Estimates of Correct and Erroneous Enumeration with Duplicates in the 2010 U.S. Census

Rachel Bray[1]

U.S. Census Bureau, 4600 Silver Hill Rd, Washington, D.C. 20233

Abstract

Although historically there has been low net undercount in the census, using a single measure to describe coverage may mask additional errors. The 2010 Census Coverage Measurement (CCM) program produced estimates of the components of census coverage in addition to estimates of net error. The components of census coverage for persons in the 2010 Census are correct enumerations, erroneous enumerations, whole-person imputations, and omissions. The data collected in 2010 CCM support the breakdown of correct and erroneous enumerations into various subgroups. The estimates include correct enumerations at the national level, for all states, and large counties and places. A key subgroup of erroneous enumerations is duplicates. This paper explores the results of 2010 CCM estimation, with particular emphasis on duplicates. We report estimates by the distance separating the linked records, including duplicates at the same address, or in the same block, county, or state. We also estimate the count of duplicates according to their characteristics, such as whole households of duplicates, and whether the duplication is a result of mailback or non-mailback census enumerations.

Key words: coverage measurement, dual system estimation, official statistics

1. Introduction

For the first time in 2010, the census' post-enumeration survey, Census Coverage Measurement (CCM) produced extensive estimates of the components of census coverage, in addition to estimates of net error. These estimates are used not only to describe the coverage of the 2010 Census, but also in planning for 2020. One component that is of particular interest is persons who were erroneously enumerated due to duplication, or duplicates. A duplicate is someone who was erroneously included in the census because he or she was counted more than once. Through matching and followup, CCM identified duplicates in the census and linked the enumerations that represent the same person together. For each linked pair of records, one is attempted to be resolved as correct, i.e., where the person should have been counted, and the other is considered to be erroneously included in the census. The erroneous duplicates account for 8.5 million of the 10.0 million total erroneous enumerations in the census household population count. For reference, the distribution of the correct and erroneous enumerations for persons at the national level is given in Table 1. Table 1 displays a subset of the components of census coverage, and the distribution of the remainder of the components along with more information on the component estimates can be found in Mule (2012).

---

[1] Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.

Table 1: Selected Components of Census Coverage for the United States Household Population (in Thousands)

Component of Census Coverage	Estimate	Standard Error	Percent	Standard Error
Correct enumerations[1]	284,668	199	94.7	0.07
Enumerated in the same block cluster[2]	280,852	220	93.4	0.07
Enumerated in the same county, though in a different block cluster	2,039	55	0.7	0.02
Enumerated in the same state, though in a different county	830	34	0.3	0.01
Enumerated in a different state	948	31	0.3	0.01
Erroneous enumerations	10,042	199	3.3	0.07
Due to duplication	8,521	194	2.8	0.06
For other reasons	1,520	45	0.5	0.01
Whole-Person Census Imputations	5,993	0	2.0	0

1. A correct enumeration is a person who was counted correctly in the census, anywhere in the nation.
2. A block cluster is a group of one or more contiguous census blocks, which contains an average of 30 housing units. Block clusters are the primary sampling units for CCM.

2. Methods

The estimates of duplication were produced using a subset of the CCM Enumeration (E) Sample. The E Sample is a sample of approximately 6,000 block clusters and the persons and housing units within them, used to generate direct estimates of correct and erroneous census enumerations. Persons enumerated in the census in housing units selected in the E Sample went through additional interviewing, matching, and followup to determine whether or not they were counted correctly in the census. This results in E-Sample persons being classified into one of the six component outcomes given in Table 1, under the larger categories of correct and erroneous. While duplication is typically referred to in the context of a single pair of records, these methods generalize to persons duplicated more than once in the census. Matching assigned one of the enumerations within a linked pair as the correct enumeration, or primary; while the other record in the linked pair is considered the erroneous duplicate. Each pair of linked records consists of a primary and one of its erroneous duplicates.

