2010 Census Coverage Measurement Research on Person Coverage Estimates by Housing Unit Enumeration Status

Vincent Thomas Mule Jr.
U.S. Census Bureau

Abstract

For the 2010 Census, the goals and objectives of the Census Coverage Measurement (CCM) program have been expanded to include components of coverage errors that include erroneous enumerations and omissions. Since one of the goals and objectives is to estimate the omissions by whether the housing unit was included in the census, we are researching ways to estimate the true population by whether the housing unit was enumerated or not. We present our research findings on methods for estimating the population in enumerated housing units. For this population, we examined whether we can use the housing unit matching to allow census outcome covariates to be used in the modeling of the match status.

Introduction

As part of Census 2000, the Census Bureau conducted the Accuracy and Coverage Evaluation (A.C.E.) to evaluate the net coverage error of the population in housing units. The objective was to evaluate the net error of the overall population and determine if there was a differential undercount among race/origin, tenure or age/sex groupings. To evaluate the net error for these population groups, we estimated the true population by using dual system estimation.

For the 2010 Census, the goals and objectives of the Census Coverage Measurement (CCM) program have been expanded to include components of coverage errors that include erroneous enumerations and omissions. One of the goals is to estimate omissions by whether a) the housing unit was included in the census or b) the housing unit was missed.

One of the estimators of omissions is to sum the estimate of erroneous enumerations and the estimate of net error by the following equation. By using this approach, the estimate of omissions is a function of how well we can estimate the true population.

Omissions = Net Error + Erroneous Enumerations
Where Net Error = True Population - Census and
the True Population is estimated by Dual System
Estimation

Since one of the goals and objectives is to estimate the omissions by whether the housing unit was included in the census, we are researching ways to estimate the true population by whether the housing unit was enumerated in the census.

We are researching using the same logistic regression modeling and estimation approach as was suggested by Habermann et al. (1998) and used in the initial results documented in Mule and Olson (2005). The estimation methods shown in those two documents were used to estimate the overall population for the nation and by groupings of race/origin, tenure and age/sex.

In order to estimate the population in enumerated housing units, we are exploring if we can utilize the results that are the basis of the Housing Unit Coverage Study (HUCS). By using the housing unit matching results, we can identify which Population- (P-) sample housing units were included in the census. We then restricted the logistic regression modeling and population estimation to use only the P-sample person data in the housing units that matched to the Census. We began this research by trying to use the 2000 Accuracy and Coverage Evaluation data to do this.

Potential Covariates for Modeling to Estimate Population in Enumerated

One option is similar to the post-stratification for the original A.C.E. estimates released in March 2001. This approach used person-level and housing unit-level variables that were collected independently by both systems. This approach also used geographic area covariates. One example of the geographic area covariates was what region or type of enumeration area (TEA) the housing unit was located in. Another example of a geographic area covariate was aggregating

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical and methodological issues are those of the author and not necessarily those of the U.S. Census Bureau.

the census data to form groupings based on mail return rates or the population size if the housing unit was in a Metropolitan Statistical Area (MSA). These person-level, housing unit-level or geographic-level covariates can be used in this new research. Other similar variables can be identified and used as well.

A second option is similar to the post-stratification for the A.C.E. Revision II that used separate post-stratification of the E and P sample data. While we saw that using census variables like proxy and mail return were good covariates in discriminating the correct enumeration status of the E sample, we saw that the separate post-strata resulted in some large over count results for some small areas.

A third option expands on an idea that William Bell presented to the National Academy of Science panel in August 2004. The idea is that different factors may be relevant for explaining if a housing unit is enumerated in the census as compared to a person being enumerated given that the housing unit is enumerated. If this is so, then conditioning on the housing unit being included in the census may have some benefit in the modeling and estimation. By using the housing unit matching results, we have a way of using information available for census housing units in both the E and P sample modeling and can research if doing the modeling and estimation by this approach has any benefits. This may be a way to use census-outcome variables like proxy in the modeling while minimizing the concern of using separate post-strata that was seen in the A.C.E. Revision II estimates. While considering census-outcome variables, we are keeping in mind a concern expressed in Alho et al. (1993). They had concerns about using characteristics of the census operation in the modeling to explain the census itself.

