Calibrating Components of Coverage from a Post-Enumeration Survey

Timothy Kennel[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Any views expressed are those of the authors
and not necessarily those of the U.S. Census Bureau

Abstract
Components of census coverage include correct enumerations, erroneous inclusions, and whole-unit imputations. The sum of these three components should equal published census counts. In this paper, we explore methods to calibrate post-enumeration survey estimates of enumeration status so that sums of estimated correct and erroneous enumerations equal census totals for all published variables. We also compare the performance of the calibrated estimators for totals and rates. Although not calibrated for all variables, we argue for a ratio adjustment because it is simple and results in sufficient mean squared error reductions for estimates of correct and erroneous enumeration rates.

Key Words: Post-Enumeration Survey, Calibration, Correct Enumerations, Erroneous Enumerations, Coverage Measurement

1. Introduction

The goal of the 2020 Post-Enumeration Survey (PES) is to measure the coverage of the 2020 Census. This includes measuring net coverage and components of coverage for housing units and people in housing units. The focus of this paper is on estimating components of coverage. There are four components of coverage:

- Correct Enumerations
- Erroneous enumerations
- Imputations
- Omissions

The sum of the first two components must equal the total number of data-defined census enumerations. For this paper, imputations are equivalent to non-data-defined records. Because of various errors in the PES, including sampling error, direct estimates of the correct and erroneous enumerations from the PES usually do not exactly equal published census data-defined totals. In this paper, we compare various methods to calibrate the PES estimates of correct and erroneous enumerations to census data-defined totals.

2. Background

2.1 PES Design

As name implies, post-enumeration surveys are usually done after the census. An area sample is selected and an independent enumeration of housing units and people in them is conducted in the sample areas. The independent lists are then matched to the addresses and people in the census. As a part of the matching process, sometimes followup operations are conducted to gather more information about the enumerations that are only in one system (either the census or the PES) and about potential matches.

As a result of extensive followup, the 2010 post-enumeration survey, called the Census Coverage Measurement (CCM) Survey was able to able to determine which census enumerations were correct and which were erroneous in the area sample.

2.2 Enumerations

After the census is complete, we have a list of enumerations. Ideally, every row on the list should correspond to a person, there should be no duplicates, and every person in the population should be included. However, inevitably, there are items on the list that should not be on the list. The erroneous enumerations are the enumerations on our list that should not have been included, such as people who died before Census Day or duplicates. The correct enumerations are rows that correspond to people who were counted once and only once in the census. So, we end the census journey with a list of enumerations, some of which correspond to people who were counted once, and only once, and some of which do not.

The census files also includes some people and housing units that have unresolved status or incomplete information. If enough information is missing, the row will be classified as not data-defined and it will be imputed. For example, even after multiple contact attempts and nonresponse followup attempts, we may not know if an address is occupied or vacant. In this case, the occupancy or vacancy stratus would be imputed and, if imputed as occupied, a count of the number of people in the housing unit would be imputed. Or, if during nonresponse followup, a proxy reports that there are three people at an address but does not give any other information about the people at the address, then three person records will be created for the housing unit and all of their characteristics will be imputed. Housing unit imputations and whole-person imputations are counted, rather than estimated, and neither classified as correct nor erroneous. Whole person imputations and omissions are not a topic of this paper.

Omissions are the final component of coverage. They are defined as the difference between the estimated true population number and the estimated correct enumerations. Since omissions are neither observed in the census nor the PES, they are indirectly estimated as the difference between the true population estimate and the estimated correct enumerations.

Table 1 shows the components of coverage for Puerto Rico by tenure from the 2010 CCM survey (Viehdorfer, 2012). The rows show the two levels of the tenure variable, owner and renter. The columns show the correct and erroneous enumeration rates as well as the whole-person imputation rate and the total number of enumerations.

Table 1: 2010 Census Coverage Measurement Survey Components of Coverage for Puerto Rico by Tenure

Tenure	Correct Enumerations	Erroneous Enumerations	Imputations	Total
Owner	90.0%	7.2%	2.2%	2,663,000
Renter	90.4%	7.0%	2.0%	1,024,800
Total	3,318,400	290,000	79,500	3,687,800

Source: 2010 Census Coverage Measurement

Table 1 is a fairly standard table from the coverage report for Puerto Rico. This paper will focus on the U.S. territory of Puerto Rico because it is a nice size for my simulation study. Correct enumerations, erroneous enumerations, and whole-person imputations are mutually exclusive and exhaustive. They must sum up to be the total census count.

