Paper SDA-06

Potential Change in Reliability Measures Based on Decreased Sample Size for the Census Coverage Measurement Survey

Vincent Thomas Mule, Jr., U.S. Census Bureau, Washington, DC

ABSTRACT

As part of the evaluation of the 2010 Census, the U.S. Census Bureau conducts the Census Coverage Measurement (CCM) Survey. This survey produces the net coverage results of undercounts or overcounts of the Census. In addition to net coverage, our program has been asked to estimate the components of census coverage that include erroneous enumerations and omissions. A proposal was made to reduce our sample from the planned 300,000 housing units to help offset the costs of the operational enhancements to try to help reduce the nonsampling error in our estimates. This paper shows the results of a simulation study to assess the potential change in reliability measures for our proposed estimates. This paper will show results of the study and how SAS® Procedures like Proc SURVEYSELECT and Proc SGPANEL were used.

INTRODUCTION

In the early fall of 2009, the Census Bureau proposed adding operational enhancements to the Census Coverage Measurement (CCM) program in an attempt to reduce nonsampling error. In order to implement and pay for these enhancements, the CCM sample size of 300,000 housing units in the Population (P) sample would be reduced. This document shows the results of a simulation study to estimate the potential impact of a sample reduction on the sampling error for the dual system estimates and the component of census coverage estimates that the 2010 CCM program will produce. In this study, we examine the ramifications of reducing the CCM sample from 300,000 housing units to 150,000 or 160,000 housing units.

This study focuses on the potential impact of a sample reduction for the estimates for people in housing units in the United States. This analysis does not try to quantify any potential reduction in nonsampling error based on the reduced sample size. As the budget work progressed, it became apparent that the budget could support a national sample size of about 170,000 housing units after accounting for these additional enhancements. While we did not re-run our analysis for a national sample of this size, the analysis laid out in this document guided our sample size and allocation recommendation.

Section I shows the sample design reduction options examined in this study. Section II shows the key survey estimates for which we look at the impact of the reduction in sample size. Section III provides some background on the 2010 CCM Block Cluster Listing sample. Section IV documents the simulation methodology. Section V shows the statistics used to quantify the potential impact of reducing the sample. Section VI lays out the limitations of these results. Section VII presents the final results of the projected potential change in reliability measures and a recommendation.

SAMPLE DESIGN REDUCTION OPTIONS

This analysis examined three possible reduction sample design options. These options utilize the random groups that were assigned during the Listing Block Cluster selection. During this operation, each block cluster was assigned to one of twenty random groups.

REDUCTION OPTION 1: ALL REDUCE HALF

This reduction design drops half the clusters in all listing sampling strata. This would be implemented by dropping the first 10 of the 20 random groups from the CCM sample. In this study, we simulate this reduction by subsampling the block clusters in every listing stratum at a rate of 1-in-2. This reduction option results in approximately 150,000 housing units being in sample...

REDUCTION OPTION 2: KEEP AIR&HI, REDUCE REST HALF

This reduction design retains all of the listing sample block clusters in the American Indian Reservation (AIR) stratum and in the state of Hawaii (HI). This is done to increase the sample size for the American Indians living on Reservations and the Native Hawaiian and Pacific Islander populations. For the remaining clusters in the medium and large sampling strata, 10 of the 20 random groups would be dropped. In this study, this subsampling is simulated by subsampling the block clusters in each stratum at a rate of 1-in-2. This reduction option results in approximately 160,000 housing units being in sample.

REDUCTION OPTION 3: KEEP AIR&HI, REDUCE REST 45%

This reduction design is similar to Option 2 by retaining all of the clusters in the American Indian Reservation stratum and the state of Hawaii. The difference is that we would only drop nine of the twenty random groups. This yields approximately 450 more medium and large block clusters to remain in sample. This option results in approximately the same 160,000 housing unit size as Option 2. Since we would have more block clusters in sample, we would select fewer housing units from block clusters with 80 or more housing units. Since we have a clustered sample, this option has a potential benefit of reducing the effect of intraclass correlation on our sampling variance. In this study, the block clusters are subsampled at a rate of 1-in-1.818 to simulate the results of dropping 45 percent of the random groups. Section IV provides details on how the simulation study accounted for having fewer housing units per cluster in the large clusters.

COVERAGE ESTIMATES

The 2010 CCM program has expanded over the past coverage measurement programs to provide more types of estimates that just the net coverage error estimates. The net error is the difference of the dual system estimate and census count. Since net error was the sole estimation objective for past programs, the key statistic for which reliability measures were monitored was the dual system estimate. The reliability measure examined was the coefficient of variation (CV) for the dual system estimate.

There are two major changes for the 2010 CCM program that need to be accounted for when assessing the impact of a sample reduction on the reliability of the survey.

1. The CCM is using logistic regression modeling instead of post-stratification to generate the dual system estimates. This difference means that our 2010 estimates will be model-based as compared to design-based as in the past. One advantage of the logistic regression approach is that higher order interactions do not necessarily have to be included in the model. The CCM staff is researching different models but a final model has not been chosen. For this simulation, one of the candidate models is used to assess the implications of each reduction option. The potential change in reliability measures is dependent on both the model and sample size.