This document will show some of the preliminary work done to investigate separate estimation of the population in enumerated housing units. We will show the results of an estimation methodology that allows us to estimate the population of enumerated housing units and show some of the issues we saw in using census-outcome variables in the modeling. We show the results of a model that uses a census-outcome variable at the housing unit level as compared to two models that do not take advantage of this covariate.

II. Research Methodology

For this research, we needed to determine if each P-sample housing unit was included in the census or not. In this analysis, we used the final housing unit matching from the A.C.E. to indicate this. Based on this matching, only 2.61 percent of occupied P-sample housing units were nonmatches. Some of these nonmatches may have been caused because the census housing unit was removed from the census matching universe because of the Housing Unit Duplication Enumeration. This removal from the census matching universe had minimal impact on the net error estimates because the corresponding unit if it is in the P sample will be coded a nonmatch. However, for this analysis these nonmatches will lead to an underestimate of the population in census housing units and is a limitation of this analysis.

This matching was only done in the limited search area that was used for net error estimation. Some of the nonmatched P-sample housing units may have been enumerated in a census block that was outside of this search area. Depending on the number of nonmatched housing units that do this, the results shown here would change.

Using this available data allows us to see what the regressions and population estimates may look like and determine the course of further analysis of this topic. **These are research estimates and are not official Census Bureau estimates of coverage of Census 2000.**

Table 1 shows, as expected, that the person match rates vary if the P-sample housing unit was included in the census. The estimates in the table use a one-cell PES-A methodology that uses just nonmovers and outmovers to determine the estimate of matches and P-sample total. Both estimates in the table use the A.C.E. Revision II data. The table shows that an estimated 1.7 million P-sample people matched a census enumeration even though their P-sample housing unit was not matched to a census housing unit in the same search area.

Table 1: Person Match Rates by Housing Unit Match Status

	Occupied HU Estimate	Person Results in these HUs		
		Matches	P-sample total	Person Match Rate
HU Matched	99,048,628	233,218,814	248,721,535	0.938
HU Nonmatch	2,650,073	1,722,751	5,822,992	0.296
Total	101,698,701	234,941,565	254,544,527	0.923

Person estimates generated using a PES-A methodology

We want to examine if using census-outcome variables of the matched housing unit will allow us to a) better model the correct enumeration status of the E sample and the match status of the P-sample people in housing units that matched to the census and b) improve the population estimates.

Estimation of the Population in Enumerated Housing Units

To show an issue that arises when you use a housing unit-level outcome variable that depends on the person-level results in the census housing unit, we will show a simple example. We aggregated person-level relationships in the census housing unit to form a "family" variable for the housing unit. In the E-sample post-stratification for A.C.E. Revision II, we saw that relationship at the person-level was a strong predictor of correct enumeration status. We realize that a "family" variable for the match regression could be determined using the P-sample data but we are interested in examining using the census outcome variables in the match regressions. We are presenting these example results to show an issue we saw when trying to use person-level results aggregated to the housing unit level.

We used a simply-defined family variable that had three levels. The third level was vacant housing units. This is a category because we are using the P-sample housing unit matching results to assign this variable. Some of these P-sample housing units had nonmovers or outmovers in them but the housing unit matched to a vacant census housing unit. Any P-sample person matches in these housing units matched to a census enumeration in another housing unit.

This "family" variable had three levels:

1. Married (head of householder and spouse)
2. Not Married
3. Vacant (No people present in housing unit)

If we want to use a family variable in the actual post-stratification, we realize that we could form more levels that help discriminate the correct enumeration and match rate better. One example is whether children are present or not in the housing unit. This simple example shows how using the census-outcome variable to model the P-sample person data does help discriminate but raises an issue.

Before any modeling of this variable using logistic regression, we examined the variable as if it was a single post-stratification variable to see its ability to discriminate the correct enumeration and match status. Table 2 shows the correct enumeration and match rates by the three "family" types. The match results were calculated using a PES-A methodology and all of the estimates used the A.C.E. Revision II data. The P-sample person data are only those in the housing units that matched a census housing unit.

Table 2 shows that this type of variable is a good predictor of the correct enumeration and match status. We can see that married families have higher correct enumeration and match rates than unmarried households.