The last column in this table shows the total number of owners and renters in Puerto Rico. According to the 2010 Census, there were 3,687,800 census enumerations in Puerto Rico. Of them, 2,663,000 owned their house.

2.3 Calibrating Estimated Components of Coverage

If we choose to estimate the total number of owners and renters from our data, it is likely that our estimates will not exactly equal census counts of owners and renters. Some people might not care if the survey estimates of owners and renters don't exactly match the census totals. Others may be deeply troubled or confused that the estimated marginal totals do not match the census. To avoid such confusion and to improve our estimates, we calibrate our weights so that they sum up to known census totals.

Calibrating survey estimates of total enumerations to census counts has the potential to greatly improve our estimates of correct and erroneous enumerations.

3. Methodology

In this section, we introduce the five estimators that we compared. We also describe the simulated population used in this study and the methods use to analyze our results.

3.1 Estimators

Horvitz-Thompson

The Horvitz-Thompson estimator, also called the expansion estimator, is a simple weighted sum using the inverse probabilities of selection. This estimator is simple and design-unbiased, but it is not statistically efficient when there are population controls. Furthermore, Horvitz-Thompson estimates generally will not equal marginal totals form the census. Thus, the Horvitz-Thompson estimator is not calibrated.

Simple Ratio Adjustment

A simple ratio adjustment could be used to force survey estimates to equal some census counts. A simple ratio adjustment can be created by multiplying design-based sampling weights for specific groups by the ratio of data-defined census counts to survey estimates. For example, in the 2010 CCM, 18 adjustment cells were created by cross classifying nine age-sex groups by two tenure groups (owner, renter). In each cell, totals from the census were tabulated and Horvitz-Thompson estimates for the totals based on the CCM survey were created. In each cell, the census total was divided by the Horvitz-Thompson estimate. The weights for all of the CCM cases in each cell were multiplied by the adjustment factor (Fox et al, 2013).

Using this method, domain totals for these 18 cells and all marginal totals based on them will match census totals, but estimates for other domains will not necessarily match census totals.

Full Ratio Adjustment

An extension of the simple cell-based ratio adjustment would be to include more variables. In this simulation, we create 72 cells by crossing:

- Nine-level age-sex groups,
- Two-level tenure groups (owner, renter)
- Two-level indicator for San Juan (in San Juan, balance in Puerto Rico)
- Two-level indicator for Black race (Black, not black)

Domain totals for all four variables will equal census totals.

Raking

Another option would be to rake the results. First, we calculate the total number of data-defined owner and renter census enumerations. We also estimate weighted totals of owners and renters from our survey. Taking the ratio of the census renters and survey estimated renters gives us an adjustment factor for renters. Similarly, we calculate a ratio adjustment for owners. Multiplying our weights by these adjustment factors calibrates our survey estimates to the census totals of owners and renters.

However, when we do this adjustment, estimated totals for other marginal will also change. So, we do the same raking process for every variable of interest. Each time, we calibrate the weights with the previous adjustment to a new domain. After calibrating once through all of the variables of interest, we repeat the process until all of the marginal estimates equal census totals.

Synthetic Estimation

Another method would be to produce synthetic estimates. Here, we fit a survey-weighted logistic regression model using our sample data after removing the whole-person imputations. Our dependent variable is an indicator if the enumeration is a correct enumeration or not. After fitting our model in the survey data, we use the estimated coefficients from our model to predict a correct enumeration propensity on the full census file, without the whole-person imputations. Estimates of correct enumerations are computed by summing up the predicted propensities on the census file. We can produce estimates for any domain on the census file.

3.2 Simulation Design

Population

To empirically compare the five estimators, we created a population to closely resemble Puerto Rico. We combined the Census 2010 Redistricting Data Files with the 2006 – 2010 American Community Survey (ACS) Public Use Microdata Areas (PUMA) data of people and housing units to create a simulated population of housing units and people in Puerto Rico. Specifically, we first read in the Census 2010 Redistricting Data Files for Puerto

Rico. These files contained the total number of people counted in the 2010 Census for each block. For each block, we created empty person records based on the census count. For example, if a block had 16 people, we created 16 blank records in the block. Then, we sorted the ACS PUMA file of people by housing unit ID and person ID. We cycled through the ACS PUMA file, inserting people into the census universe until it had 3,676,054 people. Altogether, the simulated frame had 50,796 blocks with 3,676,054 people in 1,633,387 housing units. Because the source of the person records was the ACS PUMA, we had a wealth of information about each person.

