Accounting for Changes from the 1990 Post Enumeration Survey Methodology in the 2000 Accuracy and Coverage Evaluation Sample Design

Vincent Thomas Mule, Jr., U.S. Bureau of the Census, Washington, D.C. 20233

Abstract: The Accuracy and Coverage Evaluation (A.C.E.) survey will have a different methodology than the 1990 Post-Enumeration Survey (PES). This research was done prior to the Supreme Court ruling when the Integrated Coverage Measurement (ICM) survey was being designed. Since the A.C.E. sample will be a subsample of the ICM design, studying differences between the ICM and PES will address differences between the A.C.E. and the PES and provide information for the A.C.E. survey design. Previous ICM sample design research used data from the PES while not considering these differences. This research focused on accounting for the changes in methodology when simulating coefficients of variation. The sample design and operational differences between the ICM and the PES were the primary changes investigated. While some differences could be accounted, other 1990 conditions are identified that could not. While this design will not be used in 2000, this research investigated how different variance estimations might have affected the simulated reliability. The effect of this design on minority and non-minority estimates is also discussed.

I. Introduction

This paper presents methods to calculate variance estimates and simulate coefficients of variation (CV) for the Integrated Coverage Measurement (ICM) design that are based on 1990 PES data. Previous sample design research assumed 1990 methods instead of reflecting the 2000 ICM methodology. These methods account for differences between the 1990 PES and the ICM. The first method was the differential weighting of the 1990 design. It did not account for the 2000 ICM design where proportional allocation should lead to more equal weighting. Other changed methods are 1) surrounding block search will not be performed for all blocks and 2) the effects of small block cluster[1] weighting. This analysis attempted to account, to some extent, for the methodology changes.

Results are examined for the ICM sample size of 750,000 housing units. This sample had been allocated to the 50 states and the District of Columbia. Within each state, a proportional allocation was planned. This design would produce efficient direct state total population estimates. However, we wanted to examine the reliability for state subgroup population estimates. This research examined four demographic group estimates in each state. While these were not going to be the estimates produced in 2000, it allowed us to see how this ICM sample design might have affected minority and non-minority estimates.

This analysis presents a variance estimate methodology for accounting for these changes in this 2000 ICM design. Three types of variance estimate methods are examined: direct calculation, synthetic groupings and a mixture of these two methods.

Section II describes the differences in methodology between the PES and ICM. Section III discusses the research methodology used in this analysis. Section IV provides a brief summary.

II. Difference in Methodology between PES and ICM

Changes between the 1990 and 2000 methodologies that we attempted to reflect in the reliability estimates are:

- Sample Design: The ICM plan was to conduct a state-based self-weighting design. This will produce more efficient estimates for state total population estimates. This methodology involves removing the effect of differential weighting of the 1990 PES design and replacing it with the self-weighting of the ICM plan.

[1] Small block clusters have between 0 and 2 housing units.

The author is a mathematical statistician in the Decennial Statistical Studies Division of the US Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. Research results and conclusions expressed are those of the author and do not necessarily indicate concurrence by the Census Bureau. It is released to encourage discussion.

- Surrounding block search: In 1990, a surrounding block search of 1 to 2 rings was performed for all sampled blocks. In 2000 ICM, the plan was to only perform a 1 ring surrounding block search for 20% of the sampled blocks. An adjustment has been made to compensate for the decrease in surrounding block search.

Since this research uses 1990 Census and PES data, the methodology assumes that certain results from 1990 occur again in 2000. These results are:

- The Master Address File is 99% complete.
- There was a 98% ICM Response Rate. (This was the response rate for the 1990 PES).
- The total Census 2000 estimated undercount is 1.8%.
- 100% accurate data capture in the Census.

If any of the above conditions are not met then the reliability estimates will increase.

One difference between the 1990 and 2000 ICM methodologies that we have not reflected in these estimates is:

- Handling movers: There is a period of time between Census day and when the ICM interview would have been conducted. During that period, people can and do move. In 1990, a procedure known as PES-B accounted for the movers in the estimation. In the current 2000 plan, a procedure known as PES-C will be used[2]. This methodology does not reflect how the difference in handling movers procedures can affect the reliability estimates.

