



May 22, 2012

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-09

MEMORANDUM FOR David C. Whitford
Chief, Decennial Statistical Studies Division

From: Patrick J. Cantwell (*Signed*)
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Rachel Bray
Colt Viehdorfer
Decennial Statistical Studies Division

Subject: 2010 Census Coverage Measurement Estimation Report: Missing
Data for Components of Census Coverage

This report is one of twelve documents providing estimation results from the 2010 Census Coverage Measurement program. This report focuses on Missing Data for Components of Census Coverage.

For more information, contact Colt Viehdorfer on (301) 763-6796 or Rachel Bray on (301) 763-2631.

cc:
DSSD CCM Contacts List

Census Coverage Measurement Estimation Report

Missing Data for Components of Census Coverage

Prepared by
Rachel Bray, Statistical Support Branch
Colt Viehdorfer, Variance Estimation Branch

Decennial Statistical Studies Division

Table of Contents

Executive Summary	1
1. Introduction.....	2
2. Methods.....	2
2.1 Overview.....	2
2.2 Status Imputation for Persons	3
2.3 Status Imputation for Housing Units	4
3. Limitations	4
4. Discussion of Results.....	5
4.1 Person Missing Data Results for Components of Census Coverage	5
4.2 Housing Unit Missing Data Results for Components of Census Coverage.....	6
References.....	8

Executive Summary

This document summarizes the missing data results for the components of census coverage in the United States for both persons and housing units produced by the 2010 Census Coverage Measurement program. Previous census coverage programs were designed to provide estimates of net coverage error. The 2010 Census Coverage Measurement program produced estimates of components of census coverage to aid in improving future census data (Singh 2005). Component estimation is unique in that E-sample records are classified into one of several component outcomes instead of the binary outcome of correct or erroneous census enumeration. For example, erroneous person enumerations are further categorized by whether they are duplicates or erroneous for other reasons. Overall, the amount of missing data for the components of census coverage is relatively low and missing data procedures should have only a minor effect on the estimation.

For component enumeration status about 6.02% of the E-sample persons were unresolved at the national level. For resolved records at the national level, a determination was made as to whether or not the record was counted once and only once. The largest component outcome of correct enumerations consisted of persons counted correctly in the same block cluster. Unresolved cases were assigned an average probability of 0.91 of correct enumeration in the same block cluster. The largest component outcome of erroneous enumerations consisted of persons who were duplicated in the census. Unresolved cases with duplicate links, which account for 0.28% of all E-sample people, were assigned an average probability of 0.42 of being erroneous as a result of duplication.

About 0.08% of the E-sample housing units had unresolved component enumeration status. The most common component outcome for housing units was correct enumeration in the sample block cluster. Unresolved cases were assigned an average probability of 0.89 of correct enumeration in the sample block cluster.

1. Introduction

The purpose of this document is to provide an overview of the amount of missing data for 2010 Census Coverage Measurement (CCM) component estimation. As a result of the methodology, there are different amounts of missing data for each component enumeration status. This document provides information on the average probabilities imputed for missing cases of each component enumeration status as a part of the missing data methodology to support component estimation.

While CCM continued to produce net error estimates like earlier coverage measurement programs, for the first time, the CCM program provided components of census coverage. The 2010 CCM program expanded the interviewing, matching, and followup operations to gather the additional information to support the estimation of the census coverage components. The component results evaluate the 2010 Census and provide information that aids in planning the 2020 Census. Coverage components for the household population and housing units were estimated for major demographic groups, census operational areas, states, large counties, and large places, as appropriate. As this is the first effort to provide detailed component estimates on a production basis, this report does not provide any data for earlier censuses. However, research was conducted after the 2000 Accuracy and Coverage Evaluation (A.C.E.) pertaining to certain components, including duplicate person records. For more information on this research, see Fenstermaker and Haines (2002) or Feldpausch (2001).

2. Methods

2.1 Overview

To produce more accurate estimates of the components of census coverage, the strict definition of a correct enumeration used for implementing dual system estimation (DSE) and estimation of net error was loosened. The definition used for DSE and for estimating net error overstates the number of erroneous enumerations and omissions at the national level. For example, a person counted once and only once but outside of the correct block cluster search area is considered to be erroneously enumerated for net error estimation. For component estimation, the enumeration is correct at the national level if it is counted once and only once.

Another way in which the component missing data methodology deviates from the net error missing data methodology is in the handling of cases with insufficient information¹ for DSE processing. Net error treats census records with insufficient information for DSE processing as erroneous enumerations. To avoid introducing bias to the DSE through incorrect match status or incorrect enumeration status, no attempt is made to match these cases for net error. While some of these cases may be correct enumerations, they likely correspond to P-sample nonmatches. Therefore, for estimating net error, the errors balance and bias is not introduced. To better estimate the components of census coverage, an attempt is made to match and assign an enumeration status to the cases with insufficient information for DSE processing. Research shows that many of the cases with insufficient information can be matched and an enumeration

¹ Enumerations lacking a complete name and two characteristics were said to have insufficient information for matching and followup. They do not include whole-person census imputations.

status can be determined. More details are found in Livermore Auer (2005). Those cases with unresolved enumeration status have a probability of each status imputed using the method of cell means. Standard errors of the imputed means were computed using a Taylor series method, unlike other 2010 CCM estimates that used a delete-a-group jackknife method of variance estimation.