Dependent Variable

The ACS PUMA did not contain an indicator for the component of coverage. To simulate the component of coverage, we randomly assigned each person to be a correct enumeration, erroneous enumeration, or whole person imputation within various domain groups. Table 2 shows the result of the correct enumeration assignment.

Table 2: Simulated Correct Enumeration Rates for Frame

Domain	Total	Correct Enumerations	Correct Rate
All People	3,676,054	3,306,716	90.0%
Renters	994,444	885,109	89.0%
San Juan	2,445,240	2,208,670	90.3%
Black	552,819	456,765	82.6%
College	1,202,271	1,080,943	89.9%
English	3,183,822	2,865,130	90.0%
Mover	284,048	203,846	71.8%

Source: Simulation

The rows in Table 2 show various characteristics of people. The first column shows the domain total. Then, the second column shows the total correct enumerations. The third column shows the correct enumerations divided by the total. As we see, the correct enumeration rate for most groups is about 90 percent, with the exception of people who identify as Blacks and people who have moved in the last year. The total, renters, and San Juan coverage rates are derived from the 2010 CCM reports for Puerto Rico (Viehdorfer, 2012), the last four rates in the table are made up rates that are not based on the prior reports.

Sampling

After creating the simulation frame, we selected 100 samples from it. Similar to the design of the 2010 post-enumeration survey in Puerto Rico, we stratified the blocks based on size and selected systematic samples of size 316 blocks, about the same number of blocks selected for the 2010 CCM survey.

For each of the 100 samples, we estimated the components of coverage for a variety of domains using the five estimators presented. To summarize across the 100 estimates, we calculated the root mean squared error and present them in the results section.

Estimation Models, Cells, and Domains

To explore how the estimation methods performed under a variety of situations, we used different combinations of variables for the adjustments. Table 3 summarizes the variables used in each estimation method. The rows in this table show different variables. A "yes" indicates the variables was used. The Mechanism column indicates which variables were used to determine whether an enumeration was correct or erroneous. The simple ratio column indicates that the tenure and age sex variables were used to form cells for the simple ratio.

In general, including a variable in the adjustment will improve the estimate because including the variable will calibrate the marginal totals. For example, the simple ratio, full ratio, raking, and synthetic estimates should all produce estimates of owners and renters that are exactly the same number on the simulated frame. As we saw in Table 2, the mover variable had a very low correct enumeration rate. Since this variable was not included in any of the adjustments, we would expect all of the estimators to be biased.

Table 3: Summary of Variables in Estimation Methods

Variable	Mechanism	Simple Ratio	Full Ratio	Rake and Synthetic
Tenure	Yes	Yes	Yes	Yes
Age and sex	Yes	Yes	Yes	Yes
San Juan indicator	Yes	-	Yes	Yes
Black indicator	Yes	-	Yes	Yes
College indicator	-	-	-	Yes
Mover Indicator	Yes	-	-	-
English Spoken at Home	-	-	-	-

Evaluation

To evaluate each of the estimators, we calculated the root mean squared error. The root mean squared error combines the bias and variance of an estimator. It is measured by averaging the squared difference between the survey estimate and the simulated true population value across all 100 samples. Estimators with large root mean squared error are undesirable. Specifically, the root mean squared error is defined as:

$$RMSE = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\hat{t}_i - T)^2}$$

Where

$\hat{t}_i$ is the estimated total (or rate) for sample *i*

T is the total (or rate) on the frame

and *i* indexes each of the 100 samples.

4. Results

4.1 Variables included in all models

Table 4 shows the RMSE for the total number of correctly enumerated renters, erroneously enumerated renters, and total renters using each estimation method. Each row corresponds to one of the estimation methods. The correct column shows the RMSE for the estimated number of correctly enumerated renters. The next column shows the RMSE for the estimated number of erroneous enumerations. Lastly, we see the root mean squared error for the renters.