2. The CCM is providing estimates of the components of census coverage. This is a new objective, and includes erroneous enumerations and omissions. There may be different reliability implications for the components of census coverage estimates as compared to those seen for the dual system estimates. The estimate of erroneous enumerations is a design-based estimate. The estimation plan uses two-stages of ratio adjustment to help control variances. For this study, we estimate omissions by adding the net error and erroneous enumeration estimates. If a different estimator were used, say for estimating omissions, the impact on the potential change in reliability measures may differ from what we see in this study. Though, if the only change is in how census imputations are treated when estimating omissions, there should be no difference from the results we see here.

In this study, we examined the potential increase by examining three estimates:

1. Dual System Estimates
2. Erroneous Enumerations
3. Omissions

2010 CCM LISTING SAMPLE

Mule et al. (2009) describes the 2010 CCM sampling methodology. The first phase of sampling is the selection of the block cluster sample for which independent listing is done. The CCM Listing Sample for the 50 states and the District of Columbia was selected in early 2009. Table 1 shows the block cluster distribution across the six sampling strata.

Table 1: 2010 CCM Block Cluster Listing Sample

Stratum	Block Clusters
Small	2,500
Medium/Owner	3,689
Medium/Non-Owner	1,523
Large/Owner	1,900
Large/Non-Owner	1,867
AIR	356
Total	11,835

Source: Davis (2009)

The small block clusters are not part of the proposed reduction plans and are left aside for this simulation research. The small block clusters will be subsampled in the previously planned Small Block Cluster Subsampling operation. We also set them aside because their contribution to the variance estimates is similar for each of the three proposed reduction plans.

Leaving aside the 2500 small block clusters means the sampling frame for this reduction consists of 9,335 medium and large block clusters. Two of the three sample reduction plans under consideration in section I retain all of the clusters in two areas: a) American Indian Reservations and b) Hawaii. There are 356 block clusters in the American Indian Reservation sample and 132 medium and large block clusters in Hawaii.

SIMULATION METHODOLOGY

This simulation study assesses the potential change in reliability based on the three sample reduction designs presented earlier. This section documents how the simulation study is carried out. For each of the three sample reduction designs, we simulate the resulting point estimates, standard errors and coefficients of variation based on 100 simulation samples.

Since we are interested in determining the potential impact on the reliability measures, we do a fourth simulation to serve as a baseline for comparison. This fourth set is based on simulating the results as if the 2010 Listing Sample was not reduced.

The following is a general synopsis of the methodology:

- Draw 100 block cluster samples from the 2010 Listing sample
- For each of the 100 simulation block cluster samples, draw a sample of persons from the 2000 Accuracy and Coverage Evaluation (A.C.E.). For each 2010 block cluster in a particular sample, an A.C.E. block cluster with similar characteristics is selected to provide the E- and P-sample person data.
- For each simulated sample, generate the point estimates, standard errors and coefficients of variation.

BLOCK CLUSTER REDUCTION SAMPLE SELECTION

In this study, we select 100 separate block cluster subsamples from the 2010 listing sample for each of the three reduction sample designs. Option 1, reducing all block clusters by half, results in an expected 4,668 block clusters in each simulation. Option 2, keeping all AIR and Hawaii block clusters, results in 244 more expected block clusters in sample as compared to Option 1. Option 3, reducing the remaining strata by 45 percent as compared to half, results in 442 more expected block clusters in sample as compared to Option 2.

Each simulation sample is selected by taking a systematic sample within each of the Listing sampling strata in each state. The plans for the 2010 reduction will use the random groups assigned during the listing sample. This study started when selecting a systematic sample within the listing strata was still a possible means for implementing the reduction. Despite the decision to use random groups, a systematic sample is used in this research because it approximates the results of the random groups and allows different 2010 clusters to be in the resulting simulation samples. The subsampling rates are one of the following: a) 1-in-2 when half the random groups are dropped, b) 1-in-1.818 when 45 percent of the random groups are dropped and c) 1-in-1 when all clusters are retained. While I used my own SAS code that I had written before Proc SURVEYSELECT was introduced to do this sampling, Proc SURVEYSELECT could have been used to select this systematic sample.

PERSON DATA SELECTION FOR THE REDUCTION BLOCK CLUSTERS

After a simulation block cluster sample is selected, the person data for each block cluster is drawn. For both the 2010 simulation block cluster sample and the 2000 A.C.E. block clusters, we had the following information about the block cluster:

a) If the cluster has over 40 percent renters based on the previous census data,
b) If the cluster size classification is medium or large, and
c) If the cluster is located on an American Indian Reservation?

Using these three characteristics, we assign the 2000 A.C.E. block clusters to one of these five 2010 sampling strata because the A.C.E. sample design had a different stratification and differential sampling plan. Using Probability Proportional to Size (PPS) sampling with replacement, we select a sample of A.C.E. block clusters from each of the 2010 sampling strata. The SAS SURVEYSELECT procedure is used to implement this sampling. The measure of size is the 2000 A.C.E. block cluster weight. This takes into account that the A.C.E. block clusters had a differential sample design.

