i referring either to population levels (numbers of people) or to shares (fraction of population in area or group i). The sums are taken over geographic areas, but the methodology extends without modification to arbitrary subgroups such as residents of geographic areas. Below, we will refer to areas.

The MSE is the sum of the variance and the squared bias. To estimate the squared bias, we need to allow for the variance in the estimate of bias, because the expectation of the square of a random quantity is equal to its variance plus the square of its mean. We show that, under certain conditions, the point estimate of difference in expected loss between the unadjusted census and the adjusted census estimates does not need to incorporate an allowance for variances in the estimates of bias of the adjustment factors.

In this section, we provide an overview of logic of the analysis. In section 2 we describe the total error model and its implementation for the loss function analysis. Two different methods may be used to compute variances for use in the loss function analysis. One method, used in sections 2 and 3, below, is to store replicates; in some cases the replicates are based on sample replicates (as in jackknife or other pseudo-replication estimation of variance) and in other cases they are based on use of alternative models or assumptions. A second method, described in sections 4 and 5, is to use estimated variance-covariance matrices as the basis for deriving the variances and covariances needed in the loss function analysis. This method has been used in the past but it is somewhat cumbersome in the current situation, when the matrices may have dimensions of $10^4 \times 10^4$ or even greater. *I do not expect that the method described in sections 4 and 5 will be used for the loss function analysis of the A.C.E. Revision II estimates in 2002.* The description is included here for completeness.

To make clear the logic of the analysis, consider any single area or subgroup and let the following notation refer to population level or share, as the case may be. Here we suppress the subscript i for simplicity.

θ true quantity

C census count (unadjusted)

D adjusted estimate

υ $= \theta - C$, net undercount

U $= D - C$ is the estimate of υ

V_D variance of D

$\hat{V}_D$ estimate of V_D

β bias of D

B estimate of β

V_B variance of B

$\hat{V}_B$ estimate of V_B

V_{BD} covariance of B and D

$\hat{V}_{BD}$ estimate of V_{BD}

The MSE for the unadjusted census is υ^2, i.e., net undercount squared. If B is an unbiased estimator of β, then an unbiased estimate of υ is given by U - B. Recall that the expected value of the square of a random variable is equal to the sum of its variance and the square of its expectation. The variance of U - B is $V_D + V_B - 2V_{BD}$ and thus the expected value of $(U - B)^2$ is $\upsilon^2 + V_D + V_B - 2V_{BD}$. We therefore estimate the MSE for the unadjusted census by

(1) $$(U - B)^2 - (\hat{V}_B + \hat{V}_D - 2\hat{V}_{BD}).$$

The expected value of the estimator (1) is $[\upsilon - (EB - \beta)]^2 + (V_B - E(\hat{V}_B)) +$

$(E(\hat{V}_D) - V_D) - 2(E(\hat{V}_{BD}) - V_{BD})$. Thus, if B is an unbiased estimator of β and

$\hat{V}_B + \hat{V}_D - 2\hat{V}_{BD}$ is an unbiased estimator of $V_B + V_D - 2V_{BD}$, the estimator of MSE for the

census given by (1) is unbiased. Note that if the covariance V_{BD} is estimated to be negligible, then $\hat{V}_{BD}$ drops out of (1), and the estimator of MSE of the unadjusted census simplifies to

(2) $$(U - B)^2 - (\hat{V}_B + \hat{V}_D).$$

The MSE for the adjusted estimate, D, is $\beta^2 + V_D$. To estimate this MSE we use

(3) $$B^2 - \hat{V}_B + \hat{V}_D,$$

which is unbiased if B is an unbiased estimator of β and $\hat{V}_D - \hat{V}_B$ is an unbiased estimator of $V_D - V_B$.

The excess MSE for the unadjusted census (C) relative to the adjusted estimate (D) is $\upsilon^2 - (\beta^2 + V_D)$, which we may estimate by (1) minus (3), or simply

(4) $$(U - B)^2 - B^2 - 2(\hat{V}_D - \hat{V}_{BD}).$$

If the covariance V_{BD} is estimated to be negligible, then we may use (2) instead of (1) and estimate the excess MSE by

(5) $$(U - B)^2 - B^2 - 2\hat{V}_D.$$

In this case, the point estimate for difference in expected loss between the adjusted and unadjusted census does not need to incorporate an allowance for $\hat{V}_B$. Previous loss function analyses have assumed that V_{BD} would be relatively small and could be ignored. The method

described in section 3, below, for estimating MSEs does not rely on an assumption that V_{BD} is negligible.

The following sections describe the calculation of expected loss in more detail. We use "area" as a general concept, and some care may be needed in practice. For example, if focusing numbers in a demographic group in an area, the area should be taken to exclude the groups not of interest.

2. Total Error Model

2.1. Overview

The following sources of error are considered in the total error model for the DSE:

a. Sampling variance (section 2.5.1)

b. Ratio-estimator bias (sections 2.5.2-2.5.3)

c. Bias due to inconsistency in the E-sample and P-sample reporting of the characteristics used in assigning the poststrata (sections 2.5.2-2.5.3)

d. Bias from error in identifying P-sample and E-sample cases that are duplicate enumerations in the census (sections 2.5.2-2.5.3)

e. Bias from error using inmovers for outmovers in PES-C (sections 2.5.2-2.5.3)

f. Correlation bias (sections 2.5.2-2.5.3)

g. Error due to choice of imputation model (section 2.5.4)

h. Error due to choice of modeling assumptions concerning duplication probabilities and duplicate "survival probabilities" (section 2.5.5)

2.2. Input Variables from Production

The following variables are produced by the census and the A.C.E. for population estimation and are inputs for the loss function analysis. We note that the post-strata used for adjusting for erroneous enumerations and for undercoverage will not be the same; we use the term post-stratum to refer to an estimation cell that might involve a different E-sample and P-sample poststratum.

