Comparison of Estimates of Small Area Variances

Mark E. Asiala, US Census Bureau

Mark E. Asiala, 2403-2 DSSD, Washington DC 20233-7613

Key Words: coefficient of variation, reliability, variance, Census 2000

1   Introduction

In our planning for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.), simulated data of the expected small-area variances was needed for benchmarking and testing the generalized variance operation. We started with the 1990 Census and Post-Enumeration Survey (PES) data and made adjustments for population growth and sample design changes. With the estimated census counts by post-stratum and a new covariance matrix, we produced synthetic estimates of variance and coefficients of variation (CVs) for state and various sub-state geographical areas. This paper details that process and makes the comparison between our simulated CVs and the actual 2000 A.C.E. estimated CVs.

2   Methodology

The methodology consists of two parts, the first details the simulated CVs and the second details the 2000 A.C.E. estimated CVs.

2.1   Simulation of CVs

There were three fundamental differences between the 1990 PES and the 2000 A.C.E. which affected variances. These are:

1. Sampling methodology, including sample size

2. Post-Stratification variables, including the classification of out of scope persons

3. Census counts

The means for adjusting for differences in sampling methodology and sample size was adapted from previous research (Mule[2], Sands[4]). For each post-stratum group t, an adjustment factor R_t is made

---

This paper reports the results of reseach and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

---

to account for the differences as follows

$$R_t = \frac{\text{Var}_{\text{ACE}}\left(\hat{X}_{tA}\right)}{\text{Var}_{\text{PES}}\left(\hat{X}_{tA}\right)} \approx \frac{1.56 \sum_{r=1}^{11303} n_{rt} w_{ACE,r}^2}{\sum_{j=1}^{n_t} w_{PES,j}^2}$$

The numerator is the sum over all clusters r of the final A.C.E. large block subsample weight $w_{ACE,r}$ squared times the 1998 estimated number of persons n_{rt} for that cluster and post-stratum group. That sum is then multiplied by a factor of 1.56 to approximate the increase in variance due to the Census 2000 Targeted Extended Search which takes a random sample of blocks to perform a surrounding block search. While the 2000 process may be more efficient and yield a reduction in bias over the 1990 procedure, it could introduce a slight increase in the variance. The factor 1.56 was based on preliminary research performed on the 1990 data in preparation for 2000. The actual value is likely to be different once evaluation of the 2000 Targeted Extended Search (TES) is performed. The denomenator is the sum over all n_t 1990 PES E-sample persons in post-stratum group t of the final PES weight $w_{PES,j}$ squared. The resulting factor R_t can then be used later to adjust the covariance matrix.

To account for the differences in post-stratification, we first retabulated the 1990 PES files using the 2000 A.C.E. post-stratification variables. The post-stratification schemes used in 1990 and 2000 are different enough that one cannot simply map from one to the other. Instead, the original variables including the detailed race codes were needed in order to properly post-stratify them. Some difficulties arise, however, because the 2000 post-stratification scheme does not follow the sampling methodology of the 1990 census. This caused problems of having too little sample for some 2000 post-stratum groups. Most notable among these were Non-Hispanic White or Some Other Race, Non-Owner, Low Return Rate Tract, All Other TEAs in the Northeast, South, and West Regions which had no P-sample persons. This, of course, makes it impossible to simulate variances for these post-stratum groups. When we retabulated the 1990 Census files, however, we found very few census people in these post-stratum groups. As a result, we did not collapse over these groups and treated the variance as zero.

There were some difficulties with the Native Hawaiian or Pacific Islander, Non-Hispanic Asian, and American Indian or Alaska Native (AIAN) on reservation domains because of their small population sizes. In 1990 the Native Hawaiian and Asian race groups were combined. When they were separated we found that there was not adequate sample of the Native Hawaiians. Any simulated variance, therefore, would not reflect the increased sampling rate planned for the 2000 A.C.E. Likewise, the American Indians on Reservation Domain by tenure groups also did not have adequate sample, since in 1990 this group was collapsed across tenure. For this reason, we collapsed Native Hawaiians and Asians back together and we also collapsed across tenure for the Reservation Indians.