For Reduction Options 1 and 2, all of the person records in the selected 2000 sample block clusters are used in the simulation. For Reduction Option 3 that reduces 45 percent of the block clusters as compared to 50 percent, we want to reflect that the number of housing units that will be in sample in large block clusters will be smaller. To do this, we subsample the housing units using SAS SURVEYSELECT for both the E and P samples. Our approximate calculations determined that we would expect to select about 35 housing units if we only reduced 45 percent of the block clusters. This compares to an expected sample size of 42 housing units with the 50 percent block cluster reduction for the other options. To simulate this smaller expected number of housing units selected per block cluster, the housing units in the large block cluster were subsampled using a simple random sample with a probability of selection of 0.83 which is the rounded outcome of 35/42. The result is that each 2010 block cluster in a particular simulation sample is associated with a 2000 A.C.E. block cluster. The 2000 A.C.E. person data provides the E-sample and P-sample data to be the basis of our dual system and component error simulation estimates

For the simulated E-sample and P-sample data for dual system estimation and the simulated E-sample data for component estimation, the final step is to adjust the sampling weights for each person record.

$$SIMWGT = (2000 A.C.E.\,Weight) \times \frac{CWEIGHT_{2010}}{CWEIGHT_{2000}} \times \frac{1}{HUSRATE} \qquad (1)$$

Where SIMWGT is the simulated person weight,
2000 A.C.E. Weight is the final sampling weight (including TES),
$CWEIGHT_{2010}$ is the simulated reduction block cluster sampling weight,
$CWEIGHT_{2000}$ is the A.C.E. block cluster sampling weight and
HUSRATE is the housing unit sampling rate.

For variance estimation purposes, each 2010 block cluster is systematically assigned to one of twenty random groups. This allows the standard error estimates to be generated using a delete-a-group jackknife procedure.

DUAL SYSTEM ESTIMATION

For the dual system estimation, we generate estimates by our proposed 2010 estimation approach. Mule (2008) documents generating dual system estimates using logistic regression models. This research focuses on dual system estimates prior to the adjustment for correlation bias. Mule (2008) documents the planned correlation bias adjustment[1].

Our dual system estimation formula prior to correlation bias adjustment is:

$$DSE = \sum_{i \in Census} p_{dd,i} \frac{p_{ce,i}}{p_{m,i}} \qquad (2)$$

Where $p_{dd,i}$ is the predicted data-defined rate
$p_{ce,i}$ is the predicted correct enumeration rate
$p_{m,i}$ is the predicted match rate.

---

[1] This simulation work was unable to incorporate this adjustment into the results at this time. Such an adjustment can be examined in future work, if warranted.

For each simulation, we use three logistic regression models. Table 2 summarizes the data, outcome variable and weights (if necessary) that are used in each regression.

Table 2: Logistic Regression Runs for Each Simulation

Run	1	2	3
Data	2000 Census Person Records	2010 Simulated E-sample Person Data	2010 Simulated P-sample Person Data
Outcome	Data-defined status	Correct enumeration	Match
Weights	Not needed	Simulated E-sample Weight	Simulated P-sample Weight

For all three models, a logistic regression model with 76 main effects and interaction terms is used. The main effects are Census Region, Race/Origin domains, tenure, sex and age. Age is represented by a 6 piece spline that allows a quadratic from 0 to 17, linear from 17 to 20, quadratic from 20 to 50 and then linear from 50 to a top code of 80.

The coefficients from the logistic regression models documented in Table 2 are used to generate the predicted data-defined rate, predicted correct enumeration rate and the predicted match rate shown in equation (2) for each census case. The standard error of the dual system estimates are estimated using a delete-a-group jackknife estimator.

The dual system estimates for the estimation domains by restricting the summation of census cases in equation (2) to only the census cases in the corresponding estimation domain. For example, only the census cases in the Non-Hispanic Asian domain are used to generate the dual system estimate for the Non-Hispanic Asian domain.

ESTIMATES OF ERRONEOUS ENUMERATIONS AND OMISSIONS

For people in housing units, the four components of census coverage are correct enumerations, erroneous enumerations, whole-person census imputations, and omissions. The reporting of whole-person census imputations will be a tally of census records so reducing the sample size has no impact on this result. The data-defined cases are classified into the binary results of whether they are correct or erroneous. We focus on the erroneous result. We also want to assess the impact of a sample reduction for the estimate of omissions as well.

Our estimation methodology for 2010 has two stages of ratio adjustments of the sampling weights. The first ratio adjustment will adjust the weighted E-sample estimates to the data-defined counts for specified cells. Research is ongoing to determine the cells for the first stage. The second stage of ratio adjustment will take the resulting weights from the first stage adjustment and adjust them to the estimation domains for which correct and erroneous enumerations are being estimated.

In this simulation research, we implement just the second stage of ratio adjustment. The following equation shows the estimator for the erroneous enumerations for the *j*th estimation domain.

$$EE_j = \left(\frac{DD_j}{\sum_{i \in j} SIMWGT_i} \right) \times \sum_{i \in j} \left(SIMWGT_i \times PR_{compee,i} \right) \quad (3)$$

Where DD_j is the data-defined count for estimation domain j,
$PR_{compee,i}$ is the probability of being an erroneous enumeration for component estimation.

