Assessing the Impact of the Final Housing Unit Followup on the 2010 Census Coverage Measurement Housing Unit Estimates[1]

Michael Beaghen[1], Anne Wakim[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract
The 2010 Census Coverage Measurement (CCM) program measured the coverage of the 2010 Census enumeration of persons and housing units. The CCM independently listed housing units and persons in a sample of geographies. The census and CCM listings were compared and matched where possible, and field operations were conducted to resolve differences. The Final Housing Unit Followup (FHUFU) was the last of the CCM field operations; it primarily processed late changes to the census inventory of housing units. The research described in this paper assessed what the impact would be of not conducting the FHUFU field operation on the 2010 CCM estimates of housing unit coverage. It informs a decision whether to remove the FHUFU from the 2020 Coverage Measurement program.

In the study, housing units which went to FHUFU for resolution were assigned unresolved match and housing unit statuses. These unresolved match and housing unit statuses were then imputed to simulate what would have happened if FHUFU had not been conducted. The modified data with the imputations were used to generate alternative estimates. The alternative estimates were compared to the 2010 Census coverage estimates to assess the impact of removing the FHUFU.

Key word: imputation, Post-Enumeration Survey, Dual System Estimation

1. Introduction

Starting with the 1950 Census, the U.S. Census Bureau has conducted post-enumeration surveys (PES) to evaluate the census coverage of the population of persons and housing units (Fay et al., 1980). Similarly, the Census Bureau will conduct a PES to assess the coverage of the 2020 Census and to aid in the design of future censuses. The 2020 PES will be a probability sample of about 180,000 housing units nationwide (Trang, 2017). Remote areas of Alaska, group quarters facilities, and persons residing in group quarters facilities are out of scope for the PES. The PES will also have a sample of about 10,000 housing units in Puerto Rico. The PES will support the estimation of census net coverage and components of coverage for the populations of housing units and people.

The FHUFU will be the last field operation of the 2020 PES and will be essential to producing PES estimates of the 2020 Census housing unit coverage. The PES will first conduct the Initial Housing Unit operation, which will begin with a field operation to create an address list of all housing units in the geographic areas selected for the 2020 PES. The

[1] Any views expressed in this report on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

listing will be conducted in the winter of 2020, and will be completely independent from the 2020 Census operations. This independent listing of housing units will be matched to a preliminary inventory of 2020 Census housing units available at that time. The PES Final Housing Unit operations will process changes to the census inventory of housing units from the preliminary census listing to the final census listing, that is, the final 2020 Census listing of housing units. The Final Housing Unit processing will begin by matching the PES addresses to the census enumerations during the Final Housing Unit Before Followup Matching operation. A clerical review will be conducted to resolve discrepancies between the files and determine if addresses were correctly listed. Those PES addresses and census enumerations that remain unresolved will be sent to FHUFU. The data collected in FHUFU will then be clerically reviewed to determine their match status and their Census Day (April 1, 2020) housing unit status.

The purpose of this study was to investigate the feasibility of imputing for housing unit data that previous Census Bureau PES programs have collected in the Final Housing Unit Followup (FHUFU) operation. In an effort to reduce costs and respondent burden and produce more timely estimates, the Census Bureau considered not conducting a FHUFU operation for the 2020 PES. If the 2020 PES program eliminated FHUFU, it would still plan to conduct the Final Housing Unit Before Followup Matching operation. However, housing units with unresolved status that would have gone to FHUFU would remain unresolved, resulting in the need to impute for these cases to produce estimates. In addition, any other changes to match codes as a result of FHUFU would not be reflected in the data provided to produce estimates.

The research described in this paper assessed the impact of not conducting the FHUFU field operation on the 2010 Census Coverage Measurement (CCM) estimates of net housing unit coverage. It informed the decision whether to remove the FHUFU from the 2020 PES. The 2020 PES will have a design similar to that of the 2010 CCM survey, which included a FHUFU operation. We believe we can safely generalize from the results of this study with 2010 CCM data to the 2020 PES.

In the study, we assigned an unresolved status to those housing units which went to FHUFU, simulating what would have happened if FHUFU had not been conducted. We then imputed for the unresolved enumeration, housing unit, and match statuses using the 2010 CCM imputation methods. We used these modified data to generate alternative estimates. We compared the alternative estimates to the 2010 CCM estimates to assess the impact of removing the FHUFU.