For the estimation of duplicates in CCM, persons that were duplicated to people living in a group quarters (GQ), such as a college dorm or nursing home, were considered erroneous in the housing unit in which they were enumerated. For the household population that was duplicated to persons in housing units, CCM estimation relied on computer and clerical matching to determine which half of the duplicate pair is erroneous. One exception to this occurred as a result of the subsampling that occurred within some block clusters. Persons for which the housing unit containing the other half of their pair was subsampled out of the E Sample had an adjustment performed so that the in-sample record was considered half erroneous and half correct. There are also records with duplicate links for which we were unable to determine in matching whether or not they were erroneous. These in-sample records were imputed a probability of being erroneous duplicates, and a proportion of their weight contributed to the duplicate estimate.

The 13,191 E-sample records that contributed to the duplication estimate in Table 1 were extracted from the E-sample data, along with a pointer to the other half of the pair of linked records. The records contributing to this estimate fall into one of three classifications: persons that are resolved to be erroneous duplicates, unresolved records with duplicate links, and primaries for which one of their duplicates was subsampled out of the E Sample. In some cases, a primary had more than one duplicate, and as a result was associated with more than one pair of linked records. In these cases, for the tabulation of results presented here, an additional record was made for each duplicate pair, and the weight contributing to the duplicate estimate was split evenly among all the pairs associated with the primary. The resulting analysis file contains one record for each of the 13,336 pairs of linked records, for a weighted total of 8.5 million duplicates.

After creating a file with the duplicate pairs and appropriate weights, I pulled variables describing the geography as well as operational and demographic characteristics of each half of the pair from various census files. These variables include those for the form type of the selected return, the age/sex group, an indicator of whether the person was in a group quarters, and the type of group quarters for persons duplicated to GQs. Once I had the characteristics for each half of the pair, I used this information to aggregate the estimates by various characteristics associated with the pair, such as the geographic distance between the linked records.

For persons that were duplicated to housing units in the same block cluster, it is possible to pull their housing unit outcomes from CCM by matching each half of the duplicate pair of persons to their corresponding housing unit records. The information taken from the housing unit data may provide context for how the persons came to be enumerated more than once in the census. For example, using this information enables one to see if the person duplication was accompanied by housing unit duplication. The housing unit outcomes of the duplicate pairs are broken into four summary categories, as follows:

1. Both halves of the duplicate pair of person records are in correctly enumerated housing units, no duplication of the housing units associated with the pair of records[2]

2. At least one half of the duplicate pair is in a housing unit that was duplicated

3. No housing unit duplication associated with the duplicate pair of persons, but one half of the pair is in a housing unit that is erroneously enumerated for other reasons, including non-existent housing units, those that are demolished or burned down, and commercial properties that are erroneously listed as housing units.

4. All other situations. This includes when a person was found in a correctly enumerated housing unit, and the housing unit to which they were duplicated was subsampled out of the E Sample. With housing units subsampled out of the E Sample, it was not necessary to determine whether or not they were correctly enumerated.

To put the duplicate-pair specific housing results into context, the overall distribution of housing unit components at the national level is given in Table 2. Of all housing units

---

[2] Correctly enumerated housing units consist of housing units that existed and were counted on the census. This housing unit outcome does not imply that the housing units were correctly classified as occupied.

counted in the census, 97.3 percent are correctly enumerated. We estimate that 0.9 percent of all housing units are duplicates, while 1.8 percent of all housing units are erroneously enumerated for other reasons. For more information on the components of census coverage for housing units, see Mule and Konicki (2012).

Table 2: Components of Census Coverage for Housing Units

Component of Census Coverage	Percent	Standard Error
Correct enumerations	97.3	0.1
Enumerated in the same block cluster	96.1	0.2
Enumerated in the surrounding blocks	1.2	0.1
Geocoded outside the search area	0.1	0.03
Erroneous enumerations	2.7	0.1
Due to duplication	0.9	0.07
For other reasons	1.8	0.09

3. Limitations

The limitations of these results are similar to the limitations of the CCM program. As with CCM, these results exclude persons in remote Alaska. These estimates also exclude persons in Puerto Rico. Although estimates presented may involve group quarters persons, they pertain to the coverage of the housing unit population in the census. Duplication to a person in a group quarters means that the person in a housing unit was erroneously enumerated for this reason, and that is all. One cannot make inferences about the coverage of persons in group quarters from these estimates.