Table 4: Root Mean Squared Error for Renters (Total)

Estimator	Correct	Erroneous	Total Renters
HT	50,737	8,631	57,668
Simple Ratio	4,857	4,857	0
Full Ratio	4,743	4,743	0
Rake	4,833	4,833	0
Synthetic	4,813	4,813	0

Source: Simulation

We note that the simple ratio, full ratio, rake, and synthetic methods all have zero RMSE for the estimated total enumerations. This of course is by design because there is no variability of the estimated total number of enumerations across the 100 samples, and the estimates are unbiased. The Horvitz-Thompson estimator is also unbiased, but the sum will vary from sample to sample.

Controlling to the total number of renters reduces the root mean squared error of cell estimates. Even though the synthetic estimate of the joint cell total could be biased, we do not see much evidence of that from this table. Of course, tenure was included in our model, so we would expect low bias.

Table 5 shows the RMSE for the correct enumeration rate and erroneous enumeration rate. Since the correct and erroneous enumeration rates are mutually exclusive and exhaustive of the set of data defined enumerations, their sum is always 100%.

Table 5: Root Mean Squared Error for Renters (Percent)

Estimator	Correct	Erroneous
HT	0.49	0.40
Simple Ratio	0.49	0.40
Full Ratio	0.48	0.38
Rake	0.49	0.40
Synthetic	0.48	0.40

Source: Simulation

As seen in Table 5, the RMSE for all five estimates is less than half of one percent. In terms of the RMSE, all five estimates perform similarly.

As Table 4 and Table 5 show, the performance of the five estimates depends on whether levels or percents are being estimated. For percents, all five estimators perform similarly. But, for levels, there are clear advantages to calibrate the survey weights to marginal totals.

4.2 Variables included in full ratio, rake, and synthetic models

The general guideline to control to the estimation domain is reinforced by Table 6. In this case, the San Juan indicator was used in the Full Ratio, Rake, and Synthetic estimators, but was not used in the Horvitz-Thompson and Simple Ratio estimators.

Since the RMSE is zero for the full ratio, rake, and synthetic estimates, it is abundantly clear from Table 6 that controlling to a variable results in design-unbiased estimates with zero variance for the marginal totals. Even though the Horvitz-Thompson estimator and Simple Ratio estimators are design-unbiased, their estimates of the total number of people in San Juan changes from sample to sample. As a result of sampling error, the RMSE is not zero for the estimated number of people in San Juan. We also see that the RMSE for the Simple Ratio is less than the total for the Horvitz-Thompson estimator. Even though the weights were only calibrated to age-sex and tenure status, the RMSE is lower for the Simple Ratio than the Horvitz-Thompson estimator. Calibrating to a small set of variables can reduce the variance of a large set of estimates that are correlated to the variables being controlled.

Table 6: Root Mean Squared Error for San Juan Indicator (Total)

Estimator	Correct	Erroneous	Total San Juan
HT	127,635	16,179	142,707
Simple Ratio	81,867	11,065	91,182
Full Ratio	5,569	5,569	0
Rake	5,691	5,691	0
Synthetic	5,588	5,588	0

Source: Simulation

A flag indicating if the enumeration was in San Juan was included in the full ratio, raking, and synthetic methods, but not in the simple ratio. Even though the simple ratio has a lower root mean squared error than the Horvitz-Thompson estimator, it does not perform as well as the other methods. So, our takeaway from this table is that we should include as many variables as we can in our estimation process.

A table showing the RMSE of the correct and erroneous enumeration rates for San Juan is not shown, but the results are similar to those in Table 5. All estimators have similar empirical RMSE estimates.

4.3 Variables included in rake, and synthetic models

Table 7 shows results where the raking and synthetic estimate estimators are controlled to the domain totals, but the other methods do not control to the variable. As we saw in Table 6, the RMSE for the ratio adjustments is lower than the Horvitz-Thompson estimator. In fact, for the outcome of attending some college, the simple and full ratio adjustments reduce the RMSE considerably, compared to the Horvitz-Thompson estimator. Since the Horvitz-Thompson, Simple Ratio, and Full Ratio methods are all unbiased, this reduction in RMSE is due to variance alone.

Table 7: Root Mean Squared Error for Some College (Total)

Estimator	Correct	Erroneous	Total College
HT	58,565	8,782	66,263
Simple Ratio	12,683	4,519	13,841
Full Ratio	12,936	4,528	13,991
Rake	4,199	4,199	0
Synthetic	4,185	4,185	0

Source: Simulation

Rake and Synthetic estimators control to population totals of some college attendance, so their estimates of total college attendance have zero RMSE. As we would hope, calibrating to this variable results in reductions in the RMSE for the total number of correct and erroneous enumerations.