2. Background

Since we used only 2010 CCM data in this study, we describe the 2010 CCM methodology.

The 2010 CCM program measured the coverage of the 2010 Census enumeration of persons and housing units. The 2010 CCM sample in the U.S. (excluding Puerto Rico) consisted of about 6,000 block clusters. Each block cluster consisted of one or more geographically contiguous census collection blocks grouped together to form an average of about 30 housing units. The CCM independently listed housing units and persons in this sample; these listings constituted the population or P sample. The 2010 CCM housing unit matching operation comprised two overlapping samples, the P sample and a sample of census housing unit records (the E sample) enumerated in the same set of census blocks selected for the P sample.

2.1 Census Coverage

Coverage refers to how completely and accurately the Census enumerates the population. Coverage errors include omissions and erroneous enumerations. Erroneous enumerations include duplicated enumerations of housing units and those that should not have been enumerated anywhere on Census Day, such as an empty lot. Furthermore, to be correctly enumerated, housing units should have been counted in their basic search area, an area defined by the sample block cluster and a surrounding ring of blocks. If a housing unit was enumerated outside of its basic search area it was a geocoding error, another type of erroneous enumeration.

The CCM estimate for the population was referred to as the dual system estimate, or DSE. The net coverage error is defined as the net undercount, where

$$Net\ Undercount = DSE - Census\,Count$$

2.2 The 2010 Census Coverage Measurement

We discuss in more detail the 2010 CCM housing unit operations (Whitford, 2008).

2.2.1 CCM Initial Housing Unit Operations

In the autumn of 2009 the Census Bureau conducted an independent listing of all of the housing units within the CCM sample block clusters. This independent listing was matched to a preliminary 2010 Census listing of housing units in the same sampled geographic areas. This preliminary census listing was the 2010 Census's most current inventory of housing units as of about January, 2010. The independent listings were compared to the preliminary census listing to determine which housing units matched, a requirement of dual system estimation. Housing units from the two lists were first linked using a computer matching operation. All matches were reviewed in a clerical matching operation, where additional matches were identified. Nonmatched independent listings and census enumerations, and matches that required more information, were sent to the Initial Housing Unit Followup, where interviewers collected more information in the field. This additional information was processed by clerks in an Initial Housing Unit After Followup Clerical Matching operation.

2.2.2 CCM Final Housing Unit Operations

The Final Housing Unit operations were a last stage of CCM processing to address the changes to the census inventory of housing units from the preliminary census listing to the final census listing. These changes were of two types: 'adds', or housing units new to the final census listing: and 'deletes', housing units that appeared on the preliminary census listing but were not on the final census inventory of housing units. The census and CCM independently-listed housing units were rematched. Again, differences were resolved with a field interview, the FHUFU. In the 2010 CCM, the FHUFU took place in the spring of 2011.

Most addresses included in the 2010 Final Housing Unit Matching operation did not require FHUFU for resolution because their match status was fully resolved in the Final Housing Unit Before Followup Clerical Matching operation. However, certain addresses required followup to obtain additional information to be resolved. If FHUFU was unable to provide

the necessary data, the case remained unresolved and it was assigned an appropriate code to reflect the unresolved status.

2.3 Census Coverage Measurement Imputation Methodology

A census housing unit enumeration in the E sample or a CCM independently-listed P-sample housing unit could have had an unresolved status because of a noninterview or because there was not enough information collected in the Initial Housing Unit Followup or the FHUFU to provide resolution. If CCM operations could not determine whether a P-sample listing was a valid housing unit in the sample area on Census Day, then it was assigned an unresolved housing unit status. If CCM operations could not determine whether a P-sample housing unit listing was matched to a census housing unit enumeration, then it was assigned an unresolved match status. If CCM operations could not determine the housing unit status of an E-sample housing unit for Census Day, it was assigned an unresolved enumeration status.

For those records with unresolved statuses, the CCM used logistic regression models to impute for probabilities required for estimation (Konicki et al., 2013). There were three models, one for each housing unit status. Table 1 shows the independent variables used in each of the models. An 'X' indicates that the variable was used in the model for the status listed at the head of the column.