2.2 Status Imputation for Persons

For component missing data calculations, resolved E-sample persons are classified into eight enumeration outcomes. The outcomes, along with their correct or erroneous classification by the national definition² are listed below:

1. Correctly Enumerated in the Block Cluster Search Area (BCSA)³, which consists of the block cluster and the surrounding blocks
2. Correctly Enumerated in the same County and Place but Outside of the BCSA
3. Correctly Enumerated in the same County and a different Place
4. Correctly Enumerated in a different County and the same Place
5. Correctly Enumerated in a different County and Place but the same State
6. Correctly Enumerated in a different State
7. Erroneously Enumerated as a result of Duplication³
8. Erroneously Enumerated for reasons other than Duplication

For component outcomes for persons, we apply the following steps to assign enumeration status. For each of the eight component outcomes, records are assigned a probability of 1 if the status is “yes,” and a probability of 0 if the status is “no.” For any component outcome for which a person is unresolved, we impute a probability of that outcome using the method of cell means. The probability for some of the component outcomes is adjusted to account for duplication to records that are subsampled out of the E sample. For further discussion of this adjustment, see Mule (2008). Then, the probability for each outcome undergoes an adjustment so that the eight component outcomes for any record sum to one.

For any person record some statuses may be resolved while others are unresolved. For example, only records with a duplicate link to another census record were considered unresolved duplicates, and as such they are the only cases where a probability of being erroneously enumerated as a result of duplication was imputed. For the remainder of the unresolved records without a duplicate link, this probability is forced to be 0. There are some records where it is determined that the person should have been enumerated in a different location but we have incomplete information on the address at which they should have been counted. These records are considered resolved as a “no” for outcomes 1, 7, and 8 but unresolved for a combination of the remainder of the outcomes, dependent upon how much information we have on the address where they should have been counted.

²The eight outcomes are classified as either correct or erroneous depending on the geography which one considers. For example, persons who are correctly enumerated in a different state are considered correct by the national definition but are considered erroneous when considering enumerations at a state level.

³ The probability of this outcome requires an adjustment for duplication to persons in units in the sample block that are subsampled out of the E sample.

Due to the geography of where we find the person in sample, some of the component outcomes may not be applicable. For example, Washington, D.C. is considered a single county, place, and state. Outcomes 3, 4, and 5 are not applicable to person records in Washington, D.C. because an unresolved person in D.C. cannot be enumerated in a different county or place without also being enumerated in a different state. The probabilities of the non-applicable outcomes are forced to 0 and the person is considered resolved with respect to these outcomes.

While the component missing data methodology relies on a cell mean model to impute a probability of each unresolved component outcome, net error estimation uses logistic regression to impute cases with a missing enumeration status. The use of logistic regression to impute an enumeration status was also considered for component status imputation. Research on the 2000 A.C.E. data concluded that cell mean imputation and logistic regression yielded only minor differences (Viehdorfer and Moldoff, 2011). Cell mean imputation was chosen over logistic regression due to the ease with which the method is implemented and understood. Initial imputation cells for persons and housing units were based on 2000 A.C.E. Revision II data, and the cells were further refined using 2010 CCM data. The cells were chosen in such a way as to fulfill two conditions: a minimum number of resolved cases in each cell, and cells that discriminate well among the resolved cases according to their probabilities for each status outcome. For more information on the cell selection and various adjustments, see the forthcoming methods document.

2.3 Status Imputation for Housing Units

For component missing data calculations, E-sample housing units are classified into five enumeration outcomes, listed below:

1. Correctly Enumerated in the Block Cluster⁴
2. Correctly Enumerated in the Surrounding Ring of Blocks
3. Geocoding Error
4. Erroneously Enumerated as a Duplicate⁴
5. Erroneously Enumerated for reasons other than Duplication

Unlike a person record that can be resolved for some outcomes and unresolved for others, each housing unit is either resolved for all five outcomes or unresolved for all five outcomes. The probability for each outcome is assigned using the same methodology as is used for the person records, though the cells are defined differently.

3. Limitations

The results presented in this document can be affected by certain limitations to the component missing data methodology and the implementation of procedures. Potential limitations are listed below:

⁴ The probability of this outcome requires a similar adjustment to what was done for persons.

- These results assume that the data are missing at random within each imputation cell.
- In some instances a person record was counted in location different from the sample block cluster only once and the address where the person should have been counted was only able to be assigned to a general area. If this general area overlapped with the county, place, or state where the person record should have been counted then it was assumed the person was counted in the same county, place, or state.
- For housing unit estimation, housing units deleted by the census were selected in sample but omitted from these results and from component missing data processing.