A table showing the RMSE of the correct and erroneous enumeration rates for some college attendance is not shown, but the results are similar to those in Table 5. All estimators have similar empirical RMSE estimates.

4.4 Excluded variables

First, we investigate a variable that is not related to our outcome variable, English not spoken well at home. None of the methods use this variable. As we see from Table 8, there are substantial reductions in RMSE for the calibrated estimates, even when this variable is not directly calibrated to. Indeed, the reduction in RMSE for the calibrated estimators compared to the Horvitz-Thompson estimator is rather dramatic.

Table 8: Root Mean Squared Error for English Not Spoken Well at Home (Total)

Estimator	Correct	Erroneous	Total Non-English
HT	151,265	18,957	168,692
Simple Ratio	12,597	6,991	10,900
Full Ratio	12,248	7,213	10,954
Rake	11,913	7,213	10,621
Synthetic	7,634	7,634	0

Source: Simulation

The synthetic estimates use the full population file to produce estimates. Since the domain indicators are all on the simulated population file, domain estimates will equal the population total for all domains. Thus, Table 8 shows zero RMSE for the Synthetic estimator of total people who do not speak English well at home.

A table showing the RMSE of the correct and erroneous enumeration rates for English not spoke well at home is not shown, but the results are similar to those in Table 5. All estimators have similar empirical RMSE estimates.

A second case is for variables that are not calibrated to, but are correlated with correct enumeration status. Although it would be desirable to list all of the domains in advance and calibrate to them all, this is not realistic. After producing initial estimates, there are often requests to see results for new unplanned domains. If there is not enough time to reweight the data before producing tables, we may need to produce estimates that have not been calibrated. Table 9 shows the RMSE for the correct and erroneous estimates of total people who have moved in the past year. Recall from Table 3 that the movers have a correct enumeration rate of 71.8%, well below the population average of 90.0%.

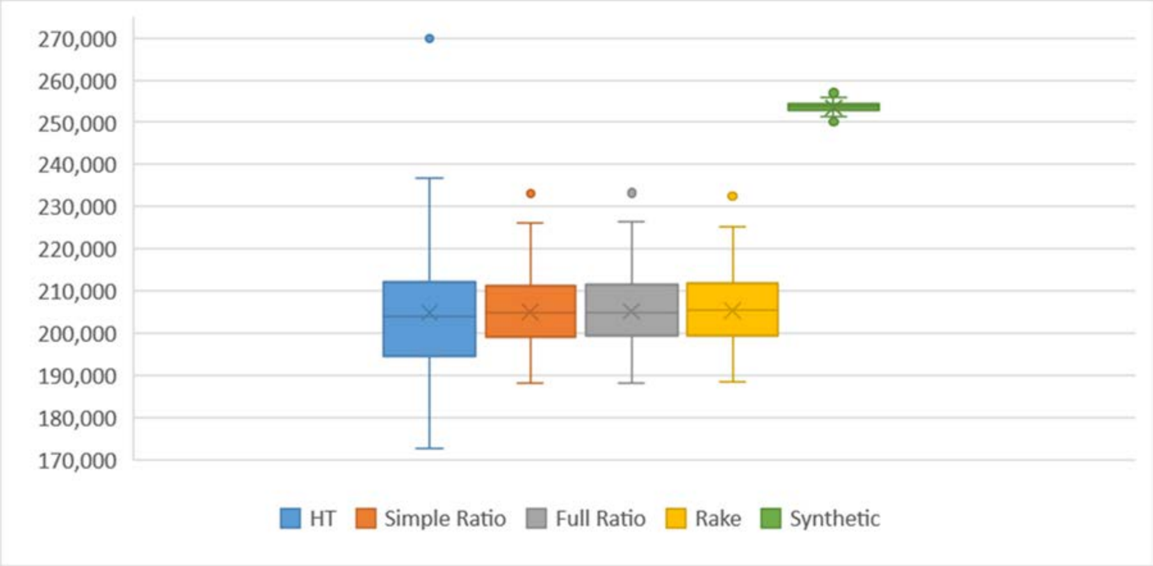
Table 9: Root Mean Squared Error for People who have Moved in the Past Year (Total)

Estimator	Correct	Erroneous	Total Movers
HT	14,247	6,849	19,880
Simple Ratio	8,906	4,670	11,639
Full Ratio	8,882	4,682	11,629
Rake	8,884	4,747	11,704
Synthetic	49,716	49,716	0

Source: Simulation

Although the Synthetic estimator always correctly estimates the number of movers exactly in the population, the estimates of correct and erroneous enumerations for the movers suffer from a large RMSE. As we see in Figure 1, the Synthetic estimate of correct enumerations who are movers is very precise, but quite biased. This bias explains the large RMSE. In general, the Synthetic estimator preforms very well. However, for domains that are not included in the synthetic model as covariates and are correlated with the dependent variable, the Synthetic estimator can be biased.