Table 1: Model Variables Used in Status Imputation for Housing Unit, Match, and Correct Enumeration Status

Variable	Housing Unit	Match	Correct Enumeration
Transformed Address Canvassing Rate	X	X	X
Transformed Enumeration Rate	X	X	X
Metropolitan Statistical Area/Type of Enumeration Area Group	X	X	X
Occupancy/Tenure	X	X	X
Region	X	X	X
Recoded Housing Unit Type of Address	X	X	

Table 2 describes in greater detail the variables used in the imputation model. It also includes a variable, E-sample Before Followup Match Code Group, which was not used in the 2010 CCM imputation models for housing units. We included this variable in the enhanced imputation, which we discuss in Section 3.

Table 2: Model Variable Descriptions for Status Imputation for Net Coverage of Housing Units

Variable Description	Variable Name	Valid Values
Transformed Address Canvassing Rate of the Tract	TRADCAN_RT	Numeric value
Transformed Enumeration Rate of the Tract	TRENUM_RT	Numeric value
Metropolitan Statistical Area by Type of Enumeration Area Groups	MSATEA	0: Large MSA, Mailout/Mailback TEA 1: Medium MSA, Mailout/Mailback TEA 2: Small MSA, Mailout/Mailback TEA 3: Non-MSA, Mailout/Mailback TEA 4: Large, Medium, or Small MSA, Update/Leave TEA 5: Non-MSA, Update/Leave TEA 6: Update/Enumerate TEA
Occupancy/Tenure	OCCTEN	1: Occupied by owners 2: Occupied by renters 3: Vacant or not a housing unit on Census Day
Region	REGION	1: Northeast 2: Midwest 3: South 3: West
Recoded Housing Unit Type of Address	HUTOA_NEW	1: Single units and other types of households 2: Multi-units
E-sample Before Followup Match Code Group	EBFUMCG	1: Resolved Before Followup 2: Possible Matches 3: Conflicting Household 4: Partial Household Nonmatch 5: Whole Household Nonmatch 6: Unresolved Inclusion Status 7: Duplicate 9: Insufficient information for dual system estimation

2.4 Census Coverage Measurement Estimation Methodology

The 2010 CCM relied on dual system estimation for its estimates of net census coverage. Dual system estimation requires two independent systems of measurement. In the CCM,

these were the P sample and the E sample, which measured the housing unit population in the same sample block clusters. After matching to the census lists and field reconciliations, the P sample provided information about the housing units missed in the census, whereas the E sample provided information about erroneous census enumerations.

The 2010 CCM used logistic regression modeling to estimate the parameters in the dual system estimation formula for correct enumeration and match probabilities (Olson and Viehdorfer, 2013), instead of the cell-based post-stratification used for previous coverage estimates (U.S. Census Bureau, 2004).

The DSE for housing units in estimation domain C can be expressed as

$$DSE_C = \sum_{j \in C} \frac{\pi_{ce(j)}}{\pi_{m(j)}}$$

With respect to the given estimation domain C, the predicted correct enumeration and match probabilities for census case j ($\pi_{ce(j)}$ and $\pi_{m(j)}$, respectively) were obtained through logistic regression modeling.

We refer to DSE_C as a synthetic estimate of the domain C. The parameters in the model were based on a national sample and then applied to each individual census case. Information collected at the individual housing unit level could be easily used in conjunction with information collected at a more aggregate level to provide estimates even for small domains with little or no sample.

The main effects used in the logistic regression models for the DSE included

- Structure type and size of the dwelling
- Occupancy and tenure
- Region of the country
- Metropolitan Statistical Area size by Type of Enumeration Area
- Measures of the number of address list changes in the neighborhood near to Census Day
- Bilingual and Replacement Questionnaire Mailing Areas

2.5 Variance Estimation Methodology

The 2010 CCM used delete-a-group jackknife replication to estimate standard errors of net coverage (Imel et al., 2013). There were 100 groups formed from the block clusters. The CCM did not attempt to account for the variance of imputation. In our study we used the 2010 CCM variance estimation methodology.

3. Study Methodology

There were two basic steps in our study

1. Identify the cases that went to FHUFU and recode them as unresolved.
2. Impute for the unresolved cases and produce estimates.

The recoding to simulate eliminating the FHUFU was described in detail in Beaghen and Wakim (2017). In the 2010 CCM, there were 72 unweighted E-sample housing unit enumerations with an unresolved enumeration status after FHUFU (this number excludes

Puerto Rico). After the recoding for this study there were 2,336 unweighted E-sample housing unit enumerations with unresolved enumeration status that required imputation. These 2,336 cases weighted up to over 1.7 million housing unit enumerations. Due to a one-time anomaly in the 2010 CCM final housing unit processing (Mule, 2011), most P-sample housing units that should have been sent to FHUFU were not. Consequently, the study recoding had only a modest effect on the P-sample housing unit coding and imputation.