Using the dual system estimate prior to correlation bias adjustment and the estimate of erroneous enumerations, we generate an estimate of omissions for estimation domain *j* shown in Table 1 by the following formula.

$$Omissions_j = DSE_j - Census_j + EE_j \quad (4)$$

Where DSE_j is the dual system estimate prior to correlation bias adjustment and $Census_j$ is the 2000 census count.

The standard error of the erroneous enumeration and omissions estimate are estimated using a delete-a-group jackknife estimator.

STATISTICS TO ASSESS POTENTIAL IMPACT

In assessing the impact of sample designs, we focused on three measures of reliability for each of the three reduction alternatives and the baseline comparison.

1. Coefficient of Variation of the Dual System Estimate, CV(DSE)
2. Coefficient of Variation of the Erroneous Enumeration Estimate, CV(EE)
3. Standard Error of the Omission Estimate, SE(Omission)

The coefficient of variation is a standard measure that is used to assess the reliability of a survey estimate. The coefficient of variation is the ratio of the standard error over the point estimate. This measure is used to assess the implications on the dual system estimates and the erroneous enumeration estimates. For omissions, the change in the standard error is used because the coefficient of variation is sensitive to how close the point estimate is to zero which can be the case for estimates of omissions. The sensitivity of the point estimate being close to zero is also why we examine the coefficient of variation of the dual system estimate and not the coefficient of variation of the net error or percent undercount estimates.

In order to assess the potential change in reliability for a reduction sampling option, we examine the ratio of the average estimate of the reliability measure of the 100 simulations for the reduction sampling option to the average estimate of the reliability measure estimate for the 100 baseline simulations. The following formula shows the ratio of the potential increase in the coefficient of variation of the DSE. The formulas for the other two measures have a similar form. The standard error of the percent increase was calculated using the delta method.

$$Increase_j = \frac{\sum_{1}^{100} CV\left(DSE_j\right)_{reduction\ option,i}}{\sum_{1}^{100} CV\left(DSE_j\right)_{baseline,i}} - 1 \qquad (5)$$

These potential increase estimates are generated for each of the three reduction options stated earlier. Figures are generated that show the average percent increase for each option and also include the 95 percent confidence interval based on the 100 simulations. Since the guidance is to not reduce the sample size below 150,000, a line has been added to the figures that shows a 41.4 percent increase. This is the approximate increase in standard errors that you would expect to see when cutting the sample in half.

Proc SGPANEL was used to show the figures in the attachment. This procedure was able to show the potential changes in reliability for the three sample reduction options and the three estimates (dual system estimates, erroneous enumerations and omissions) for various estimation domains. Two records were created for each simulated estimate so the procedure would generate a 95 percent confidence interval of the predicted change. This was done by using the RESPONSE, STAT and LIMITSTAT options when using the VLINE statement.

LIMITATIONS

This simulation study projects the potential percent change in the reliability measures. The results shown here are based on the following assumptions and limitations. Any violation or difference of these could lead to different results than those projected here.

- The 2010 CCM person results are like the 2000 A.C.E.
- This simulation used the 2000 A.C.E. person weights that had the targeted extended search weighting. The 2010 CCM will not have this feature so some of the differential weighting seen in this simulation will not be present in 2010.
- This simulation assumed that the small block clusters would impact the three reduction options and the baseline approach equally.
- This simulation uses a candidate logistic regression model in all results. Any differences of the final model to this candidate could result in different changes in reliability than those shown here.

CONCLUSIONS

Figure 1 shows the overall national-level projected percent change in reliability measures. Option 3, keep all of the AIR and Hawaii block clusters and reduce 45 percent of the remaining block clusters, shows average potential increases of 37 to 39 percent. The other two options show potential increases that average between 38 percent and 51 percent. The results for all options include the 41.4% reference line.

Figures 2a shows the results for the one of the seven race/origin domains. For the American Indians on Reservation domain, figure 2a shows that Options 2 and 3 are still projecting average increases of 25 percent and 18 percent, respectfully, for the dual system estimates. These two options keep the entire AIR sample so this result might seem surprising at first. The reason for the increase is that we now have a model-based estimate for this group as compared to a post-stratified estimate as in the past. Since we used the same model in all simulations, the American Indians on Reservation estimates are sensitive to the changes in the coefficients for the region, tenure, age and sex coefficients. The reduction in sample size for estimating these other coefficients is leading to some potential increase. The potential increase in the CV(DSE) for the American Indians on Reservation domain is not as large as those seen in the figures for the other domains. The potential impact for this domain could change if the 2010 model has major differences from the candidate model used in this research.

For the American Indians on Reservation domain, we see that keeping all of the AIR clusters in sample is showing no potential increase for the erroneous enumeration estimates. This is because the erroneous enumeration estimate is a design-based estimate.

Figures 3a shows the potential percent increases in reliability measures for the owner population. The figure show comparable results for all three reduction options. Figures 4a shows the potential increases in measures for the Northeast census regions. The three reduction options are showing similar results for the region. While Options 2 and 3 retain all of the clusters on AIR and Hawaii, we are seeing results where the confidence intervals for all three options are overlapping. Figures 5a shows the potential increase in reliability measures for the states in the Northeast census region. The confidence intervals of the average potential increase based for the three options all overlap.