III. Research Methodology

Step 1: Obtain Direct Variance Estimates from 1990 PES

The direct variance of the Dual System Estimate (DSE) for a Census division/poststrata were calculated. For purposes of this work, each division had $i = 28$ possible poststrata. The 28 poststrata were formed by the cross-classification of 7 age/sex, 2 race/ethnicity (minority, non-minority) and 2 tenure (owner, renter) categories. The variance was calculated by:

$$\text{Var}\,(\text{DSE}_{\text{i, E-Sample Estimate}}) = E_i^2\ \text{Var}(\text{CF}_i)$$

where E_i is the ratio adjusted weighted E-sample size estimate and Var(CF_i) is the coverage factor variance for the *i*th division/poststratum using a jackknife methodology on 1990 PES data. The ratio adjusted weighted E-sample size was used as an estimate of the unadjusted Census count in the above calculation.

The actual Census count should have been used instead. Because of this, Census counts for each division/poststratum state were obtained. Since the DSE variance is a function of the Census count squared, we recalculated the DSE variance with the actual Census count.

The variance of the Dual System Estimate was set equal to DSE variance using the E-Sample estimate times the Census count squared divided by the ratio adjusted weighted E-sample estimate squared.

$$\text{Var}\,(\text{DSE}_i) = \frac{C_i^2\ \text{Var}\,(\text{DSE}_{\text{i, E-Sample Estimate}})}{E_i^2}$$

where Var (DSE_i) is the variance of the Dual System Estimate, C_i is the Census count, E_i is the ratio adjusted weighted E-sample estimate and Var($\text{DSE}_{\text{i,E-Sample Estimate}}$) is the DSE variance using the E-sample estimate in the *i*th poststratum in a division.

Step 2: Obtain Variances for the Four Collapsed Poststrata in a Division

In addition to looking at the reliability of the total state estimate, we wanted to investigate the reliability of

[2] PES-B matches the census enumerations at the inmover's census day address to estimate movers. PES-C estimates match rate by matching outmovers but estimates number and characteristics from inmovers.

certain groups within a state. The four groups examined were Majority Owners, Majority Renters, Minority Owners and Minority Renters. The Minority groups consisted of Black, Non-Black Hispanic, Asian and Pacific Islander, and American Indians on Reservations. These four groups were formed by collapsing the 28 poststrata from the previous step across the seven Age-Sex levels. There was a significant amount of covariance among the Age-Sex levels that was reflected in the DSE variance of the four collapsed groups. The national correlation between the Age-Sex levels for each of the four collapsed groups was estimated using 1990 PES data.

The four collapsed poststrata DSE Variances were calculated as follows:

$$\mathrm{Var}(\mathrm{DSE}_{i*}) = \sum_{a=1}^{7} \mathrm{Var}(\mathrm{DSE}_{i*,a}) + 2\sum_{a<b}^{7}\sum \mathrm{Cov}(\mathrm{DSE}_{i*,a}, \mathrm{DSE}_{i*,b})$$

where $\mathrm{Var}(\mathrm{DSE}_{i*})$ = DSE Variance for Division/Collapsed Poststratum (4 Levels), $\mathrm{Var}(\mathrm{DSE}_{i*,a})$ = DSE Variance for Division/Collapsed Poststratum/Age-Sex Levels (7), $\mathrm{Cov}(\mathrm{DSE}_{i*,a}, \mathrm{DSE}_{i*,b})$ = DSE Covariance between Age-Sex Levels within a collapsed poststratum for a Division.

The national covariance structure across the age-sex levels was assumed appropriate for a collapsed poststratum. The DSE covariance between age-sex levels was calculated as follows:

$$\mathrm{Cov}(\mathrm{DSE}_{i*,a}, \mathrm{DSE}_{i*,b}) = \hat{\rho}\sqrt{\mathrm{Var}(\mathrm{DSE}_{i*,a})\,\mathrm{Var}(\mathrm{DSE}_{i*,b})}$$

where $\hat{\rho}$ = Correlation between national age-sex levels for collapsed poststratum based on the 1990 PES.

Step 3: Obtain Variance Component Independent of Sample Size and Weights

The 1990 PES design had differential weights based on the sampling strata. This step factored out of the DSE variance in a Division/collapsed poststratum, the effect of sample size and the differential weighting of the 1990 PES. This variance component of the DSE in a Division/collapsed poststratum, σ_{i*}^2, was calculated from the following formula.

$$\sigma_{i*}^2 = \frac{\mathrm{Var}(\mathrm{DSE}_{i*})}{\sum_{i=1}^{n_{i*}} w_{i*,j}^2}$$

where $w_{i*,j}$ = the inverse probability of selection in the 1990 PES, n_{i*} = the number of E-sample people in the *i**th Division/poststratum in the 1990 PES.