The study imputation and estimation methodologies were the same as those for the 2010 CCM, which were described in the previous sections. The one place where we deviated from the 2010 CCM estimation methodology was that we explored a second, enhanced imputation for the E-sample enumeration status. In other CCM imputations, the Before Followup Match code had proven to be of predictive value (Konicki et al., 2013, and U.S. Census Bureau, 2004). Thus we added the variable "E-sample Before Followup Match Code Group," or EBFUMCG, to the independent variables already in the model. The results with this additional variable are presented and discussed in Section 5, in Table 4.

4. Limitations

The most important limitation may be that because of an anomaly in 2010 CCM data processing (Mule, 2011), the study's recoding resulted in few new unresolved cases in the 2010 CCM P-sample housing units. With more unresolved cases, the impact of eliminating the FHUFU could have been greater.

In addition, the following two considerations about the data should be noted when reading this document.

4.1.1 Sampling Error

Because the CCM estimates were based on a sample survey, they were subject to sampling error. The standard errors provided with the data reflect variation due to sampling.

4.1.2 Synthetic Error

In calculating the DSE of the population, we created a synthetic estimator, as described in Section 2.4. The estimation domains were subject to a potential synthetic bias. The bias in the synthetic estimator represented the difference, if any, in the domain's population estimate one would obtain by applying the synthetic model versus by simply tabulating over the true population (if it were known). For most estimation domains, main effects and interactions related to the domain were included in the dual system estimation models to minimize the synthetic bias in the population estimates. Otherwise, in our study, we did not account for synthetic error.

5. Results

We refer to the DSEs with no FHUFU as the alternative DSEs. Tables 3 and 4 show alternative DSEs broken down by occupied/vacant, census type of enumeration area, and census region. Table 4 also includes alternative DSEs with an enhanced imputation. The enhanced imputation includes the Before Followup Match Code Group variable in the imputation model. All results exclude Puerto Rico and remote areas of Alaska.

In Table 3, we see the breakdown by occupied and vacant housing units. Note that the net undercount is the DSE minus the census count, 132,467,000 minus 131,676,000, or 790,000[2] (all figures were rounded to the thousands to be consistent with the tables). Since the census count was a constant, the standard error of the net undercount equals the standard error of the DSE, 266,000.

As seen in the Total line in Table 3, the alternative DSE was 133,285,000. Thus, with no FHUFU, the net undercount would have about doubled from 790,000 to 1,609,000, a difference of 819,000. Also, we see the effect of the alternative method was in the same direction for both occupied and vacant housing units. For example, for occupied housing units the alternative DSE was 590,000 greater than the official DSE, while for vacant housing units the alternative DSE was 229,000 greater than the official DSE. As will be seen later in Table 4, this pattern was consistent across various breakdowns of the estimates.

Table 3: Alternative DSEs by Occupied/Vacant Housing Units[3]

	Census (×1000)	Official DSE (×1000)	SE (×1000)	Alternative DSE (×1000)	SE (×1000)	Official DSE Minus Alternative DSE (×1000)	SE (×1000)
Occupied	116,699	116,735	160	117,325	168	-590	76
Vacant	14,977	15,732	174	15,961	177	-229	57
Total	131,676	132,467	266	133,285	267	-819	111

In Table 4, we see the alternative DSEs with the 2010 CCM imputation methodology and the alternative DSEs with the enhanced imputation, broken down by census region. The alternative DSE for Total with the enhanced imputation was 132,728,000, leading to an estimated net undercount of 1,052,000. Thus, the enhanced imputation reduced the difference between the alternative and official net coverage errors from 819,000 to 261,000.

The alternative DSE with enhanced imputation included the additional variable "E-sample Before Followup Match Code Group" in the logistic regression model. We did not investigate additional imputation models for two reasons. First, we had already included the variables that historically were found to be predictive, and there was not much prospect of continued improvement. And second, it went beyond the scope of the study to develop new imputation methodologies. Nevertheless, even with the enhanced imputation, the difference between the alternative DSE and the official CCM DSE, 261,000, was still large compared to the official CCM estimate of net coverage of 790,000.

[2] The quoted figures do not add up exactly because of rounding error.

[3] Figures in the table may not add up exactly because of rounding error.