Figure 1: Box and Whisker Plot of Correct Enumeration Level for Movers (Levels)



Source: Simulation

For all of the other tables, the RMSE for the rates were similar, regardless of the estimator. This is not the case for the movers. In this case, the Synthetic estimator is biased and has a RMSE much larger than the other estimators. Table 10 shows the RMSE for the estimated percent correct and erroneous enumerations for people who moved in the previous year. The Synthetic estimator is clearly different from the other estimators.

Table 10: Root Mean Squared Error for People who have Moved in the Past Year (Percent)

Estimator	Correct	Erroneous
HT	1.13	1.14
Simple Ratio	1.13	1.13
Full Ratio	1.13	1.13
Rake	1.14	1.13
Synthetic	17.50	17.50

Source: Simulation

4.5 Summary

In this section, we explored how five estimators performed under a variety conditions. It is clear that only the Synthetic estimator always results in calibrated marginal totals. So, if exactly matching population totals is necessary, the Synthetic estimator is the best estimator. However, for domains that are not included in the synthetic model as covariates, estimates of correct and erroneous enumerations can be biased and highly variable. If we do end up producing estimates for domains that are not included in the adjustment, we may want to compare several estimators and suppress the tables where the estimates are not similar.

On the other hand, if rates and percentages are of primary interest, then all estimators produce similar estimates, with the exception of the synthetic estimator. When a domain is not included in the model, but is correlated with the dependent variable, the synthetic estimator can suffer from large root mean squared error. Since the simple and full ratio adjustments are simple to perform and offer some calibration for levels, the ratio adjustment might be the best choice for rates.

5. Conclusion

Only the synthetic estimator will force all domain estimates to match census totals. Since census totals exist for all domains and inference is made about the universe of data-defined enumerations in the census, it is important for the sum of correct and erroneous enumerations from the PES to equal published counts of census enumerations. However, in 2010, the post-enumeration survey results focused on the correct and erroneous enumeration rates, rather than totals. Only a few tables showed levels for components of coverage. Thus, for the 2010 post-enumeration survey, a simple ratio adjustment would likely have resulted in estimates with similar properties and been simpler to compute. Since correct and erroneous enumeration rates are of greater interest, the recommendation for the 2020 PES is to use a ratio adjustment for rates.

We saw evidence of reductions in RMSE when more domain totals were calibrated to. Further research is needed to determine which domains should be controlled to for the 2020 PES. It may be advantageous to include more variables than what was used in the 2010 ratio adjustment.

Although synthetic estimation is attractive because all domains are inherently calibrated to census totals and the synthetic method is already used for net coverage estimates, there are clear disadvantages for domains that are correlated with enumeration status, but not included in the model. Certainly, one can produce direct estimates of the associated between various domains and enumeration status to determine if this is the case. And, the variables that are determined to be strongly correlated with enumerations status can be included in the synthetic model. However, after the model is fit and estimates are produced, one must expect biased estimates for new domains that are strongly correlated with enumeration status. For, rerunning the model to include the new variable would result in changes to all of the previously reported estimates.

References

Fox T., Keller A., and Davis, P. (2013). “2010 Census Coverage Measurement Estimation Methods: Component Estimation Methodology.” DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-J-05.

Mule, T. (2008). “2010 Census Coverage Measurement Estimation Methodology.” DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-E-18, Retrieved from https://www.census.gov/coverage_measurement/pdfs/2010-E-18.pdf.

Viehdorfer, C. (2012). “2010 Census Coverage Measurement Estimation Report: Results for Puerto Rico.” DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-13, Retrieved from https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g13.pdf.

Disclosure

This paper meets all of the U.S. Census Bureau’s Disclosure Review Board (DRB) standards and has been assigned DRB approval number DRB-B0002-DSSD-20180918.