N_i census count (unadjusted estimate) for area i

H number of poststrata (or estimation cells)

H_E number of E-sample poststrata

H_P number of P-sample poststrata

C_{ih} census count for area i, poststratum h, against which adjustment factor is applied, $1 \le h \le H$

$C^E_{i\,h}$ census count for area i, E-sample poststratum h, against which E-sample adjustment factor is applied, $1 \le h \le H_E$

$C^P_{i\,h}$ census count for area i, P-sample poststratum h, against which P-sample adjustment factor is applied, $1 \le h \le H_P$

C_i vector of area i census counts across poststrata, to be multiplied by adjustment factors $= (C_{i1}, \ldots, C_{iH})^T$

C^E_i vector of area i census counts across E-sample poststrata, to be multiplied by E-sample adjustment factors $= (C^E_{i\,1}, \ldots, C^E_{i\,HE})^T$

C^P_i vector of area i census counts across P-sample poststrata, to be multiplied by P-sample adjustment factors

$= (C^P_{i\,1}, \ldots, C^P_{i\,HP})^T$

f^E_j adjustment factor for E-sample poststratum j, $1 \le j \le H_E$. Note: these are assumed to include adjustments for II cases.

f^P_k adjustment factor for P-sample poststratum k, $1 \le k \le H_P$

f_h adjustment factor for poststratum h based on E-sample poststratum j and P-sample poststratum k

$= f^E_j f^P_k$

f^E vector of E-sample adjustment factors for E-sample poststrata

f^P vector of P-sample adjustment factors for P-sample poststrata

f vector of adjustment factors for poststrata

$= (f_1, \ldots, f_H)^T$

D_i adjusted estimate for area i

$= f^T C_i$

2.3. Bias Estimates

The following variables related to bias will be produced during the total error analysis.

b_h estimated bias in f_h

b estimated bias of f

$= (b_1, \ldots, b_H)^T$

B_i estimated bias in adjusted estimate for area i

$= b^T C_i$

T_i "target" estimate for area i

$= (f - b)^T C_i = D_i - B_i$

The vector b will be estimated as the sum of two components, $b = b_{meas} + b_{dup\text{-}modeling}$, with

b_{meas} estimate of net bias due to errors (b) - (f) in section 2.1.; see section 2.5.2,

$b_{dup\text{-}modeling}$ estimate of net bias from error (h) in section 2.1; see section 2.5.5.

2.4. Variance Estimates

Estimates of variance will be produced for D_i, T_i. and B_i. Two methods of estimation will be considered. Primarily, we will consider the use of replicates to develop the variance estimates for levels and shares. That method will avoid explicit use of a variance-covariance matrix. A second method will be described in section 4 that will explicitly use a variance-covariance matrix.

$\hat{V}_{sampling}(D_i)$ estimate of variance due to error (a) in section 2.1; see section 2.5.1

$\hat{V}(B_i)$ estimate of variance due to error (a) in section 2.1; see section 2.5.3

$\hat{V}_{sampling}(T_i)$ estimate of variance due to error (a) in section 2.1; see section 2.5.3

$\hat{V}_{imputation}(D_i)$ estimate of variance due to error (g) in section 2.1; see section 2.5.4

$\hat{V}_{dup-modeling}(D_i)$ estimate of variance due to error (h) in section 2.1; see section 2.5.5

The following overall variances are estimated as in section 2.5.6.

$$\hat{V}(D_i) = \hat{V}_{sampling}(D_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i)$$

$$\hat{V}(B_i) = \hat{V}_{sampling}(B_i)$$

$$\hat{V}(T_i) = \hat{V}_{sampling}(T_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i).$$

2.5. Summary Statistics Used to Develop Bias and Variance Estimates

2.5.1. Sampling Variance

The Census Bureau will prepare K replicates of the factors, based on sample replicates. It is possible that the factors may be vectors with full vector for f, or alternatively that the vectors will consist of a subvector of P-sample factors and a subvector of E-sample factors – the latter will involve less computer storage. The replicates of the factors may be used to compute variances of f and D_i's.

To generate the sampling variance of D_i, one would compute K estimates, say $D_{i(k)}$, $1 \leq k \leq K$, and derive the variance estimate accordingly, say $\hat{V}_{sampling}(D_i)$. (This same technique applies whether D refers to a population level or a share.)

2.5.2. Point Estimates of Bias Related to Data, Bias Related to Sampling, and Correlation Bias

There are a number of sources of bias in f. Mulry (2002) describes the estimation of error due to inconsistency in the E-sample and P-sample reporting of the characteristics used in assigning the poststrata, the error in identifying P-sample and E-sample cases that are duplicate enumerations in the census, error due to using inmovers for outmovers in PES-C, ratio estimator bias, and correlation bias. (If the A.C.E. Revision II estimates incorporate adjustments for

correlation bias, the remaining correlation bias will be assumed to be negligible.) An estimate of b reflecting those sources of error will be produced by taking the difference between the production f and a version of f adjusted for the errors described earlier in this paragraph; the estimate will be denoted by b_{meas}.

2.5.3. Sampling Error of Point Estimates of Bias Related to Data and Sampling

A set of 32 sample replicates for use with the simple jackknife procedure has been developed at the Bureau by Katie Bench. Corresponding to each replicate, a value of f and a value of b_{meas} will be computed. The replicates may be used to compute sampling variances of b and T_i's. They may also be used to compute variances of f and D_i's, as an alternative to the replicates discussed in section 2.5.1 above. The replicates do not account for uncertainty in correlation bias estimates.