Table 4: Alternative DSEs with Enhanced Imputation by Census Region[4]

Census Region	Census (×1000)	Official DSE (×1000)	Alternative DSE (×1000)	Enhanced Imputation for E-sample Status DSE (×1000)	Official DSE Minus Alternative DSE (×1000)	Official DSE Minus Enhanced DSE (×1000)
Northeast	23,647	23,531	23,688	23,590	-157	-59
Midwest	29,483	29,702	29,856	29,791	-154	-89
South	49,980	50,399	50,819	50,523	-420	-124
West	28,564	28,835	28,922	28,823	-87	11
Total	131,676	132,467	133,285	132,728	-819	-261
			Standard Errors (×1000)			
Northeast	N/A	112	113	113	44	36
Midwest	N/A	129	123	123	36	37
South	N/A	179	176	193	81	118
West	N/A	109	116	114	50	46
Total	N/A	266	267	265	111	136

6. Conclusions and Recommendations

With no FHUFU, the estimated net undercount of housing units more than doubled from 790,000 to 1,609,000. With the enhancements to the 2010 CCM imputation, the estimated net undercount increased by a smaller amount, from 790,000 to 1,052,000. However, even this smaller increase, 261,000[5], was large compared to the official CCM estimated net undercount of 790,000.

We conclude that to eliminate the FHUFU we would need to develop stronger imputation methods. We are not confident this can be achieved. Considering which variables were useful in past approaches to imputation for unresolved statuses, we may have approached the best imputation model we could have built. However, even with stronger imputation models, the large number of weighted cases requiring imputation may make an imputation risky. In our simulation, we had about 1.7 million weighted housing units that required imputation. In addition, in the 2020 PES we would have more unresolved P-sample cases to impute than the 2010 CCM did because of the data processing anomaly. With these many unresolved data, any imputation model might be risky. Thus, the authors' recommendation is to conduct the FHUFU for the 2020 PES.

A second recommendation is that the Census Bureau conduct a similar study after the 2020 PES. One reason is that the data-processing anomaly made our assessment of the impact of dropping the FHUFU less clear. Another reason is that we hope to conduct

[4] Figures in the table may not add up exactly because of rounding error.

[5] The quoted figures do not add up exactly because of rounding error.

research with the 2020 PES data to include new variables in our models to strengthen the imputation methodology and make relying on a large-scale imputation less risky.

References

Beaghen, M., and Wakim, A. (2017). "Assessing the Impact of the Final Housing Unit Followup on the 2010 Census Coverage Measurement Housing Unit Estimates." DSSD 2020 Post-Enumeration Survey Memorandum Series #2020 B-07.

Fay, R.E., Passell, J.S., and Robinson, J.G. (1980). "The Coverage of Population in the 1980 Census". (See pages 54-57). Bureau of the Census. https://books.google.com/books?id=1xI2XvV3T7IC&pg=PA112&lpg=PA112&dq=fay+passell+robinson+1980&source=bl&ots=1PSmPFd5Pd&sig=Sr4fP9wIhdKgXRvYadv8WmzFlM&hl=en&sa=X&ved=0ahUKEwj2kMHMyJrTAhWM7SYKHfMyB4kQ6AEIKjAE#v=onepage&q=fay%20passell%20robinson%201980&f=false

Imel, L., Mule, V., Seiss, M. (2013). "2010 Census Coverage Measurement Estimation Methods: Measures of Variation." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-J-03.

Konicki, S., Keller, A., Bray, R., Seiss, M., and Viehdorfer, C. (2013). "2010 Census Coverage Measurement Estimation Methods: Missing Data." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-J-03.

Mule, V. (2011). "Document Describing How We Handle Deletes in E-sample: May 3, 2011 Draft." Unpublished U.S. Census Bureau note.

Olson, D., and Viehdorfer, C. (2013). "2010 Census Coverage Measurement Estimation Methods: Net Coverage Estimation." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-J-04.

U.S. Census Bureau (2004). "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology." https://www.census.gov/prod/2004pubs/dssd03-dm.pdf

Whitford, D. (2008), "Overview of the 2010 Census Coverage Measurement Program" DSSD 2010 Census Coverage Measurement Memorandum Series #A-19, Washington, D.C.

Trang, T. (2017). "2020 Post-Enumeration Survey: High-Level Design Requirements for Sampling Operations." DSSD 2020 Post-Enumeration Survey Memorandum Series #2020-C-09.