Based on the results of this simulation study, we recommend Option 3 among the three options studied. This design keeps all of the AIR and Hawaii block clusters and cuts the sample in the remaining strata by 45 percent. The results show some benefits to the race/origin domain estimates. By only cutting 45 percent of the block clusters as compared to 50 percent, this approach should also be able to minimize the impact of the clustered observations on the variance estimates.

REFERENCES

Davis, Peter (2009), “2010 Census Coverage Measurement: 2010 Block Cluster Sampling Results” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-C-16,
October 28, 2010.

Mule, Thomas (2008), “2010 Census Coverage Measurement Estimation Methodology” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-18, October 30, 2008.

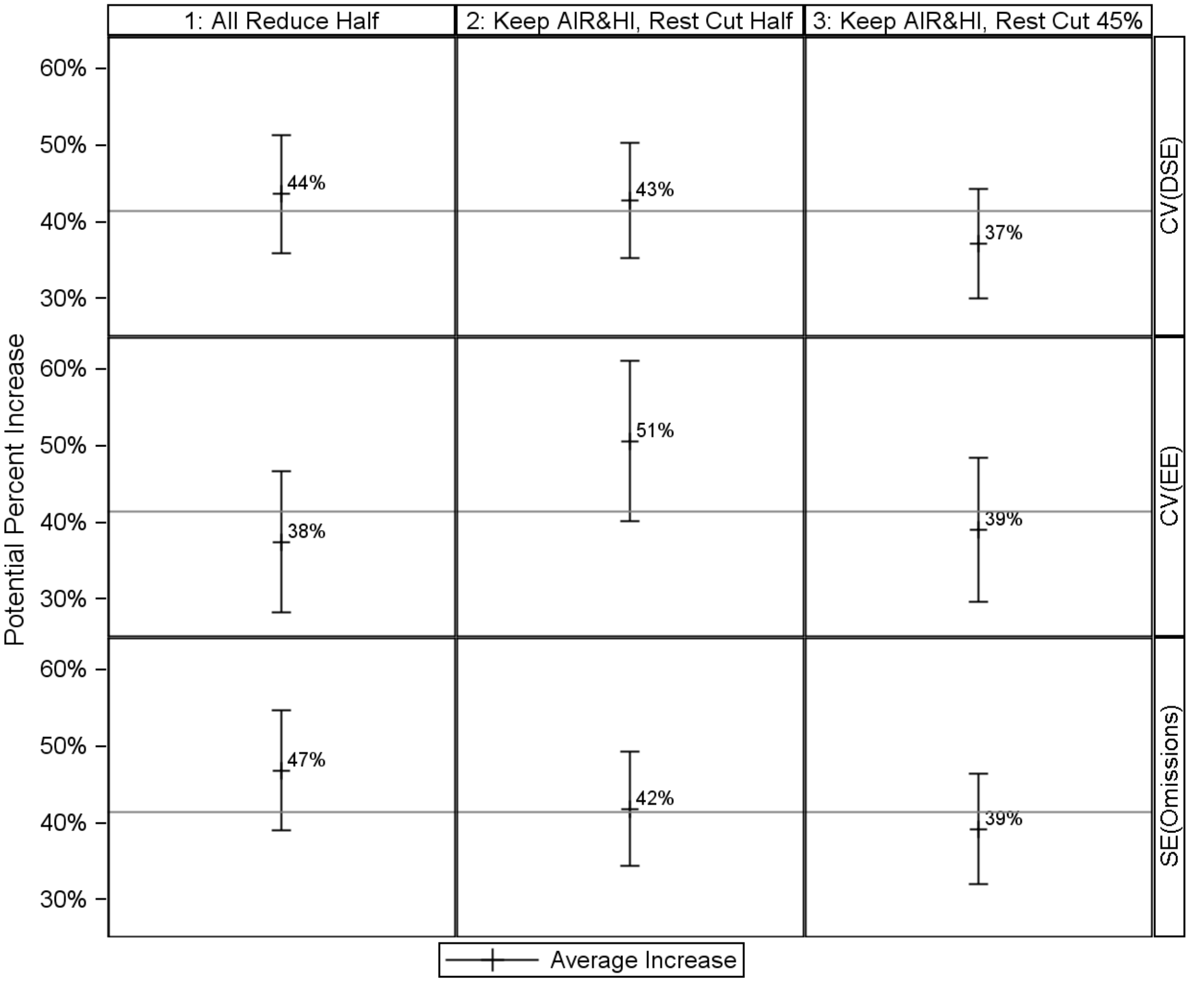
Mule, T., Davis, P. and Mulligan J. (2009), “2010 Census Coverage Measurement: Sample Design” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-C-13,
May 28, 2009.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