To generate the sampling variance of B_i and T_i, use the 32 replicates to develop $B_{i[k]}$, $D_{i[k]}$, and $T_{i[k]} = D_{i[k]} - B_{i[k]}$, $1 \le k \le 32$. Then derive the variance estimates accordingly, say $\tilde{V}_{sampling}(B_i)$, $\tilde{V}_{sampling}(D_i)$, $\tilde{V}_{sampling}(T_i)$. We will ratio-adjust these in accordance with the variance estimate of D_i based on K replicates,

$$\hat{V}_{sampling}(T_i) = \lambda_i \tilde{V}_{sampling}(T_i) \text{ and } \hat{V}_{sampling}(B_i) = \lambda_i \tilde{V}_{sampling}(B_i),$$

with $\lambda_i = \hat{V}_{sampling}(D_i)/\tilde{V}_{sampling}(D_i)$. (The purpose of λ_i is to ratio-adjust the variance estimates for D_i using as controls the variance estimates based on larger numbers of replicates, while ensuring consistency among the sampling variance estimates for D_i, T_i, and B_i.)

2.5.4. Error from Choice of Imputation Model

Spencer (2002, section 3, step 7) provides for the construction of replicates that reflect error due to choice of imputation model. There are 128 replicates of vectors of factors, which we will denote by $f_{impute(k)}$, $1 \le k \le 128$. Note: it is assumed that the vectors of factors include adjustments for II cases.

To generate the variance of D_i from choice of imputation model, one would first compute 128 estimates of $D_{i,}$ one from each of $f_{impute(k)}$, say $\mathbf{D}_{\mathbf{i(k)}}^{\mathbf{imp}}$, $1 \le k \le 128$. For example, if $f_{impute(k)}$ is a vector referring to the adjustment factors and D_i refers to a population level for area i, one sets $\mathbf{D}_{\mathbf{i(k)}}^{\mathbf{imp}} = (f^{impute(k)})^T C_i$ and if D_i refers to a population share, one sets $\mathbf{D}_{\mathbf{i(k)}}^{\mathbf{imp}} = (f^{impute(k)})^T C_i / \Sigma_j (f^{impute(k)})^T C_j$. One would then derive the variance estimate accordingly, say $\hat{\mathbf{V}}_{\mathbf{imputation}}(\mathbf{D_i}) = \sum_{\mathbf{k=1}}^{\mathbf{128}} (\mathbf{D}_{\mathbf{i(k)}}^{\mathbf{imp}} - \bar{\mathbf{D}}_{\mathbf{i}}^{\mathbf{imp}})^2/\mathbf{127}$, with $\bar{\mathbf{D}}_{\mathbf{i}}^{\mathbf{imp}} = \sum_{\mathbf{k=1}}^{\mathbf{128}} \mathbf{D}_{\mathbf{i(k)}}^{\mathbf{imp}}/\mathbf{128}$. (This same technique applies whether D refers to a population level or a share.)

2.5.5. Failure of Assumptions in Modeling Duplication

An additional source of bias arises from error in the modeling assumptions concerning duplication probabilities and duplicate "survival probabilities" in the A.C.E. Revision II estimates (See Kostanich (2003), Chapter 6). To reflect this source of error, L alternative modeling assumptions will be used to generate alternative estimates of f, say $f_{dup\text{-}modeling(\ell)}$, $1 \le \ell \le L$. (This is to occur during the production of the production factors, f.) The hypothetical bias due to modeling error in the production estimate when the alternative model k is true is

$b_{dup\text{-}modeling(\ell)} = f - f_{dup\text{-}modeling(\ell)}$.

It is reasonable to consider that no particular alternative model is correct, but still that the model used in the production estimate is incorrect. In this case, the modeling error may be treated as a random bias, in a manner similar to the treatment of the failure of the model underlying imputation of unresolved match, correct-enumeration, or residency status (Spencer 2002), and $b_{dup\text{-}modeling(\ell)}$ will be set to zero. To generate the variance of D_i from failure of assumptions for modeling duplication, one would take the L estimates, $f_{dup\text{-}modeling(\ell)}$, $1 \le \ell \le L$, compute the corresponding L values of D_i, say $\mathbf{D_{i(\ell)}^{dup}}$, $1 \le \ell \le L$, and derive the variance estimate accordingly, say and derive the variance estimate accordingly, say

$\hat{V}_{dup-modeling}(D_i) = \sum_{\ell=1}^{L} (D_{i(\ell)}^{dup} - \bar{D}_i^{dup})^2/(L-1)$, with $\bar{D}_i^{dup} = \sum_{\ell=1}^{L} D_{i(\ell)}^{dup}/L$. (This same technique applies whether D refers to a population level or a share.)

2.5.6. Overall Variances

The overall variance of D_i is the sum of the sampling variance, variance from choice of models of duplication, and variance from choice of imputation model. Set

$\hat{V}(D_i) = \hat{V}_{sampling}(D_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i)$.

The variance of B_i is estimated as $\hat{V}(B_i) = \hat{V}_{sampling}(B_i)$.

The variance of T_i is estimated by the sum of the sampling variance of T_i, the variance in D_i from choice of models of duplication, and variance in D_i from choice of imputation model. Set $\hat{V}(T_i) = \hat{V}_{sampling}(T_i) + \hat{V}_{dup-modeling}(D_i) + \hat{V}_{imputation}(D_i)$.

3. Loss Function Calculations Based on Replicates

Aggregate loss functions for levels and shares are based on weighted sums and differences of $M_{C,i}$ and $M_{D,i}$. Define

$M_{C,i}$ Mean-Square Error (MSE) for Unadjusted Estimate of Level or Share for Area i

$M_{D,i}$ Mean-Square Error (MSE) for Adjusted Estimate of Level or Share for Area i.

$\hat{M}_{C,i} = (N_i - T_i)^2 - \hat{V}(T_i)$.

$\hat{M}_{D,i} = B_i^2 + \hat{V}(D_i) - \hat{V}(B_i)$.

These estimates may be used to estimate the corresponding MSEs in loss functions.

4. Variance-Covariance Matrices

Define the following variance-covariance matrices for adjustment factors.

$\Sigma_{f,sampling}$ estimated sampling variance-covariance matrix of f, of dimension H × H

$\Sigma_{f,imputation}$ estimated variance-covariance matrix of f, of dimension H × H, reflecting error due to choice of imputation models for unresolved status.