4. Results

Estimates of erroneous duplicates in the census by geographic distance between linked records and whether the duplication was to a person in a housing unit or GQ are given in Table 3. Only 515,000 of the duplicates within the household population were to a person in a GQ, while about 8 million duplicates were to people in housing units. Duplicates to persons in housing units increased at every level of geography from 2000 levels. Of the 8.0 million duplicates to persons in housing units, 3.8 million were located within the same block. This is a substantial increase from the 2000 estimate of 2.9 million duplicates in the same block. Later tables will provide additional breakdowns of the 2010 estimate to attempt to explain the increase in within-block duplication. Meanwhile, the duplicates to persons in group quarters decreased overall, and either decreased or had no significant change from 2000 at every geographic level except for those in the same state, different county. The reduction in duplication to persons in group quarters is concentrated within the same county areas and smaller geographic distances.

Table 3: Total Duplicates in the Census, 2000 and 2010 by Geographic Distance and if the Duplication is to a Person in a Housing Unit or Group Quarters (in Thousands)

Geographic Distance	Duplicates to Persons in Group Quarters (GQs)		Duplicates to persons in Housing Units	
	2000	2010	2000	2010
Within Block	73	41	2,907	3,801
	(15)	(16)	(84)	(160)
Within Tract,	42	16	625	1,018
Different Block	(21)	(4)	(31)	(106)
Within County,	219	138	1,210	1,649
Different Tract	(38)	(12)	(35)	(67)
Within State,	190	237	664	852
Different County	(7)	(18)	(24)	(36)
Different State	92	83	549	686
	(6)	(9)	(31)	(37)
	616	515	5,955	8,006
Total	(43)	(29)	(113)	(195)

4.1 Estimates of Duplication of Persons in HUs to Persons in GQs

Although the 2010 CCM is not designed to evaluate coverage of persons in group quarters, and is limited to coverage estimates of the household population, we are able to estimate how much of the household population was erroneously included because they were also enumerated in a GQ. Tables 4 and 5 show breakdowns of duplication of the household population to persons in GQs. Table 4 shows the duplicates by the type of group quarters and their age/sex groups. Table 5 shows the breakdown of duplicates by the geography between the primary and the duplicate census enumerations, and the type of group quarters. The blank cells in Tables 4 and 5 represent domains in which we had no sample.

Table 4 shows that about 260,000 persons were erroneously duplicated to people in college dorms. Many of these are in the 18 to 29 age group. A small number of the duplicates to college dorms are in the 10 to 17 age group. The 10 to 17 age group contains about 14,000 of the approximately 19,000 persons duplicated to juvenile institutions. The duplicates to correctional institutions are concentrated in the 18 to 49 year old males, with some additional duplicates in the 50+ males and 18 to 49 year old female age/sex groups. Many of the duplicates to nursing homes are in the 50+ male and female age/sex groups. The 18 to 29 males and 30 to 49 year old males make substantial contributions to the estimate of duplicates to military group quarters.

Table 4: Estimate of Erroneous Enumerations Due to Duplication from Census HUs to Census GQs by Type of Group Quarters and Age/Sex Categories (in Thousands)

Age/ Sex Category	Type of Group Quarters						Total
	College Dorm	Correctional Institution	Juvenile Institution	Nursing Home	Military	Other[1]	
0 to 4	2 (1)	1 (1)	<1 (<1)	1 (1)		<1 (<1)	3 (1)
5 to 9	2 (1)		2 (2)				4 (2)
10 to 17	10 (3)		14 (4)	2 (1)		2 (1)	27 (5)
18 to 29 Males	103 (10)	37 (6)	2 (2)		9 (3)	17 (7)	167 (13)
18 to 29 Females	141 (12)	5 (2)			1 (1)	6 (3)	154 (13)
30 to 49 Males		23 (5)			6 (3)	8 (3)	37 (6)
30 to 49 Females	1 (1)	3 (1)	<1 (<1)	2 (1)		3 (2)	10 (3)
50+ Males	1 (1)	8 (3)	<1 (<1)	38 (7)		6 (3)	53 (9)
50+ Females		1 (1)		50 (8)		8 (4)	59 (8)
Total	260 (16)	77 (8)	19 (4)	92 (12)	16 (4)	51 (14)	515 (29)

1. The Other category includes Transitional Shelters, Group Homes intended for Adults, workers group living quarters, etc.