As a result, we were able to define 60 post-stratum groups (and hence 420 post-strata) out of the 64 defined A.C.E. post-stratum groups. We then retabulated the 1990 census files using these 420 post-strata. We next updated these counts using 1998 demographic estimates. These estimates contain 1998 state population estimates for Non-Hispanic Whites, Non-Hispanic Black, Non-AIAN Hispanic, Non-Hispanic Asian and Pacific Islander (API), and all AIAN. These five basic demographic groups were then crossed by the age/sex post-stratification groups (children under 18, males/females 18–29, males/females 30–49, and males/females 50+). This process creates a growth factor F_{dS} for each of the 35 demographic groups d within each state S defined as follows

$$F_{dS} = \frac{\text{1998 Est Pop for group d, state S}}{\text{1990 Census Count for group d, state S}}$$

To create an estimate for some post-stratum counts by geography in a state S, we would then apply the growth factors F_{dS} appropriate for that state and post-strata.

Using the 1990 PES data with the 2000 A.C.E. post-stratification, a covariance matrix Cov for the coverage correction factors CCF_X by 2000 post-strata X was calculated. A stratified jackknife over the 1990 block clusters was used to calculate the matrix as follows,

$$\text{Cov}\left(CCF_X, CCF_{X'}\right) =$$
$$\frac{1}{C_X C_{X'}} \sum_h \sum_k \frac{n_h - 1}{n_h} \left(\widehat{DSE}_{X,h}^{(k)} - \widehat{DSE}_{X,h}\right)$$
$$\times \left(\widehat{DSE}_{X',h}^{(k)} - \widehat{DSE}_{X',h}\right)$$

where C_X is the 1990 census count for 2000 post-stratum X, n_h is the number of block cluster in

sampling stratum h, $\widehat{DSE}_{X,h}$ is the DSE estimate of A.C.E. post-stratum X in sampling stratum h, and $\widehat{DSE}_{X,h}^{(k)}$ is the kth replicate of the dual system estimate (DSE). The resulting covariance matrix was then adjusted to account for the 2000 sampling methodology by

$$\text{Cov}^*\left(CCF_X, CCF_{X'}\right) = \text{Cov}\left(CCF_X, CCF_{X'}\right)$$
$$\times \sqrt{\frac{\sum_t C_{Y(X)t} R_t}{\sum_t C_{Y(X)t}} \frac{\sum_t C_{Y(X')t} R_t}{\sum_t C_{Y(X')t}}}$$

where $C_{Y(X)t}$ is the 1990 census count of persons in 2000 post-stratum group Y (to which X is a member) and also in 1990 post-stratum group t. This gives us a covariance matrix in terms of 2000 post-strata that accounts for each of the items listed in the beginning of this section.

Using the growth factor $F_{d(X)S}$ and coverage correction factors CCF_X, we can create synthetic total population estimates $\widehat{POP}_A$ for area A (in state S) as follows

$$\widehat{POP}_A = OOS_A + \sum_{X=1}^{420} F_{d(X)S} \, C_{AX} \, CCF_X$$

where OOS_A are persons in area A who are out of scope for the A.C.E. (Group Quarters persons and Remote Alaska) and C_{AX} is the 1990 census count for area A and 2000 post-stratum X.

To calculate small area variances, we use the covariance matrix along with our census counts and growth factors to calculate synthetic variances as follows

$$\text{Var}_{\text{ACE}}\left(\widehat{POP}_A\right) =$$
$$\sum_{X=1}^{420} \sum_{X'=1}^{420} \left(F_{d(X)S} C_{AX}\right)\left(F_{d(X)'S} C_{AX'}\right)$$
$$\times \text{Cov}^*\left(CCF_X, CCF_{X'}\right)$$

The small area estimates have two components, those persons who are out of scope (general quarters persons and remote Alaska) and those who are in scope (and have an assigned post-stratum).

Finally, we define the coefficient of variation CV_A for area A in the usual manner:

$$CV_A = \frac{\sqrt{\text{Var}_{\text{ACE}}\left(\widehat{POP}_A\right)}}{\widehat{POP}_A}$$

2.2 Estimated A.C.E. Variances

This subsection serves to only give a quick overview of the A.C.E. variance estimation process due to its

complexity. For a complete detailed description see Kim, et al.[1]. We concentrate here on describing the differences between the 2000 methodology as compared to the 1990 methodology with special attention to those differences which we did not model.