Step 4: Obtain the Person Sample Sizes by State

For each state, the block cluster sample sizes needed to be converted into people sample sizes. The block cluster is the basic unit of sampling for the A.C.E. The person sample size for each state was estimated by the number of block clusters allocated times the average number of E-sample cases per block cluster in the division in the 1990 PES. This accounted for different densities of people per block cluster across the United States. The state person sample size was then proportionately allocated to the 4 collapsed poststrata based on the 1990 Census population. Let $n_{i*,s}^*$ be the resulting sample for collapsed poststratum $i*$ within state s.

Step 5: Estimate Variance for the Allocation by Division/Poststrata

The variance of the DSE for a Division/collapsed poststrata was shown earlier to be a function of the variance component, σ_{i*}^2, sample size, n_{i*}, and the weights, $w_{i*,j}$. The amount of sample in each Division/collapsed poststratum will change from the 1990 PES to this design. A new estimate of the variance of the DSE for a Division/collapsed poststrata was calculated based on the new sample size and weights.

$$\text{Var}^*(\text{DSE}_{i*,\text{Division}}) = \sigma^2_{i*} \sum_{s=1}^{k} \sum_{j=1}^{n^*_{i*,s}} w^{*2}_{i*,j,s}$$

where $w_{i*,j,s}{}^*$ = the inverse probability of selection in the ICM self-weighting design and k = the number of states in a division.

The variance of the Coverage Factor for a Division/collapsed poststrata is equal to the variance of the DSE divided by the square of the unadjusted Census Estimate.

$$\text{Var}^*(\text{CF}_{i*,\text{Division}}) = \frac{\text{Var}^*(\text{DSE}_{i*,\text{Division}})}{C^2_{i*,\text{Division}}}$$

Step 6: Estimate Variances and Simulate Coefficients of Variation for States

Three types of variance calculations estimated three types of methods, direct, synthetic and mixed. The variance estimate of the DSE for direct state estimates and synthetic state estimates was calculated based on the proportional allocation of the medium and large block cluster sample. Direct estimates were calculated by only using the sample allocated to the state.

Synthetic estimates were calculated by forming groupings by Census division. A state demographic group coverage factor variance estimate "borrowed strength" by using the division group coverage factor variance estimate. One limitation of this analysis is that while synthetic variances tend to be lower than direct, the bias introduced is unknown.

A mixed estimate was calculated using a direct estimate for some of the collapsed poststrata and a synthetic estimate for the remaining. The variance estimates were adjusted to account for 1) surrounding block search not being performed for all blocks and
2) the effect of small block weighting.

The variance estimates are calculated by summing over the 4 collapsed poststrata (*i**) in each state. Since there was small correlation among the four collapsed poststrata at a national level in 1990, the covariance among the four collapsed poststrata was ignored.

For Direct State Estimates:

$$\text{Var}\ (\hat{X}_{s,D}) = \sum_{i*=1}^{4} n^*_{i*,s} w^{*2}_{i*,s} \sigma^2_{i*}\ \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Block}}$$

where $n_{i*,s}{}^*$ = the sample size (in persons) allocated to state/poststratum for the design,
$w_{i*,s}{}^*$ = inverse of the probability of selection for the design,
$\text{ADJ}_{\text{Surr. Block}}$ = adjustment for doing surrounding block search in only 20% of the blocks,
$\text{ADJ}_{\text{Small Block}}$ = adjustment for small block weighting effect.

For Synthetic State Estimates:

$$\text{Var}\ (\hat{X}_{s,SYN}) = \sum_{i*=1}^{4} C^2_{i*,s}\ \text{Var}^*(\text{CF}_{i*,\text{Division}})\ \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Block}}$$

For Mixed State Estimates:

$$\text{ar}(\hat{X}_{s,MIX}) = \sum_{i*\in \text{Maj.}} n^*_{i*,s} w^{*2}_{i*,s} \sigma^2_{i*}\ \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Block}}$$
$$+ \sum_{i*\in \text{Min.}} C^2_{i*,s}\ \text{Var}^*(\text{CF}_{i*,\text{Grouping}})\ \text{ADJ}_{\text{Surr. Block}} \text{ADJ}_{\text{Small Blk}}$$

where direct estimates are used for Non-Minority Owners and Non-Minority Renters and synthetic estimates are used for Minority Owners and Minority Renters.

Coefficients of variation can be calculated using the variance methodologies described above and the 1990 DSE estimates.