Table 5 shows estimates of duplication of the housing unit population to group quarters by type of group quarters and the geographic distance separating the halves of the duplicate pair. For college dorms, duplication occurred mainly outside of the tract and farther away, with about half of the duplicates to college dorms being within the same state but in a different county. For persons duplicated to nursing homes, much of the duplication occurred within the county but outside of the block, although some also occurred within the state but in a different county. Duplication to juvenile institutions is concentrated within the state but outside of the tract, as is duplication to correctional institutions.

The 'Other types of group quarters' domain accounts for much of the within-block duplication to group quarters. These 24,000 enumerations consist of persons duplicated to group homes intended for adults and workers group living quarters. Since one of the enumerations of the pair is located in a living quarters listed as a GQ in the census, we do not make any assessment as to the correctness of these living quarters including their classification as a GQ.

Table 5: Estimate of Erroneous Enumerations Due to Duplication from Census HUs to Census GQs by Type of Group Quarters and Geographic Distance (in Thousands)

Geographic Distance	College Dorm	Correctional Institution	Juvenile Institution	Nursing Home	Military	Other	Total
Within Block	7 (5)			10 (9)		24 (13)	41 (16)
Within Tract, Different Block	3 (2)	1 (<1)		12 (4)	<1 (<1)	1 (1)	16 (4)
Within County, Different Tract	38 (6)	31 (5)	5 (2)	49 (7)	4 (2)	11 (3)	138 (12)
Within State, Different County	147 (13)	42 (6)	13 (3)	19 (4)	6 (3)	10 (4)	237 (18)
Different State	66 (8)	3 (1)	1 (1)	2 (1)	6 (2)	5 (2)	83 (9)
Total	260 (16)	77 (8)	19 (4)	92 (12)	16 (4)	51 (14)	515 (29)

4.2 Estimates of Duplication to Persons in HUs

Most of the duplication of the household population in the 2010 Census occurred between persons in housing units. Table 6 shows the distribution of these 8.0 million duplicates by the geographic distance between the halves of the duplicate pair, the household size, and the number of duplicate links between the two households.

Duplicates for which either pair of enumerations was in a household of size one are given in the first two columns. These are presented by whether both enumerations were households of only one person, or if the duplicate or primary were located in a household of two or more persons. The remaining columns describe duplicates that were not associated with a single-person household, with both the household sizes being two or more and the number of links ranging from only one, to a partial household of duplicates, to a whole household of duplicates. The designation of whole household of duplicates is reserved here for a direct copy of the roster of an entire household to another, where all persons in one housing unit are erroneous duplicates and all persons in the other housing unit are the correctly enumerated half of the pair.

Of the 3.8 million within-block duplicates, about 60% are in housing units with more than one person and for which the whole household was duplicated. Much of the within-tract, different-block duplicates in households with two or more links were also in whole households of duplicates. For duplication beyond the tract level, comparatively fewer duplicates between households with 2+ to 2+ links occur in whole households of duplicates. In general, as the geographic distance between the linked enumerations increases, the overall amount of duplication decreases. For one-person households of duplicates, the within-tract duplication consists of more single-person households that are duplicated rather than single persons who are duplicated to larger households.

Table 6: Estimate of Erroneous Enumerations Due to Duplication between Persons in Housing Units by Household Size, Number of Links, and Geographic Distance (in Thousands)

Geographic Distance	Household Size and Number of Links					Total
	1 Person to 1 Person	1 Person to 2+	2+ to 2+			
			Whole Household	Partial Household	Only 1 Link	
Within Block	337	108	2,278	777	302	3,801
	(23)	(11)	(129)	(54)	(20)	(160)
Within Tract, Different Block	75	20	628	193	102	1,018
	(15)	(4)	(89)	(24)	(11)	(106)
Within County, Different Tract	38	62	348	591	610	1,649
	(6)	(7)	(48)	(41)	(26)	(67)
Within State, Different County	33	48	195	227	349	852
	(6)	(7)	(23)	(23)	(19)	(36)
Different State	43	50	256	121	216	686
	(7)	(9)	(27)	(16)	(15)	(37)
Total	526	288	3,705	1,909	1,579	8,006
	(29)	(18)	(171)	(77)	(44)	(195)

About one fourth of all duplicates to persons in housing units fall in to a single cell in Table 6, the whole households duplicated within the collection block. In order to understand this subgroup, I looked at the housing unit outcomes of the duplicate pairs, located in Table 7. The numbers of the housing unit outcomes correspond with the explanations of these housing unit outcomes given in the methods section. The distribution of housing unit outcomes in Table 2 provides context for the overall amount of housing unit erroneous enumerations.