4. Discussion of Results

Person results are given first, followed by housing unit results. Within each, the amount of missing data is discussed first, then the results of component status imputation are presented.

4.1 Person Missing Data Results for Components of Census Coverage

Table 1 presents the unweighted percentages of unresolved records for each component outcome under their correct or erroneous classification by the national definition. As previously discussed, a person record can be resolved for one outcome but unresolved for another. The unresolved records have a probability of each outcome for which they are unresolved imputed, and Table 1 also shows the average probabilities imputed for unresolved records.

Table 1. Amount of Missing Data and Probabilities Imputed for Component Status Outcomes for Person Records

Component Outcome	Average Probability Imputed	Standard Error	Unresolved (%)
Correctly Enumerated			
In the Block Cluster Search Area (BCSA)	0.9140	0.0011	6.02
In the same County and Place but outside of the BCSA	0.0115	0.0001	6.66
In the same County and a different Place	0.0045	<0.0001	6.10
In a different County and the same Place	0.0009	<0.0001	2.06
In a different County and Place but the same State	0.0037	<0.0001	6.56
In a different State	0.0067	<0.0001	6.58
Erroneously Enumerated			
Duplicates	0.4231	0.0031	0.28
Other Reasons*	0.0202	0.0003	6.02

*Includes Fictitious persons, those born after 4/1/10, and those that died before 4/1/10.

For most of the component outcomes, about 6% of the records are unresolved. One component outcome for which fewer records are unresolved is erroneously enumerated duplicates. There is a much smaller amount of missing data here because only records with a duplicate link to another census person were considered unresolved for the duplicate outcome. Another component outcome for which there is less missing data is the outcome of enumerated in a different county but the same place. Since many persons are in sample in locations with geography that conflicts with this outcome, only 2.06% of the records are unresolved and imputed a probability of correctly enumerated in a different county and the same place.

The average probability imputed for correctly enumerated in the block cluster search area is 0.9140. The remaining outcomes that are considered correct at the national level have low average probabilities imputed, the smallest of which is the average probability of being correctly enumerated in a different county and the same place. On average, persons that have an unresolved duplicate status are given a 0.4231 probability of being a duplicate. This probability may seem large, but it is only imputed for the 0.28% of unresolved persons with a duplicate link to a census record. Records that are unresolved for erroneous due to other reasons have an average probability of 0.0202 imputed.

4.2 Housing Unit Missing Data Results for Components of Census Coverage

Table 2 presents the unweighted percentages of housing unit records with a resolved and unresolved component status. Only one percentage is given for resolved and unresolved because a housing unit is either completely resolved or completely unresolved for all component status outcomes.

Table 2. Resolved Enumeration Status for Housing Units (Unweighted)

Component Outcome	Resolved	Unresolved
All Housing Unit Outcomes	99.88	0.12

Very few housing units have an unresolved enumeration status. Only 0.12% of unweighted housing unit records are unresolved. The few records that are unresolved have probabilities imputed for each component status outcome with the five probabilities adding to 1. The average probability imputed for each outcome is shown in Table 3.

Table 3. Probabilities Imputed for Component Status Outcomes for HU Records

Component Outcome	Average Probability Imputed	Standard Error
Correctly Enumerated		
In the Block Cluster	0.8853	0.0177
In the Surrounding Blocks	0.0253	0.0043
Erroneously Enumerated		
Geocoding Error	0.0023	0.0003
Duplicates	0.0247	0.0041
Other Reasons	0.0623	0.0110

The largest component outcome, correct in the block cluster, has an average imputed probability of 0.8853. A housing unit being correctly enumerated in the surrounding blocks is imputed at an average probability of 0.0253. A very low average probability of 0.0023 is imputed for being a geocoding error, while the average probability imputed for being a duplicate is 0.0247. Finally, the average probability imputed for being erroneous for another reason is 0.0623.

References

Feldpausch, R. (2001), "Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report Number 6: Census Person Duplication and the Corresponding A.C.E. Enumeration Status," DSSD Census 2000 Procedures and Operations Memorandum Series #T-16.

Fenstermaker, D., and Haines, D (2002), "A.C.E. Revision II Results: Estimated Correct Enumeration and Residence Probability for Duplicate Links," DSSD A.C.E. Revision II Memorandum Series #PP-52.

Livermore Auer, P. (2005), "Results of Feasibility Study to Match Census Enumerations Coded in A.C.E. as Insufficient Information for Matching and Followup," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-B-01.

Mule, T. (2008), "2010 Census Coverage Measurement Estimation Methodology," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-18.

Singh, R. (2005), "2010 Census Coverage Measurement - Updated Plans," DSSD 2010 Census Coverage Measurement Memorandum Series #A-06.

Viehdorfer, C. and Moldoff, M. (2011), "Documentation of Research on Imputing an Enumeration Status for Person Component Missing Data," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-E-43.