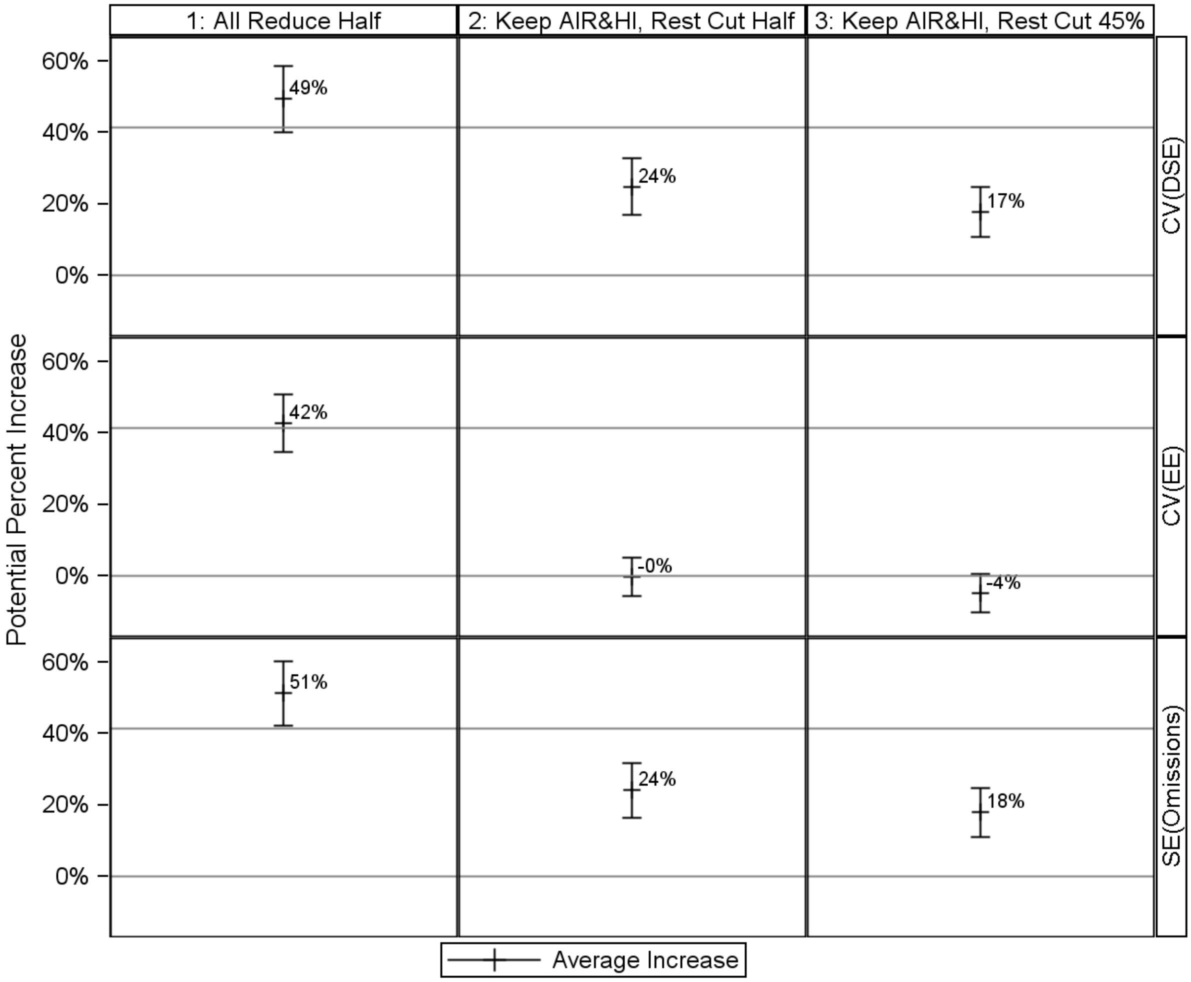
Vincent Thomas Mule, Jr.
U.S. Census Bureau
4600 Silver Hill Rd.
Bowie, MD 20716
E-mail: vincent.t.mule.jr@census.gov

Figure 1: Potential Percent Increase in Reliability Measures for
National-level Estimates
95% Confidence Interval of Average Increase for the Different Reduction Options