Table 7: Distribution of Housing Unit Outcomes for Whole Households of Duplicates in the Sample Block (in Thousands)

Housing Unit Outcomes	Duplicates (Standard Error)	Percent of Whole Household Duplicates in the block
1. Both halves of the duplicate pair of person records are in correctly enumerated housing units, no duplication of the housing units associated with the pair of records[1]	797 (50)	35.0%
2. The duplicate, the primary, or both are in a HU that was Duplicated	783 (87)	34.4%
3. No housing unit duplication associated with the duplicate pair of person enumerations, but one half of the pair is in a housing unit that is erroneously enumerated for other reasons	544 (46)	23.9%
4. Other Situation, including unable to determine whether one half of the linked pair of persons was in an erroneously enumerated housing unit	154 (20)	6.8%

1. Correctly enumerated housing units consist of housing units that existed and were counted on the census. This housing unit outcome does not imply that the housing units were correctly classified as occupied.

The housing unit outcomes of duplicates in households with 2+ to 2+ links where the whole household was duplicated within the sample block show that many of the duplicate pairs have housing unit outcomes that put them in category 2 or 3. That is, about half of the person duplication in whole households in the block involved either the primary or the duplicate being located in an erroneous housing unit. These rates suggest that although at 2.7 percent, the overall amount of erroneous housing units is low, the impact of these errors on person duplication through duplicating whole households within the block is considerable. About 35 percent of these duplicates (whole households, 2+ to 2+ links, in the sample block) are not associated with any erroneously enumerated housing unit. Table 7 does not give any information as to whether the housing units were classified properly as occupied in the Census.

Table 8 presents the estimates of duplication by census geography and the selected form types of the duplicates. This required some collapsing of the selected form types into groups that represent the operations with which they are associated. The first form type group is Mail returns, the form types which were either mailed or given to a housing unit, and then mailed back. The Mail form type group includes initial Mailout - Mailback forms, bilingual forms, replacement mailing forms, and Update/Leave forms. The second form types group is for Non-Response Followup (NRFU) form types, and consists of any form type associated with the initial, supplemental, or NRFU reinterview (used in quality control for NRFU). The third group of form types consists of those associated with Coverage Followup (CFU), an operation that conducted telephone interviews in order to collect additional or clarifying information from households identified as having

suspected coverage problems. If a pair of duplicates does not have both of its form types in these three major groups, then the pair is placed in the other category. This category consists of linked pairs with at least one of the enumerations resulting from form types associated with Telephone Questionnaire Assistance, Fulfillment, or Be Counted forms, among others. As with the housing unit outcomes, the selected form type groupings give an idea of the operational context in which duplication of the persons occurred.

Table 8: Estimate of Erroneous Enumerations Due to Duplication from Census HUs to Census HUs by Type of Return and Geographic Distance (in Thousands)

| Geographic Distance | Form Type Pair | | | | | | | Total |
	Mail-Mail	Mail-NRFU	NRFU-NRFU	Mail-CFU	NRFU-CFU	CFU-CFU	Other	
Within Block	304	2,344	604	62	282	43	163	3,801
	(26)	(114)	(55)	(13)	(35)	(9)	(22)	(160)
Within Tract, Different Block	75	660	208	12	46	3	14	1,018
	(14)	(83)	(33)	(4)	(11)	(2)	(6)	(106)
Within County, Different Tract	368	786	188	120	132	22	33	1,649
	(22)	(42)	(20)	(14)	(21)	(10)	(9)	(67)
Within State, Different County	333	319	77	61	45	11	7	852
	(23)	(20)	(11)	(10)	(8)	(4)	(2)	(36)
Different State	273	278	55	42	22	6	9	686
	(23)	(22)	(11)	(7)	(5)	(3)	(4)	(37)
Total	1,352	4,387	1,132	297	527	86	226	8,006
	(53)	(165)	(74)	(74)	(44)	(15)	(25)	(195)