Table 2: Correct Enumeration and Match Status by Census Family Type

Family Type	CE	E-sample Total (E)	CE rate	Matches	P-sample Total	Match Rate	CE rate / Match Rate
Married	163,882,254	172,862,844	0.9480	156,426,088	162,509,247	0.9626	0.9849
Not Married	83,517,011	91,716,019	0.9106	76,260,455	83,791,531	0.9101	1.0005
Vacant	0	0	0	532,271	2,420,757	0.2199	0

Note: Match rate estimates generated using a PES-A methodology.

However, we saw an issue because some of the P-sample housing units matched to vacant housing units. Table 2 shows that for this situation that there are no E-sample cases since the E-sample is a sample of the data-defined enumerations in the census. These housing units have been determined to be vacant so there are no data-defined people in them. On the P-sample side for this post-stratification value, there were 2.4 million P-sample cases in these P-sample housing units with 1.9 million of them being nonmatches.

The issue is what do to with the P-sample people who are in housing units that match to vacants in the estimation. To show the ramifications, we show the following three possible types of post-stratifications. We collapsed these two groups together for demonstration purposes and not because we believe they have similar coverage properties

1) Include All 3 Levels
2) Two levels by collapsing Married and Non-Married into one group
3) Two levels by collapsing Non-Married and vacant into one group.

1. If we continue to have individual post-strata then the population estimate for the vacant housing unit stratum will be equal to zero. The population estimate for housing units with married families is 170.255 million and the population for non-married households is 91.765 million. This produces an overall population estimate of 262.022 million people.

2. If we collapse married and non-married into one group then the results for the new combined post-stratum are 261.874 million. The estimate of zero for vacant housing units from part 1 does not change. We collapsed these two groups together for demonstration purposes and not because we believe they have similar coverage properties.

3. If we collapse vacant and non-married families into one post-stratum then the dual system estimate for the total population is 264.077 million. The estimate of married families from step 1 does not change.

Comparison of the Three Results

We see that post-stratification #3 estimates a total population of 264.077 million for the total population as compared to post-stratification #1 and #2 where the total population estimate is approximately 2 million less. The difference is roughly the same number of nonmatches from the vacant post-stratum that are being added to the non-married or vacant estimate in post-stratification #3. We see from post-stratifications #1 and #2 the usual result that collapsing introduces some correlation bias and produces a lower overall population estimate by 146,000. However, the collapsing shown in post-stratification #3 produces a higher population estimate of almost 2 million more which is counterintuitive. What is the correct population estimate of these populations? If you want to use person-level outcomes aggregated to the housing unit-level, what do you do with the P-sample housing units that match to a vacant housing unit? These results raised issues to us with how to use person-level values summarized to the housing unit level as covariates in this research.

We would like to use person-level results of the census in the modeling and will continue to examine methods to do so. If we are unable to develop a methodology then we will focus on census outcome variables that are available for all housing units and do not depend on the person-level results in the housing unit.

Preliminary Results of Using Census Outcome Variables

In this section, we want to show some preliminary results of using census-outcome variables in the modeling to estimate the population in enumerated housing units. This is not a thorough and complete model development but a presentation of a model that uses a census outcome variable to show how this approach can be used to estimate this population.

The census outcome variable used in this model is a combination of the mail return and proxy status. The coding is the same as the mail return and proxy post-stratum variables that were used in the E-sample post-stratification for A.C.E. Revision II. P-sample housing units that matched to census vacant housing units were assigned to the non-mail proxy group. Census vacant housing units are non-mail proxy returns since it is determined in the field by someone who was not the resident on April 1st.

This new variable has been created with three levels:

1. Mail return
2. Non-mail proxy return
3. Non-mail non-proxy return

This research ran the following model that used the mail/proxy census outcome variable. The census-outcome model starts with the Race/Ethnicity domain, Age/Sex groupings and Tenure (combination referred to as "ROAST") main effects model from our previous research and adds the mail/proxy variable as an additional main effect. For simplicity, this will be referred to as the "ROAST Main with Mail/Proxy" model.

For comparison purposes, two of the models in Mule and Olson (2005) were also modeled separately in order to estimate the population in enumerated housing units. To show the extremes of those models, we used the ROAST Main Effects and the 416 post-stratification models in this framework. The 416 post-stratification was the post-stratification used in the March 2001 estimates. See Mule and Olson (2005) for more information on these models.