1. The A.C.E. was a three-phase sample which necessitated a different variance methodology than what was done for the 1990 PES

2. The 2000 DSEs were calculated in a different manner than in 1990 due to differing treatment of movers and non-movers

3. The 2000 A.C.E. variance operation included estimating the variance due to missing data within the replication.

The 2000 sampling methodology had more complexity than in 1990. This arose from first of all having to reduce the original 750,000 housing unit ICM sample to the 300,000 housing unit A.C.E. sample through A.C.E. reduction and through small and large block cluster subsampling. An additional phase was added to the sampling through TES. This made the 1990 method of stratified jackknifing to get the covariance matrix okay as a quick approximation but it did not capture all of the variance. A new method which is developed in Kim, et al.[1] properly calculates the variance. There was no way to simulate the effect of having a three-phase sample using the 1990 data since it had a fundamentally different design.

Secondly, the DSE was defined differently in how it treated movers and non-movers during the A.C.E. operations. This makes for a slightly more complex DSE. The 1990 data was not compatible with the 2000 treatment of movers so this difference also could not be simulated.

Lastly, the 2000 A.C.E. variance operation included estimating the variance due to imputation except for the component due to imputation model selection. Mulry & Spencer[3] estimate that the variance in the 1990 PES due to all imputation is about 6% of the total variance with about 2% of the total variance due to imputation model selection. Subtracting, we could estimate that our simulation may miss about 4% of the total variance which is included in the 2000 A.C.E. variance operation.

3 Results

The study was designed to give a general idea of what to expect from the 2000 A.C.E. variances. The results, therefore, should not be expected to be precise for a specific geographical area but we would expect that the simulation performs well on the whole.

Since we do not want to confound our analysis with our demographic population estimates, all comparisons will be based on CVs rather than absolute variances. The 2000 results are based on data used for Starsinic, et al.[5].

There are four levels of geography on which we make a comparison:

1. State

2. Congressional District

3. Counties with Population Greater than 100,000

4. Places with Population Greater than 100,000

We do not include areas with population of less than 100,000 (including tracts) because the accuracy of the simulation is more suspect for smaller geographies where the variance of the estimates is higher. For each level of geography, the District of Columbia is included for completeness. The results are presented first in a broad fashion followed by more detailed analysis for each geographic level.

3.1 Overall Distributions

Table 1 compares the distribution of the CVs for the simulation to the estimated A.C.E. CVs. While the mean and quartiles are broad measures of distribution, they do show that the majority of the distribution matches rather nicely. The simulation appears to do well for gauging the minimum CV and the first two quartiles. The match worsens for the third quartile and the maximum as the estimated CVs have a heavier tail than their simulated counterparts. This draws the mean CV upwards for the estimated CVs as well.

3.2 State Comparison

The simulated CVs for states show the most discrepancies from the estimated values out of the four geographic levels. Figure 1 shows that the estimated CVs of 10 states are above 0.4% whereas no simulated CVs are above this threshold. Most of the states are in the Northeast or West Census Regions with large non-Mail Out/Mail Back (MO/MB) areas. They include a high percentage of Non-Hispanic White or Some Other Race or high numbers of Native Hawaiians. The post-strata corresponding to these characteristics had higher CVs than what was simulated.

In Figure 2, we see a scatter plot of simulated versus actual CVs. With those 10 high CV states included, the line of best fit is far from being the