$\Sigma_{f,dup\text{-}modeling}$ estimated variance-covariance matrix of f, of dimension H × H, reflecting error due to choice of modeling assumptions concerning duplication probabilities and duplicate "survival probabilities". It is possible that $\Sigma_{f,dup\text{-}modeling}$ will be set to zero.

Σ_f estimated variance-covariance matrix of f, of dimension H × H

$$= \Sigma_{f,sampling} + \Sigma_{f,imputation} + \Sigma_{f,dup\text{-}modeling}$$

It is possible that $\Sigma_{f,sampling}$ will be provided (e.g., by Douglas Olson), in which case Σ_f can be computed as described above. Alternatively, it may be estimated from the K replicates described in section 2.5.1. As a final, although less precise method, $\Sigma_{f,sampling}$ may be estimable from the 32 replicates of f described in section 2.5.3.

The variance-covariance matrix for imputation error will be developed as described in Spencer (2002, section 2, step 7), with one possible modification. If the variance-covariance matrix described there is for the poststratum-level DSEs, the jk entry in the matrix must be divided by $C_{.j}C_{.k}$, with the subscript "." denoting national level census count for the poststratum. The matrix may also be estimated from the 128 replicates described in section 2.5.4.

The variance-covariance matrix $\Sigma_{f,dup\text{-}modeling}$ may be estimated from the vectors $f_{dup\text{-}modeling(\ell)}$, $1 \le \ell \le L$, as described in section 2.5.5. Weighted moments may also be considered, if there is reason to consider weighting some of the alternatives more than others. We will set either (or both) of $\Sigma_{f,dup\text{-}modeling}$ and $b_{dup\text{-}modeling}$ to zero – model bias will be taken to be random (with mean zero) or fixed (with mean either non-zero or zero).

Sampling error, error due to choice of imputation model, and error due to choice of modeling assumptions concerning duplication probabilities and duplicate "survival probabilities" are taken to be independent.

Define the following variance-covariance matrices involving bias estimates for adjustment factors.

Σ_b estimated variance-covariance matrix of b, of dimension H × H

Σ_{fb} estimated cross-covariance matrix of f and b, of dimension H × H

$$\Sigma = \begin{pmatrix} \Sigma_f & \Sigma_{fb} \\ \Sigma_{fb} & \Sigma_b \end{pmatrix}, \text{ of dimension } 2H \times 2H$$

Σ_{f-b} estimated variance-covariance matrix of f - b

$= \Sigma_f + \Sigma_b - 2\Sigma_{fb}$

The source for estimating Σ_b will be 32 sample replicates developed by Katie Bench. A collection of 32 values of b will be computed with respect to the replicates, and the variance-covariance matrix computed accordingly. At the same time, 32 values of f will be computed, one per replicate, and variance-covariance matrices $\Sigma_{f,sample}$ and Σ_{fb} will be computed. (If $\Sigma_{f,sample}$ was developed separately (e.g., by Douglas Olson), as described above, it may be desirable to compute a correlation matrix, say R_{fb} by from the replicate-based estimates, Σ_b, $\Sigma_{f,sample}$, and Σ_{fb}, and then to re-estimate Σ_{fb} by using the R_{fb} and Σ_b along with the original $\Sigma_{f,sample}$. The matrix R_{fb} will be of independent interest, because if its (off-diagonal entries) are small enough then we can use (2) and (5) instead of the more complicated (1) and (4).

5. Loss Function Analysis Based on Variance-Covariance Matrices

When we are estimating loss functions for shares based on variance-covariance matrices, the calculations are more complex than in section 3.

5.1. Loss Functions for Levels

Aggregate loss functions for levels are based on weighted sums and differences of $M_{C,i}$ and $M_{D,i}$. Now we use the following estimates for levels.

$$\hat{M}_{C,i} = (N_i - T_i)^2 - C_i^T \Sigma_{f-b} C_i$$

$$\hat{M}_{D,i} = B_i^2 + C_i^T \Sigma_f C_i - C_i^T \Sigma_b C_i$$

Aggregate loss functions for levels are based on weighted sums and differences of $M_{C,i}$ and $M_{A,i}$.

5.2. Loss Functions for Shares

Consider area i's share of aggregation G, where G is a union of areas. The unadjusted share is $N_{share,i} = N_i / \sum_{j \in G} N_j$. The adjusted share is $D_{share,i} = D_i / \sum_{j \in G} D_j$. The target share is $T_{share,i} = T_i / \sum_{j \in G} T_j$. The bias of the adjusted share is estimated by $B_{share,i} = D_{share,i} - T_{share,i}$.

5.2.1. Replications

For estimating variances of shares, we use replicates. As described below, only one set of Q replicates of f and b need to be generated. That set will service all of the loss functions for shares when the same specification of underlying variances is used. The value of Q is initially set at 1000.

Generate Q 2H×1 vectors $z^{(q)}$, $1 \le q \le Q$, from a multivariate normal distribution with mean zero and variance-covariance matrix Σ. Let the vector $x^{(q)}$ denote vector of the first H components of $z^{(q)}$ and let the vector $y^{(q)}$ denote the remaining H components of $z^{(q)}$; in other words, consider $z^{(q)}$ as a two H×1 vectors stacked on each other,

$$\mathbf{z}^{(q)} = \begin{pmatrix} \mathbf{x}^{(q)} \\ --- \\ \mathbf{y}^{(q)} \end{pmatrix}.$$

Observe that $x^{(q)}$ is distributed as the random error in f, $y^{(q)}$ is distributed as the random error in b, and the covariance between $x^{(q)}$ and $y^{(q)}$ is Σ_{fb}.