Evaluation of Model Fitting

To evaluate the model fitting of the logistic regression of the correct enumeration and match status, we estimated the log penalty functions for the regressions of each model. Table 3 shows the log penalty estimates of the three models for both regressions. We see that "Roast Main with Mail/Proxy" using only 16 total degrees of freedom produces a lower log penalty estimate than the 416 post-stratification that used 416 degrees of freedom. We see that this census-outcome variable is able to help improve the model fit, especially for the match status. We still need to research if there are other variables besides those in the 416 post-stratification that can improve the model fitting.

Table 3: Log Penalty Estimates of the Three Models

	dfs including intercept	Correct Enumeration Status[1]	Match Status[2]
ROAST Main Effects	14	0.18820	0.22200
416 Post-strata	416	0.18661	0.21466
ROAST Main Effects with Mail/Proxy	16	0.18474	0.19701

[1] Only sufficient information for matching and followup cases included.

[2] Using only the P-sample data in P-sample HUs matched to census HUs.

Estimates of Population in Enumerated Housing Units

We used a version of the N2 estimator from the Habermann et al. (1998) work to generate the population estimates of people in enumerated housing units. The N2 estimator, shown below, uses the weighted E-sample data and the predictions of the correct enumeration rate and match rate based on the model. In this preliminary research, we only used the sufficient information for matching and followup E-sample cases in the modeling and estimation. As the N2 estimator in Mule and Olson (2005) used all of the P-sample data in the match regression to estimate the overall population, this estimator for population in enumerated housing units uses only the regressions of the match status for P-sample data in P-sample housing units that match to the census.

$$EST = \sum_{i \in non-KE\ Esample} RTESFINWT \times \frac{r_{ce}}{r_{m\ in\ census\ HUs}}$$

Where non-KE E-sample is the E-sample cases that have sufficient information for matching and followup,

RTESFINWT is the sampling weight,

r_{ce} is the predicted correct enumeration probability and

$r_{\text{m in census Hus}}$ is the predicted match probability based on the regression of cases in P-sample housing units that matched to the Census.

Population estimates of people in enumerated housing units were generated for each of the three models tested. Since the P-sample regressions used only nonmovers and outmovers, the estimates are similar to those from a PES-A methodology.

These estimates are underestimates of the true population in enumerated housing units because of two reasons. The first is that the PES-A methodology underestimates the number of mover matches and nonmatches as compared to the PES-C methodology that was used in A.C.E. and A.C.E. Revision II. The second is that these estimates have no adjustment for correlation bias.

Table 4 shows the national estimates for the population in enumerated housing units for the three models. The table shows that the estimate of the population from the ROAST Main with Mail/Proxy is higher than the two estimates from models that don't use any census outcome variables. These differences are statistically significant because of the high correlation of the predictions from the different models. This is a preliminary result and the results after further model development with the consideration of other variables may change.

One of the objectives of using this approach is that the addition of variables available from the census can help reduce the correlation bias present in the final estimates. One way of assessing the correlation bias in population estimates is to compare the sex ratios of the survey estimates to the sex ratios from Demographic Analysis. Table 5 shows the sex ratios of the population in housing units using Demographic Analysis (Shores 2002).

Table 6 shows the sex ratios for the Black and Non-Black populations in census housing units. The results are shown for 3 age groupings (18-29, 30-49 and 50+). The table shows the resulting sex ratios from each of the 3 models. We compared the sex ratios from the ROAST Main with Mail/Proxy estimation as compared to the other two. Only the differences of the sex ratios for the Non-Black 18-29 and Non-Black 30-49 groups were significantly different. All of the differences for the three black age groups were not significantly different.

This section showed the initial research of using a census-outcome variable of mail/proxy return in the modeling and estimation of the population of enumerated housing units. We will continue to explore the approach by examining other census-outcome variables for the housing unit like:

- Was a foreign language questionnaire requested?
- Was the housing unit in the Coverage Edit Follow-up and/or Coverage Improvement Follow-up universe?

We will also examine the planning of the enumeration of housing units and the short form questionnaire for Census 2010 to see if any new changes may be possible covariates for this modeling. Any new identified changes can hopefully be examined as part of the coverage measurement testing in the 2006 Census testing and the 2008 Dress Rehearsal.