Table 1: Comparison Summary Table

Area	Subdivision	No.	Mean Size	Mean CV	Min	Q1	Median	Q3	Max
State	Simulated	51	5,390,781	0.250%	0.143%	0.201%	0.244%	0.296%	0.347%
	Estimated	51	5,582,035	0.310%	0.159%	0.220%	0.240%	0.378%	0.804%
CDs	Simulated	436	630,573	0.311%	0.150%	0.247%	0.290%	0.350%	0.891%
	Estimated	436	653,103	0.330%	0.156%	0.250%	0.297%	0.375%	0.948%
Counties	Simulated	493	414,164	0.343%	0.186%	0.270%	0.315%	0.374%	1.124%
> 100,000	Estimated	524	400,345	0.368%	0.201%	0.274%	0.310%	0.405%	1.498%
Place	Simulated	236	330,509	0.355%	0.222%	0.291%	0.330%	0.380%	0.936%
> 100,000	Estimated	245	315,037	0.343%	0.213%	0.283%	0.314%	0.361%	1.435%

optimal line where the simulated value is equal to the estimated value. If one removes the 10 high CV states, however, the model fit improves but we still see that the simulated values tend to underestimate the variance.
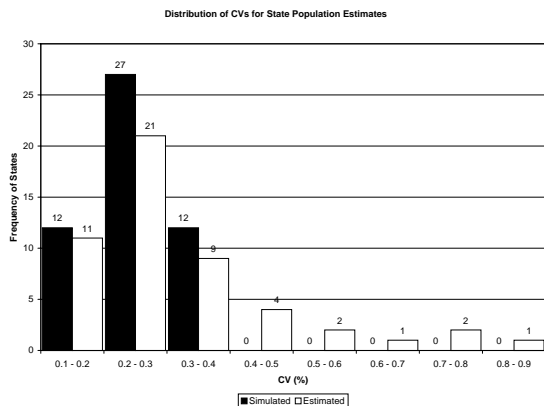

Figure 1: Distribtion of CVs for State Population Estimates

Figure 2: Plot of Simulated vs. Estimated CVs for State Population Estimates

3.3 Congressional District Comparison

For congressional districts, we see in Figure 3 that the distribution of CVs fits rather well with a slight tendency towards underestimating the variance. This histogram, however, hides the variability of simulating the variance for a particular district as seen in Figure 4. In the scatter plot, we see the slight tendency to underestimate the variance along with the fact that we may greatly underestimate or overestimate the variance for a particular district.

3.4 County Comparison

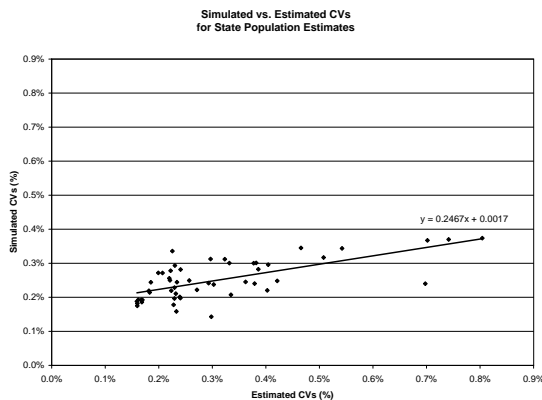To compare counties, we screened the counties to find the set of counties where both the demographic estimates and the Census 2000 counts have a population of greater than 100,000. The distribution of CVs for these comparable counties show in Figure 5 shows a similar trend as we saw with the congressional districts. The "OOS" category contains those counties which were greater than 100,000 for either the demographic estimates or the Census 2000 counts, but not both. The largest relative discrepancy in the distribution appears in the "0.6–0.8" percent category.

Figure 6 shows the detailed correlation between the simulated versus estimated CVs. We see that a subset of the counties draw the line of best fit down.

3.5 Place Comparison

The distribution of CVs for places are very similar for the simulated and estimated CVs as seen in Figure 7. The only category which contains a sizable difference is the category for "0.2–0.4". The plot in Figure 8 shows the most optimal fit of simulated versus estimated CVs of any geography.