Define replicates of the adjustment factors f and bias estimates b by $f^{(q)} = f + x^{(q)}$ and $b^{(q)} = b + y^{(q)}$ for $1 \le q \le Q$. Replicates of adjusted estimates of shares and target values of shares are based on the replicates $f^{(q)}$ and $b^{(q)}$, and the variances and covariances are derived from the replicates.

Specifically, notice that the adjusted share for area i may be written as $D_{share,i} = \mathbf{f}^T\mathbf{C}_i / \sum_{j \in G} \mathbf{f}^T\mathbf{C}_j$. The q-th replicate of the adjusted share is

$$\mathbf{D}^{(q)}_{share,i} = \mathbf{f}^{(q)T}\mathbf{C}_i / \sum_{j \in G} \mathbf{f}^{(q)T}\mathbf{C}_j.$$

The sample variance among the Q values of $\mathbf{D}^{(r)}_{share,i}$ is used to estimate the variance of $D_{share,i}$. Denote the variance estimate by $V_{D,share,i}$.

Similarly, the q-th replicate of the target share is defined by

$$\mathbf{T}_{\mathbf{share,i}}^{(q)} = (\mathbf{f}^{(q)} - \mathbf{b}^{(q)})^T \mathbf{C}_i / \sum_{j \in G} (\mathbf{f}^{(q)} - \mathbf{b}^{(q)})^T \mathbf{C}_j .$$

The sample variance among the Q values of $\mathbf{T}_{\mathbf{share,i}}^{(q)}$ is used to estimate the variance of $T_{share,i}$. Denote the variance estimate by $V_{T,share,i}$.

The q-th replicate of the bias in the adjusted share is defined by $\mathbf{B}_{\mathbf{share,i}}^{(q)} = \mathbf{D}_{\mathbf{share,i}}^{(q)} - \mathbf{T}_{\mathbf{share,i}}^{(q)}$. The sample variance among the Q values of $\mathbf{B}_{\mathbf{share,i}}^{(q)}$ is used to estimate the variance of $B_{share,i}$. Denote the variance estimate by $V_{B,share,i}$.

5.2.3 MSEs for Shares

The MSE for the unadjusted estimate of share for area i is estimated by

$$\hat{\mathbf{M}}_{\mathbf{C,share,i}} = (N_{share,i} - T_{share,i})^2 - V_{T,share,i}$$

and the MSE for the adjusted estimate of share for area i is estimated by

$$\hat{\mathbf{M}}_{\mathbf{D,share,i}} = B_{share,i}^{\ 2} + V_{A,share,i} - V_{B,share,i}.$$

If there is zero correlation between f and b, then in $\hat{\mathbf{M}}_{\mathbf{C,share,i}}$ we may replace $V_{T,share,i}$ by the sum, $V_{D,share,i} + V_{B,share,i}$. The aggregate loss functions for shares are based on weighted sums and differences of $M_{C,share,i}$ and $M_{D,,sharei}$.

References

Kostanich, D. (2003), "A.C.E. Revision II: Design and Methodology," DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-30, U.S. Bureau of the Census, Washington, DC.

Mulry, M. H. (2002) "Estimating Bias in the A.C.E. Revision II and Forming Confidence Intervals." (Appendix A of this document.)

Spencer, B.D. (2002) Draft report, "Report on Missing Data Evaluation," October 5, 2002. Prepared by Abt Associates Inc. and Spencer Statistics, Inc. for the Bureau of the Census, Activity 20 - Deliverable 4, Task Number 46-YABC-7-00001, under contract no. 50-YABC-7-66020

APPENDIX C

Bias Estimation and Loss Function Analysis for A.C.E. Revision II.

Randy ZuWallack

I. Introduction

The purpose of this document is to document the computer design for calculating loss functions as described in Bruce Spencer's October 21, 2002 draft (now Appendix B), "Total Error and Loss Function Analysis for the A.C.E. Revision II Estimates of Population." The loss functions are weighted sums of mean squared errors over the groupings of geographic areas below. Variance estimates are calculated using replication methods developed from several different sources, which are discussed in the Inputs section.

We will look at loss functions of both population levels and shares for geographic groupings of states, counties and places. The design below describes the analysis for a level. The calculations for a share are equivalent to that of a level except the estimates in each area are relative to the total of the estimates over all areas.

The geographic groupings of interest are:

Levels: All counties with population of 100,000 or less

All counties with population greater than 100,000

All places with population at least 25,000 but less than 50,000

All places with population at least 50,000 but less than 100,000

All places with population greater than 100,000

Shares within state: All counties

All places

Shares within U.S.: All places with population at least 25,000 but less than 50,000

All places with population at least 50,000 but less than 100,000

All places with population greater than 100,000

All states

The design below is general enough to encompass all the above geographic groupings as well as any others. The only piece of information that varies with the above data groupings is the level of geography for the poststratified census counts as described in the input section.

II. Loss Function analysis

To calculate the average loss, $\bar{L}$, for the geographic groupings above, we estimate the average mean squared error, $\hat{M}$, over all areas in the geographic grouping for both the Census and the DSE:

$$\bar{L}_C = \sum_i \frac{\hat{M}_{C,i}}{C_i} \text{ and } \quad \bar{L}_{DSE} = \sum_i \frac{\hat{M}_{DSE,i}}{C_i}$$

To do this, we need to first estimate the mean squared error for the areas in the data groupings.

Let C_i = Census count for area i

D_i = DSE for area i

B_i = Estimated bias of DSE for area I

$T_i = D_i - B_i$ = 'Target' estimate for area i

Estimated MSE of the census count for area i is $\hat{M}_{C,i} = (C_i - T_i)^2 - \hat{V}(T_i)$, and the estimated MSE of the DSE for area i is $\hat{M}_{DSE,i} = B_i^2 + \hat{V}(D_i) - \hat{V}(B_i)$. The variance terms for the target and the DSE are further broken down into three contributors: sampling, imputation, and duplication modeling. These variance contributors are assessed individually and then pieced together to form the two terms, $\hat{V}(D_i) = \hat{V}_{sampling}(D_i) + \hat{V}_{imputation}(D_i) + \hat{V}_{dup-modeling}(D_i)$ and

$\hat{V}(T_i) = \hat{V}_{sampling}(T_i) + \hat{V}_{imputation}(D_i) + \hat{V}_{dup-modeling}(D_i)$. The calculations for estimating these pieces as well as are discussed in the sections that follow.