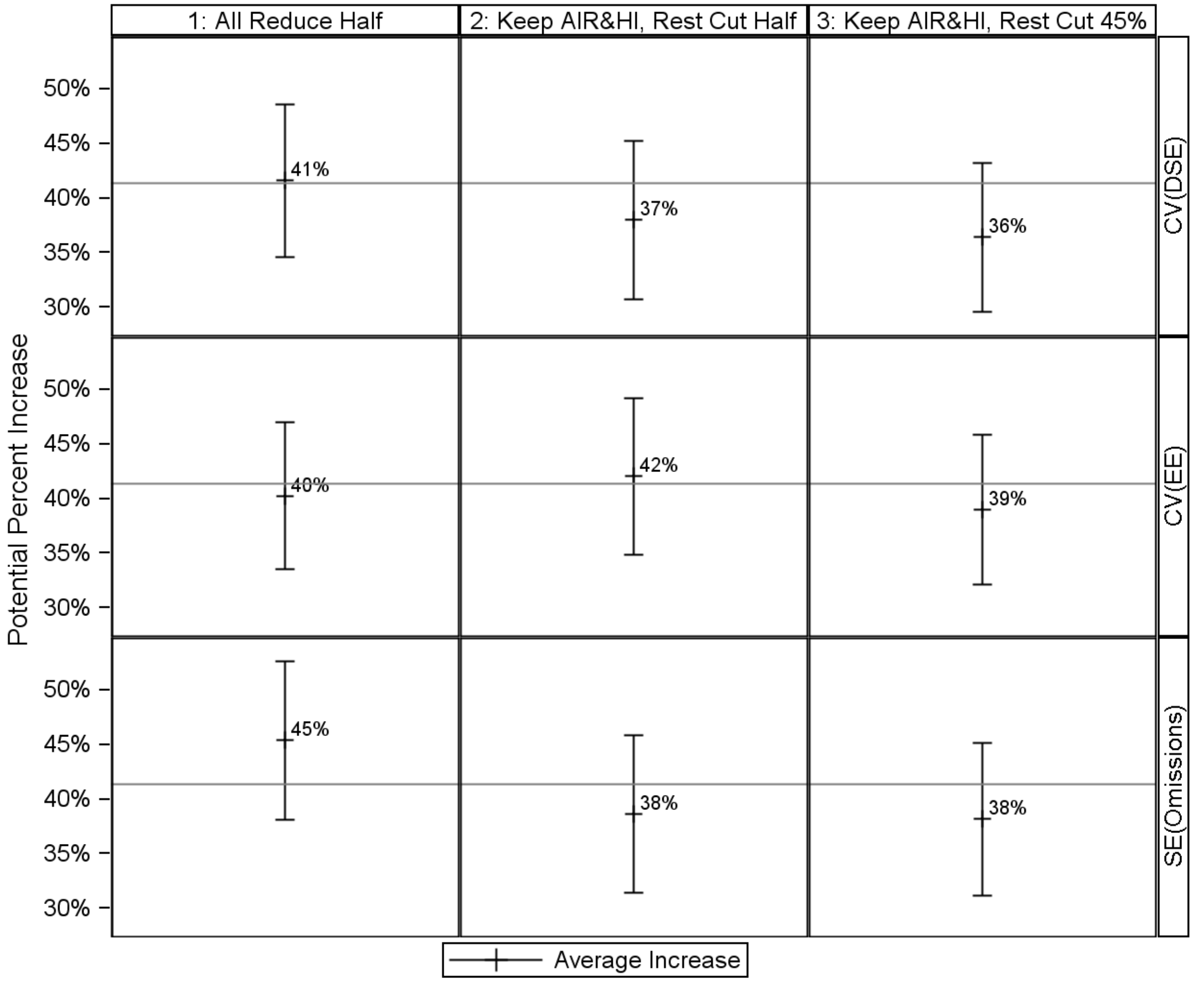
Average Increase Printed for Each Option
Reference Line at 41.4% (Approximate Increase Based on Cutting Half The Sample)

Figure 2a: Potential Percent Increase in Reliability Measures for the
American Indian on Reservation Domain
95% Confidence Interval of Average Increase for the Different Reduction Options



Average Increase Printed for Each Option
Reference Line at 41.4% (Approximate Increase Based on Cutting Half The Sample)
Reference Line at 0 Percent (No Increase)

Figure 3a: Potential Percent Increase in Reliability Measures for the
Owners
95% Confidence Interval of Average Increase for the Different Reduction Options