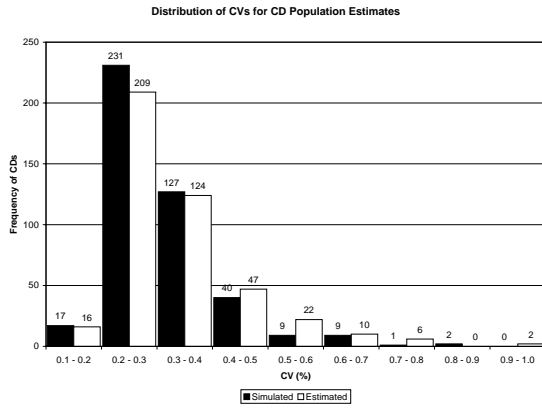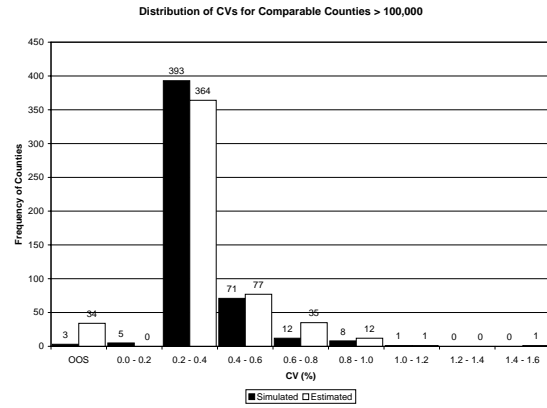Figure 3: Distribtion of CVs for CD Estimates

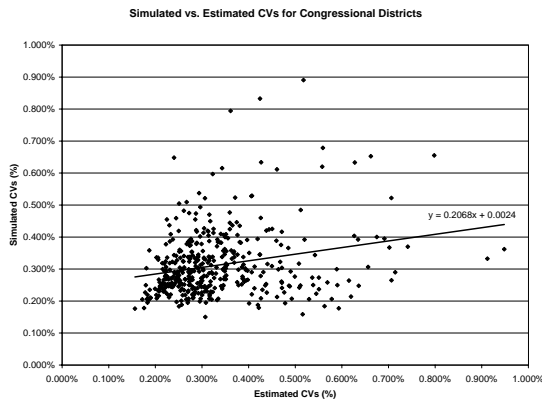Figure 5: Distribtion of CVs for Comparable County Estimates > 100,000
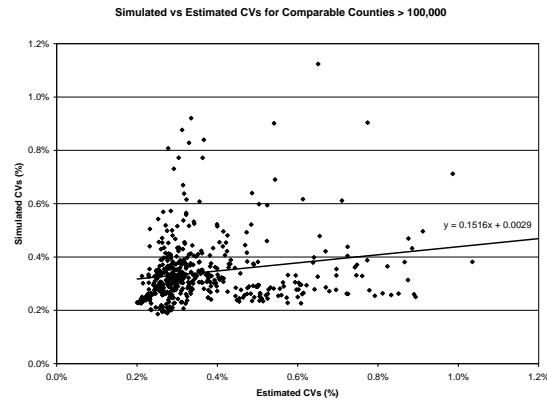

Figure 4: Plot of Simulated vs. Estimated CVs for CDs

Figure 6: Plot of Simulated vs. Estimated CVs for Counties

4 Discussion and Conclusion

Overall, the simulation did an adequate job of producing representative distributions of CVs for all geographic areas except states. While other geographical areas showed only a slight overall underestimation of the variance, there were some particular issues with states. One possible explanation for this is that in 1990, many of these high variance states had large portions of List/Enumerate Type of Enumeration Areas (TEAs). These areas were aggregrated with all other TEAs in the high return rate category since they had no mail return rate and were regarded as having a high capture probability by its nature. Many of these same areas in 2000 were Update/Leave TEA where the forms are left at the residence and are then mailed back. The Update/Leave produced more Census persons in the low return

rate category. These post-strata had higher than average CVs, around two percent for the Northeast and West Regions, whereas all the 1990 data fell into post-strata with CVs around 0.5–0.8 percent. The high return rate for these areas for Non-Hispanic Whites and Some Other race had a 2000 CV of around 1.00–1.68 percent versus our simulated 0.5–0.8 percent. These differences lead to large underestimation of the variance for these states in the Northeast and West regions with Update/Leave TEAs, which largely consist of Non-Hispanic White or Some Other Race persons.