The equations above are applicable for levels and shares alike, as are the variance estimation calculations in the sections that follow. The difference is the definitions of D_i, B_i, and T_i. For

population levels, $D_i = \sum_h D_{hi}$, $B_i = \sum_h B_{hi}$, and $T_i = \sum_h T_{hi}$, where h denotes the poststrata.

For population shares, $D_i = \sum_h D_{hi} \Big/ \sum_i \sum_h D_{hi}$, $B_i = \sum_h B_{hi} \Big/ \sum_i \sum_h B_{hi}$, and

$T_i = \sum_h T_{hi} \Big/ \sum_i \sum_h T_{hi}$.

III. Inputs

1) Census counts - A file containing poststratified census counts for all states, counties and places. There is one record per geographic area and one variable per poststrata. Randy ZuWallack will create this file from the poststratified micro-level census file.

2) Replicate CCFs used for DSE variances - A file of replicate CCFs used for DSE variance estimation. There is one record per replicate and one variable per poststrata. The double sampling factor for each replicate used in the variance estimation is also included on this file. Doug Olson will create this file as part of his DSE variance estimation system.

3) Replicate CCFs used for measurement bias estimates - A file of replicate CCFs used for measurement bias estimation. There is one record per replicate and one variable per poststrata. Katie Bench will create this file as part of her measurement bias estimation system.

4) Replicate measurement bias estimates - A file of replicate bias estimates used for measurement bias estimation. There is one record per replicate and one variable per poststrata. Katie Bench will create this file as part of her measurement bias estimation system.

5) Replicate CCFs used for imputation variance estimation - A file of replicate CCFs used for imputation variance estimation. There is one record per replicate and one variable per poststrata. Anne Kearney will create this file as part of her imputation variance estimation system.

6) Alternative CCFs due to duplicatation modeling assumptions - A file of CCFs based on various duplicatation modeling assumptions. There is one record per alternative and one variable per poststrata. Eric Schindler will create this file.

Based on the above input files, define the following matrices:

C $=[C_{ih}]_{I\times H}$, a matrix of census counts where H is the number of poststrata and I is the number of geographic areas in the grouping (states, counties, or places). This is taken from input file 1.

f $= [f_h]_{1\times H,}$ a vector of CCFs where H is the number of poststrata. Use replicate 0 from input file 2 to form this vector.

$\mathbf{f_{(K)}} = [f_{(k)h}]_{K\times H,}$ a matrix of replicate CCFs where H is the number of poststrata and K is the number of replicates. This is taken from input file 2.

DS $= [DS_i]_{K\times K}$, diagonal matrix of K double sampling factors. This is taken from input file 2.

$\mathbf{f}_{(J)} = [f_{(j)h}]_{J\times H,}$ a matrix of replicate CCFs where H is the number of poststrata and J is the number of replicates. This is taken from input file 3.

$\mathbf{b} = [b_h]_{1\times H,}$ a matrix of bias estimates where H is the number of poststrata. Use replicate 0 from input file 4 to form this vector.

$\mathbf{b}_{(J)} = [b_{(j)h}]_{J\times H,}$ a matrix of replicate bias estimates where H is the number of poststrata and J is the number of replicates. This is taken from input file 4.

$\mathbf{f}_{(m)} = [f_{(m)h}]_{M\times H,}$ a matrix of replicate CCFs where H is the number of poststrata and M is the number of replicates. This is taken from input file 5.

$\mathbf{f}_{(L)} = [f_{(l)h}]_{L\times H,}$ a matrix of alternative CCFs where H is the number of poststrata and L is the number of alternatives. This is taken from input file 5.

IV. Sampling variance

For all I geographic areas, the sampling variance needs to be calculated for three terms, the DSE, the estimated measurement bias and the estimated target. We are using two sets of replicates to estimate these three components. The first set of replicates was developed during the estimation of direct DSE variances at the poststrata level. The second, much smaller set was developed for the purpose of estimating measurement bias. For consistency, variance estimates based on the second set of replicates are ratio adjusted to conform to the larger set of replicates.

First, we estimate the sampling variance of the DSE for area I, D_i, using the simple Jackknife replication formula with the double sampling factor DS_k applied to each replicate:

$$\hat{V}_{sampling}(D_i) = \frac{K-1}{K}\sum_{k=1}^{K} DS_k (D_{i(k)} - \overline{D}_i)^2, \text{ where } \overline{D}_i = (1/K)\sum_{k=1}^{K} D_{i(k)}$$

Using matrix calculations, we estimate this variance component for all I geographic areas as follows:

Let $\mathbf{D_{(K)}} = [D_{i(k)}]_{I\times K}$, matrix of K replicated DSEs used in variance estimation for I areas

$\mathbf{COVD_{(K)}} = [cov_{samp}(D_i, D_{i'})]_{I\times I}$, estimated sampling covariance matrix for I area DSEs.

1) Calculate $\mathbf{D_{(K)}} = \begin{cases} \mathbf{Cf}'_{\mathbf{(K)}} \text{ if a level} \\ \mathbf{Cf}'_{\mathbf{(K)}}\,\mathrm{DIAG}(\mathbf{1}'\mathbf{Cf}'_{\mathbf{(K)}})^{-1}, \text{ where } \mathbf{1}' = (1,1,\ldots,1)_{1\times I} \text{ if a share} \end{cases}$

   Note: DIAG is a function that creates a diagonal matrix from the elements of a vector.