Table 4: National Research Estimates of Population in Enumerated Housing Units

		1	2	3
	Census	ROAST Main Effects	416 Post-stratification	ROAST Main with Mail/Proxy
Estimate	273,586,997	264,294,145	264,366,501	264,545,342
SE(Estimate)		2,985,176	2,989,073	2,990,379

Note: The use of 1) a PES-A methodology, 2) no correlation bias adjustments and 3) that the 2000 housing unit matching probably overstates the number of housing units not in the census are three reasons why these population estimates are probably underestimates of the true population in census housing units.

Table 5: Sex Ratios of the Population in Housing Units from Demographic Analysis

	Black	Non-Black
18-29	0.90	1.04
30-49	0.89	1.01
50+	0.76	0.86

Table 6: Sex Ratios from the Modeling Results

Model	Race/Ethnicity	Age	Male Research Population	Female Research Population	Sex Ratio
ROAST Main Only	Non-Black	18-29	18,419,340	17,608,849	1.0460
	Non-Black	30-49	36,179,997	36,461,961	0.9923
	Non-Black	50+	30,039,167	35,354,368	0.8497
	Black	18-29	2,499,733	3,005,052	0.8318
	Black	30-49	4,191,652	5,205,887	0.8052
	Black	50+	2,645,367	3,693,920	0.7161
416	Non-Black	18-29	18,488,751	17,615,830	1.0496
	Non-Black	30-49	36,196,045	36,484,442	0.9921
	Non-Black	50+	30,028,687	35,329,655	0.8500
	Black	18-29	2,453,963	2,996,384	0.8190
	Black	30-49	4,204,960	5,190,402	0.8101
	Black	50+	2,671,531	3,717,809	0.7186
ROAST Main + Mail/Proxy	Non-Black	18-29	18,521,490	17,616,789	1.0514
	Non-Black	30-49	36,202,116	36,447,614	0.9933
	Non-Black	50+	30,060,728	35,383,203	0.8496
	Black	18-29	2,518,573	3,031,054	0.8309
	Black	30-49	4,206,108	5,225,756	0.8049
	Black	50+	2,657,131	3,714,775	0.7153

We will see if we can develop a methodology to use the person-level variables aggregated to the housing unit level in the modeling and estimation. We will also examine if using census outcome variables to model both the P- and E-sample people might induce a dependence between the E- and P- samples thus creating a concern.

VI. Conclusions

We have shown our initial research into using separate estimation of the population in enumerated and non-enumerated housing units. Because of findings seen with using census person-level results aggregated to the housing unit-level, we are currently only using census outcome variables that are available for all housing units. We will continue to try to develop a methodology to use the person-level results in these research.

Our results show that using census-outcome variables is able to help improve the model fit for the regression of match status. Using this model resulted in a research population estimate for the people in enumerated housing units that was significantly higher than a model that used the 416 post-stratification for the March 2001 estimates. However, this did not result in any significant differences for the sex ratios between the ROAST Main with Mail/proxy estimate and the 416 post-stratification. Only the Non-Black 18-29 and 30-49 groups had a significant difference between the ROAST Main with Mail/Proxy model and the ROAST Main effects model. This is a preliminary result and the results after further model development with the consideration of other variables may change.

References

Alho J., Mulry, M., Wurdeman, K., Kim, J., (1993), “Estimating Heterogeneity in the Probabilities of Enumeration for Dual System Estimation,” Journal of the American Statistical Association, Volume 88 Number 423.

Habermann S.J., Jiang, W. And Spencer B.D. (1998), “Activity 7: Develop Methodology for Evaluating Model-Based Estimates of the Population Size for States Final Report,” prepared by NORC for the U.S. Census Bureau under contract no. 50-YABC-2-66023.

Mule, T. and Olson, D. (2005), “Initial Results of Preliminary Net Error Empirical Research Using Logistic Regression,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-03, U.S. Census Bureau, April 18, 2005.

Shores (2002), “Accuracy and Coverage Evaluation Revision II: Adjustment for Correlation Bias,” A.C.E. Revision II Memorandum Series PP-53, December 31, 2002.