Individual, simulated CVs were not a reliable indicator of the actual CVs in the 2000 A.C.E. Comparing the CVs at the post-stratum level showed great variability between the simulated versus the actual CVs. The exception to this trend was for places with
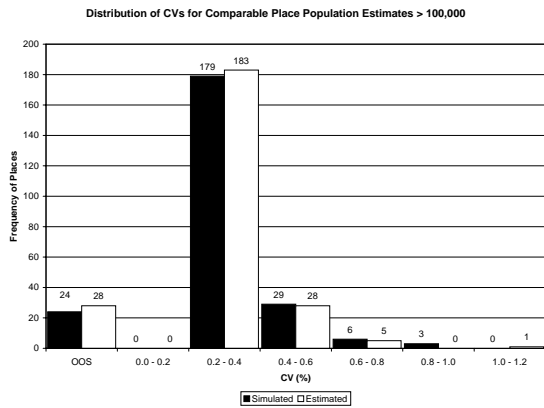
Figure 7: Distribtion of CVs for Comparable Place Population Estimates > 100,000
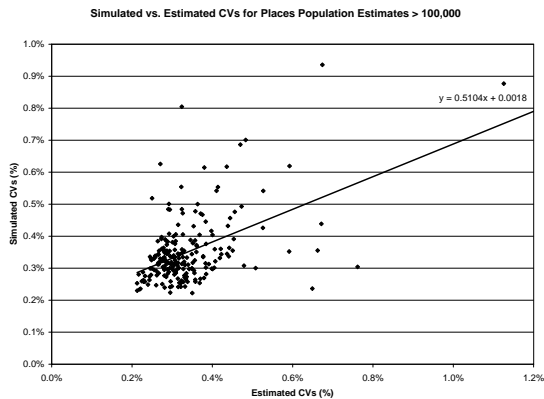

Figure 8: Plot of Simulated vs. Estimated CVs for Comparable Place Population Estimates > 100,000

population greater than 100,000. This may be due to the fact that the post-strata for Medium and Large Metropolitan Statistical Areas (MSAs) in MO/MB TEAs generally showed better agreement between the simulated and the estimated values. Much of the main trend, however, can be interpreted through the many limitations of our study. One limitation was that in 1990, Asians and Native Hawaiians had one race code and we had no way to separate them. Thus our simulation did not reflect the differing capture probability of the two races. In 2000, it was found that the CVs for these two groups were very different with Non-Hispanic Asians having CVs in the 0.87–1.00 percent range compared with Native Hawaiian or Other Pacific Islander with CVs in the 3.94–4.36 percent range. Our simulation, which had a combined CV of 0.65–1.24 percent, could not reflect

that difference. Post-strata containing Non-Hispanic Asians were reasonably simulated, while post-strata containing Native Hawaiians were poorly simulated.

A second major limitaion of our study was the inability to accurately predict the affect that the 2000 sample design would have on lowering the 1990 CVs for American Indians living on reservation. Consequently, we overestimated the CV greatly with our 3.81 percent versus the 2000 1.53 percent for owners and 1.48 percent for non-owners. The impact is greatest for areas with large concentrations of American Indians and Alaskan Natives (e.g., some southwestern states, sub-state areas which contain reservations).

In conclusion, the study did what it was intended to do. We produced general distributions of CVs which were weak where expected. This study does raise some interesting lines of further research.

1. Was the overall effect of TES adequately accounted for?

2. Would further collapsing of post-strata produce better simulated CVs?

References

[1] Kim, J., Navarro A., and Fuller W. (2000). "Variance Estimation for 2000 Census Coverage Estimates", *2000 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 515–520.

[2] Mule, V. (1999). "Accounting for Changes from the 1990 Post Enumeration Survey Methodology in the 2000 Accuracy and Coverage Evaluation Sample Design". *1999 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 507–511.

[3] Mulry, M, and Spencer, B. (2001). "Overview of Total Error Modeling and Loss Function Analysis". DSSD Census 2000 Procedures and Operations Memorandum Series B-19*.

[4] Sands, R. (1999). "Accuracy and Coverage Evaluation Survey: Variance Estimation for Post-Stratification Research". DSSD Census 2000 Procedures and Operations Memorandum Series Q-4.

[5] Starsinic, M, Sissel, D., and Asiala M. (2001). "Accuracy and Coverage Evaluation Survey: Variance Estimates by Size of Geographic Area", DSSD Census 2000 Procedures and Operations Memorandum Series B-11*.