2) Calculate $\mathbf{COVD_{(K)}} = (1\text{-}1/K)(\mathbf{D_{(K)}} - (1/K)\mathbf{D_{(K)}}\mathbf{11}')\mathbf{DS}(\mathbf{D_{(K)}} - (1/K)\mathbf{D_{(K)}}\mathbf{11}')'$, where $\mathbf{1'}=(1, 1, \ldots, 1)_{1\times K}$
3) Create $\mathbf{VD_{(K)}}$, by extracting the $\hat{V}_{sampling}(D_i)$ terms for all I geographic areas from the diagonal of $\mathbf{COVD_{(K)}}$.

Next, we estimate the sampling variance of the measurement bias and target estimate for area I, B_i and T_i respectively, using the simple Jackknife replication formula and then ratio adjusting the estimates to concur with $\hat{V}_{sampling}(D_i)$:

$$\hat{V}_{sampling}(T_i) = \lambda_i \tilde{V}_{sampling}(T_i) = \lambda_i \frac{J-1}{J}\sum_{j=1}^{J} (T_{i(j)} - \overline{T}_i)^2, \text{ where } \overline{T}_i = (1/J)\sum_{j=1}^{J} T_{i(j)}$$

$$\hat{V}_{sampling}(B_i) = \lambda_i \tilde{V}_{sampling}(B_i) = \lambda_i \frac{J-1}{J}\sum_{j=1}^{J}(B_{i(j)} - \bar{B}_i)^2, \text{where} \bar{B}_i = (1/J)\sum_{j=1}^{J} B_{i(j)}$$ and

$$\lambda_i = \frac{\hat{V}_{sampling}(D_i)}{\frac{J-1}{J}\sum_{j=1}^{J}(D_{i(j)} - \bar{D}_i)^2} = \frac{\hat{V}_{sampling}(D_i)}{\tilde{V}_{sampling}(D_i)}$$

Likewise, using the following matrix calculations, we estimate these variance components for all I geographic areas:

Let $\mathbf{D_{(J)}} = [D_{i(j)}]_{I\times J}$, matrix of J replicated DSEs used in variance estimation for I geographic areas

$\mathbf{COVD_{(J)}} = [\text{cov}_{sampling}(D_i, D_{i'})]_{I\times I}$, estimated sampling covariance matrix for I area DSEs.

$\mathbf{COVB_{(J)}} = [\text{cov}_{sampling}(B_i, B_{i'})]_{I\times I}$, estimated sampling covariance matrix for I areas bias estimates.

$\mathbf{COVT_{(J)}} = [\text{cov}_{sampling}(T_i, T_{i'})]_{I\times I}$, estimated sampling covariance matrix for I area target estimates.

$\mathbf{\Lambda} = [\lambda_i]_{I\times 1}$, ratios for adjusting $\hat{V}_{sampling}(T_i)$ and $\hat{V}_{sampling}(B_i)$ for all I areas.

First, we calculate λ_i for all I areas:

1) Calculate $\mathbf{D_{(J)}} = \begin{cases} \mathbf{Cf}'_{\mathbf{(J)}} & \text{if a level} \\ \mathbf{Cf}'_{\mathbf{(J)}}\text{DIAG}(\mathbf{1}'\mathbf{Cf}'_{\mathbf{(J)}})^{-1}, \text{ where } \mathbf{1}' = (1,1,\ldots,1)_{1\times I} & \text{if a share} \end{cases}$
2) Calculate $\mathbf{COVD_{(J)}} = (1\text{-}1/J)(\mathbf{D_{(J)}} - (1/J)\mathbf{D_{(J)}}\mathbf{11}')(\mathbf{D_{(J)}} - (1/J)\mathbf{D_{(J)}}\mathbf{11}')'$, where $\mathbf{1}' = (1, 1, \ldots, 1)_{1\times J}$
3) Create $\mathbf{VD_{(J)}}$, by extracting the $\tilde{V}_{sampling}(D_i)$ terms for all I geographic areas from the diagonal of $\mathbf{COVD_{(J)}}$.
4) For all I areas, calculate λ_i by dividing the i-th element of $\mathbf{VD_{(K)}}$ by the i-th element of $\mathbf{VD_{(J)}}$. Create $\mathbf{\Lambda} = [\lambda_i]_{I\times 1}$.

Similarly, we calculate $\hat{V}_{sampling}(T_i)$:

1) Calculate $\mathbf{T_{(J)}} = \begin{cases} \mathbf{C(f_{(J)} - b_{(J)})'} \text{ if a level} \\ \mathbf{C(f_{(J)} - b_{(J)})'}\text{DIAG}(\mathbf{1'C(f_{(J)} - b_{(J)})'})^{-1}, \text{where } \mathbf{1'} = (1,1,\ldots,1)_{1\times I} \text{ if a share} \end{cases}$
2) Calculate $\mathbf{COVT_{(J)}} = (1\text{-}1/J)(\mathbf{T_{(J)}} - (1/J)\mathbf{T_{(J)}11'})(\mathbf{T_{(J)}} - (1/J)\mathbf{T_{(J)}11'})'$, where $\mathbf{1'}=(1, 1, \ldots, 1)_{1\times J}$
3) Create $\mathbf{VT_{(J)}^{*}}$, by extracting the $\tilde{V}_{sampling}(T_i)$ terms for all I geographic areas from the diagonal of $\mathbf{COVT_{(J)}}$.
4) Calculate $\mathbf{VT_{(J)}}$ by multiplying the elements of $\mathbf{VT_{(J)}^{*}}$ by the corresponding elements of $\mathbf{\Lambda}$.