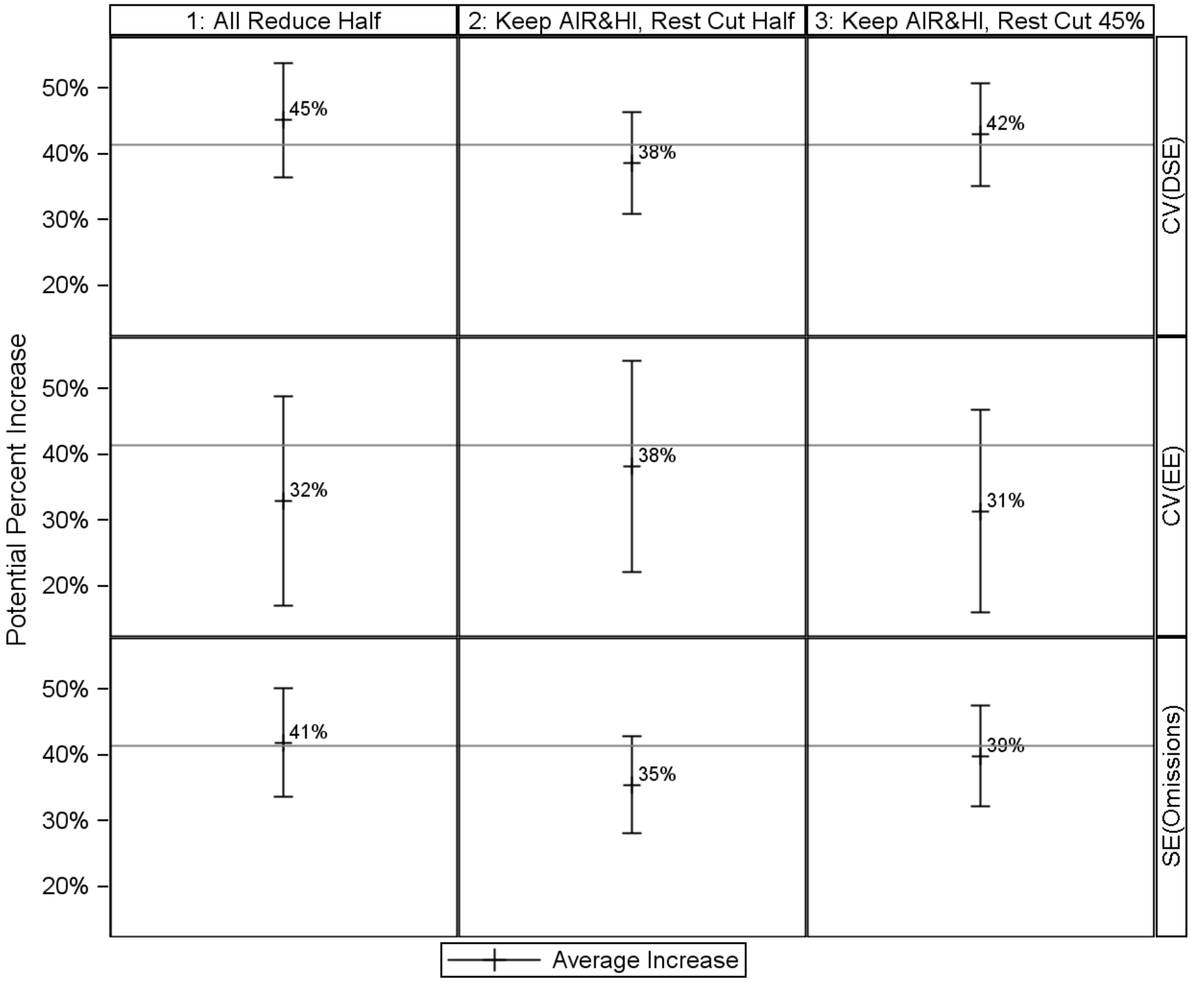


Average Increase Printed for Each Option
Reference Line at 41.4% (Approximate Increase Based on Cutting Half The Sample)

Figure 4a: Potential Percent Increase in Reliability Measures for the

Northeast Region

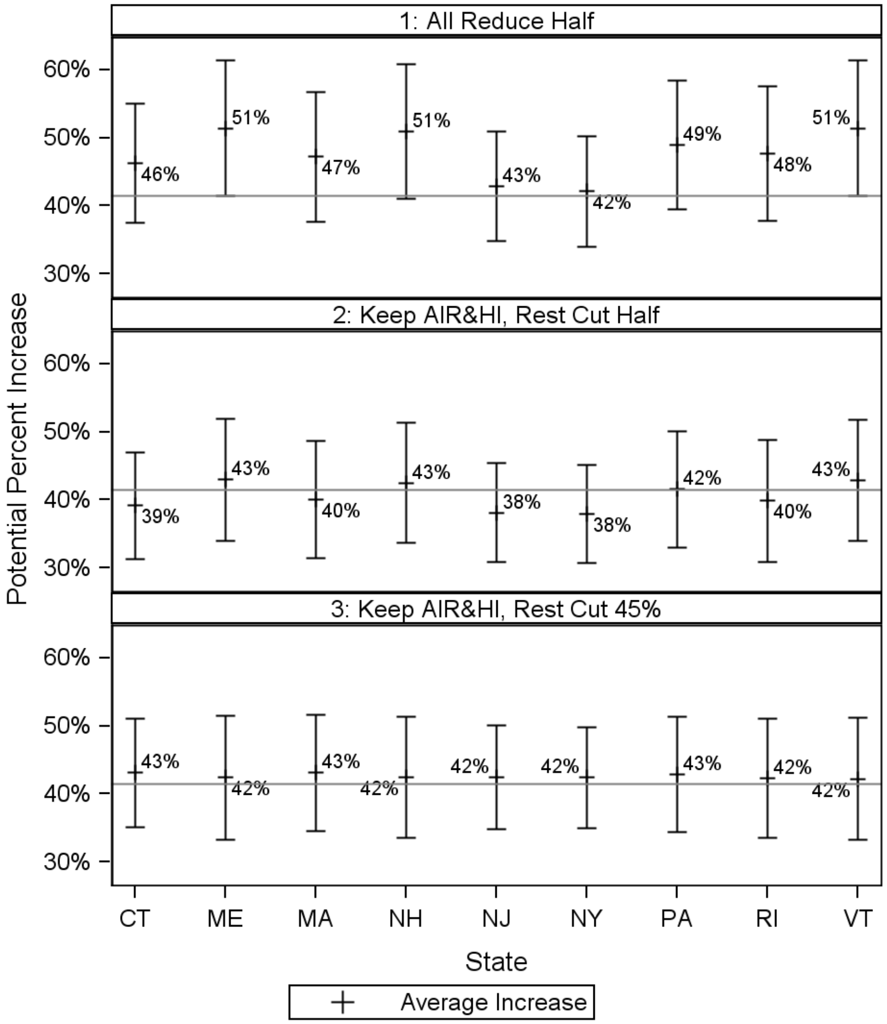
95% Confidence Interval of Average Increase for the Different Reduction Options



Average Increase Printed for Each Option

Reference Line at 41.4% (Approximate Increase Based on Cutting Half The Sample)

Figure 5a: Potential Percent Increase in
CV(Dual System Estimate) for Northeast States
Average Increase for the Different Reduction Options



Average Increase Printed for Each Option
Reference Line at 41.4% (Approximate Increase Based on Cutting Half The Sample)