IV. Summary

The graphs on the following page show simulated coefficients of variation (CV) estimates based on direct, synthetic and mixed variance methods for the 50 states and the District of Columbia. These five graphs compare the simulated CVs for the Total Population, Non-Minority Owner, Non-Minority Renter, Minority Owner and Minority Renter.

- The direct simulated CVs were all below 0.5%. This reliability would have been needed in order to use these estimates for reapportionment.

- Non-minority direct simulated CVs were all below 0.9% for the direct estimation. Using synthetic estimation lowered the reliability to between 0.1% and 0.4%.

- For Minority estimates, the direct owner simulated CVs were higher than 3.0% for 10 states. For renters, the direct simulated CVs were higher than 3.0% in 21 states. However, the synthetic CVs were only higher than 3.0% for only three states for both owners and renters.

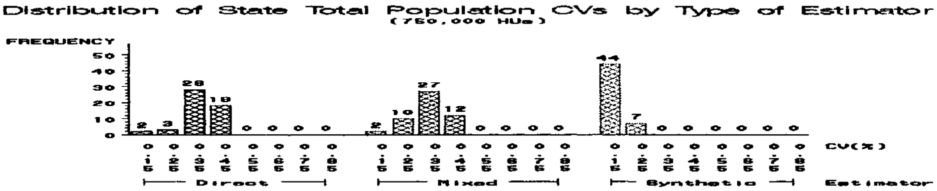
V. Future Research

Future research will involve using this methodology to simulate CVs for possible A.C.E. sample designs. Possible areas for investigation are:

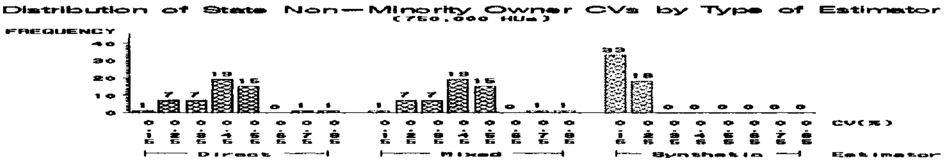
- Simulating 1990 Poststrata, synthetic state and synthetic congressional district CVs. For various A.C.E. sample designs, these simulated CVs can be estimated. This will allow comparisons of the sample design effect on demographic/tenure, state total population and substate area reliabilities.

- Using different synthetic groupings instead of Census divisions. Alternative groupings may have more similar coverage properties while less bias is introduced.

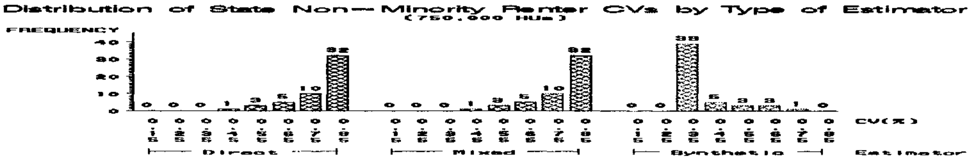
Graph 1



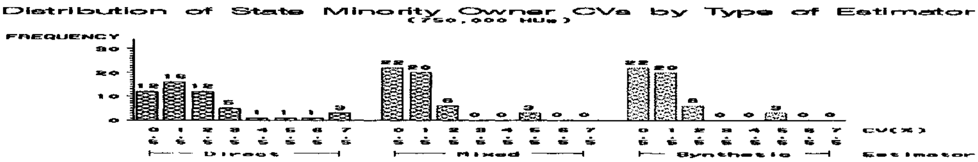
Graph 2



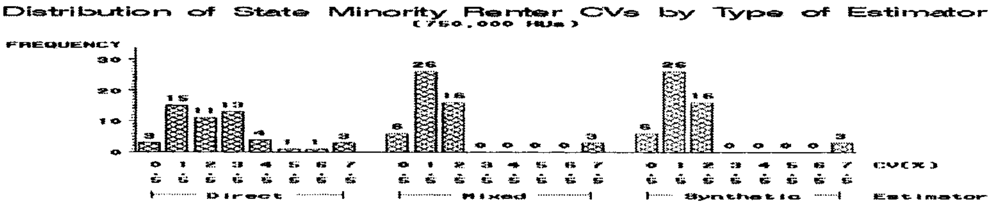
Graph 3



Graph 4



Graph 5


Distribution of State Minority Renter CVs by Type of Estimator
(750,000 HUs)
FREQUENCY
CV(%)
Direct
Mixed
Synthetic
Estimator