Finally, we calculate $\hat{V}_{sampling}(B_i)$ as follows:

1) Calculate $\mathbf{B_{(J)}} = \mathbf{D_{(J)}} - \mathbf{T_{(J)}}$
2) Calculate $\mathbf{COVB_{(J)}} = (1\text{-}1/J)(\mathbf{B_{(J)}} - (1/J)\mathbf{B_{(J)}11'})(\mathbf{B_{(J)}} - (1/J)\mathbf{B_{(J)}11'})'$, where $\mathbf{1'}=(1, 1, \ldots, 1)_{1\times J}$
3) Create $\mathbf{VB_{(J)}^{*}}$, by extracting the $\tilde{V}_{sampling}(B_i)$ terms for all I geographic areas from the diagonal of $\mathbf{COVB_{(J)}}$.
4) Calculate $\mathbf{VB_{(J)}}$ by multiplying the elements of $\mathbf{VB_{(J)}^{*}}$ by the corresponding elements of $\mathbf{\Lambda}$.

V. Imputation variance

To estimate the imputation variance of the DSE for area I, D_i, we are using a set of replicates developed for assessing the error due to the choice of the imputation model. The replication variance formula is:

$$\hat{V}_{imputation}(D_i) = \frac{\sum_{m=1}^{M}(D_{i(m)} - \overline{D}_i)^2}{M-1}, \text{where } \overline{D}_i = (1/M)\sum_{m=1}^{M} D_{i(m)}$$

We estimate this variance component for all I geographic areas using the matrix calculations that follow.

Let $\mathbf{D_{(M)}} = [D_{i(M)}]_{I\times M}$, matrix of M replicated DSEs used in imputation variance estimation for I areas.

$\mathbf{COVD_{(M)}} = [\mathrm{cov}_{imp}(D_i, D_{i'})]_{I\times I}$, estimated imputation covariance matrix for I area DSEs.

1) Calculate $\mathbf{D_{(M)}} = \begin{cases} \mathbf{Cf'_{(M)}} \text{ if a level} \\ \mathbf{Cf'_{(M)}}\,\mathrm{DIAG}(\mathbf{1'Cf'_{(M)}})^{-1}, \text{ where } \mathbf{1'} = (1,1,\ldots,1)_{1\times I} \text{ if a share} \end{cases}$
2) Calculate $\mathbf{COVD_{(M)}} = 1/(M-1)(\mathbf{D_{(M)}} - (1/M)\mathbf{D_{(M)}11'})(\mathbf{D_{(M)}} - (1/M)\mathbf{D_{(M)}11'})'$, where $\mathbf{1'}=(1, 1, \ldots, 1)_{1\times M}$
3) Create $\mathbf{VD_{(M)}}$, by extracting the $\hat{V}_{\text{imputation}}(D_i)$ terms for all I geographic areas from the diagonal of $\mathbf{COVD_{(M)}}$.

VI. Duplication modeling variance

To estimate the duplication modeling variance of the DSE for area I, D_i, we are using a set of estimates generated under different modeling assumptions. The formula for estimating this variance component using the L alternatives is:

$$\hat{V}_{\text{dup-modeling}}(D_i) = \frac{\sum_{l=1}^{L}(D_{i(l)} - \overline{D}_i)^2}{L-1}, \text{ where } \overline{D}_i = (1/L)\sum_{l=1}^{L} D_{i(l)}$$

We estimate this variance component for all I geographic areas using the matrix calculations that follow.

Let $\mathbf{D_{(L)}} = [D_{i(L)}]_{I\times L}$, matrix of L replicated DSEs used in imputation variance estimation for I areas.

$\mathbf{COVD_{(L)}} = [\mathrm{cov}_{\text{dup-modeling}}(D_i, D_{i'})]_{I\times I}$, estimated dup-modeling covariance matrix for I area DSEs.

1) Calculate $\mathbf{D_{(L)}} = \begin{cases} \mathbf{Cf'_{(L)}} \text{ if a level} \\ \mathbf{Cf'_{(L)}}\,\mathrm{DIAG}(\mathbf{1'Cf'_{(L)}})^{-1}, \text{ where } \mathbf{1'} = (1,1,\ldots,1)_{1\times I} \text{ if a share} \end{cases}$
2) Calculate $\mathbf{COVD_{(L)}} = 1/(L\text{-}1)(\mathbf{D_{(L)}} - (1/L)\mathbf{D_{(L)}11'})(\mathbf{D_{(L)}} - (1/L)\mathbf{D_{(L)}11'})'$, where $\mathbf{1'}=(1, 1, \ldots, 1)_{1\times L}$
3) Create $\mathbf{VD_{(L)}}$, by extracting the $\hat{V}_{\text{dup-modeling}}(D_i)$ terms for all I geographic areas from the diagonal of $\mathbf{COVD_{(L)}}$.

VII. Total Error and Average Loss

All components are available for calculating MSEs as described in section II. First, calculate the census counts, the bias estimates, and the targets for the I areas:

$C_i = \mathbf{C_i 1}$, where $\mathbf{C_i}$ is the i^{th} row of $\mathbf{C}$ and $1'=(1, 1,\ldots, 1)_{1\times H}$

$B_i = \mathbf{C_i}\,\mathbf{b}'$

$T_i = \mathbf{C_i}\,\mathbf{(f\text{-}b)}'$

Then, for each of the I elements in the vectors created above and in section II, calculate the estimated MSEs:

$$\hat{M}_{C,i} = (C_i - T_i)^2 - [\hat{V}_{\text{sampling}}(T_i) + \hat{V}_{\text{imputation}}(D_i) + \hat{V}_{\text{dup-modeling}}(D_i)]$$

$$\hat{M}_{DSE,i} = B_i^2 + \hat{V}_{\text{sampling}}(D_i) + \hat{V}_{\text{imputation}}(D_i) + \hat{V}_{\text{dup-modeling}}(D_i) - \hat{V}_{\text{sampling}}(B_i)$$

Finally, calculate the estimated loss of the census count and DSE by weighting the MSE estimates by the inverse of the census count and summing over all I areas:

$$\overline{L}_C = \sum_i \frac{\hat{M}_{C,i}}{C_i} \text{ and } \quad \overline{L}_{DSE} = \sum_i \frac{\hat{M}_{DSE,i}}{C_i}$$