A little over half of the duplication within the tract and lower levels of geography is associated with duplicates resulting from one Mail and one NRFU return. At the within-county, different-tract level, slightly less than half of the duplication is the result of one Mail and one NRFU return. For duplication outside of the county, two Mail returns along with one Mail one NRFU account for much of the duplication, while the estimates of duplication within the county show more of the duplication resulting from one Mail and one NRFU return.

Within the duplicates resulting from Mail-CFU form type pairs, much of the duplication occurs at the within-county, different-tract level. Within the duplicates resulting from NRFU-CFU and CFU-CFU form type pairs, much of the duplication occurs at the within block level, with the some of the duplication also occurring at the within county, different tract level.

Over half of the total duplicates between persons in housing units are coming from one Mail and one NRFU return. About a quarter of the total duplicates between persons in housing units have this form type pair and are in the same block. In order to provide some context as to how this duplication occurs, Table 9 shows the housing unit outcomes for the within block Mail-NRFU return type pair duplicates.

Table 9: Distribution of Housing Unit Outcomes for Duplicates in the Sample Block with Mail-NRFU Selected Form Type Pairs (in Thousands)

Housing Unit Outcomes	Duplicates (Standard Error)	Percent of Whole Household Duplicates in the block
1. Both halves of the duplicate pair of person records are in correctly enumerated housing units, no duplication of the housing units associated with the pair of records[1]	991 (56)	42.3%
2. The duplicate, the primary, or both are in a HU that was Duplicated	656 (74)	28.0%
3. No housing unit duplication associated with the duplicate pair of person enumerations, but one half of the pair is in a housing unit that is erroneously enumerated for other reasons	547 (45)	23.3%
4. Other Situation, including unable to determine whether one half of the linked pair was in an erroneously enumerated housing unit	150 (18)	6.4%

1. Correctly enumerated housing units consist of housing units that existed and were counted on the census. This housing unit outcome does not imply that the housing units were correctly classified as occupied.

As in the previous table of housing unit outcomes, many of the housing units fall into outcome 1. However, if we combine outcomes 2 and 3, and look at all duplicate pairs that were associated with an erroneous housing unit, then much of the within-block duplication resulting from a NRFU-Mail form type pair is associated with some erroneous housing unit. This suggests that the person duplication between Mail and NRFU operations is at least partially accounted for by housing unit errors that are being captured in these same operations.

5. Conclusions

Of the total 8.5 million duplicates in the 2010 Census, 8.0 million were to persons in housing units. The remaining 515,000 were to persons in group quarters, with about half (260,000) of the duplicates to GQs being duplicated to persons in college dorms. The duplicates to college dorms are mostly outside of the county, with 147,000 being in the same state and different county, and 66,000 being in a different state. At the within-block level, duplication to GQs is small for every type of group quarters except for the "Other" category consisting of Group Homes intended for adults and workers group living quarters. This suggests that the coordinated frame development between housing units and GQs in 2010 (they were developed separately in 2000) was effective in eliminating person duplication within the block in nearly all types of Group Quarters.

About half of the 8.0 million duplicates to persons in housing units were within the tract and lower levels of geography. These duplicates represent an absolute increase from the

number of duplicates found in similar geography in 2000. Different household compositions of duplicates, as seen by the number of links and the number of persons duplicates, are more likely to occur at different levels of geography. For the 3.8 million within-block duplicates, 2.3 million were whole households of duplicates of two or more persons. These whole households of duplicates can partially be accounted for by the erroneous inclusion of housing units associated with them.

References

Mule (2012). "Summary of Estimates of Coverage for Persons in the United States." 2010 CCM Summary Reports, G-01. http://2010.census.gov/news/press-kits/ccm/ccm.html

Mule and Konicki (2012). "Summary of Estimates of Coverage for Persons in the United States." 2010 CCM Summary Reports, G-02. http://2010.census.gov/news/press-kits/ccm/ccm.html

National Research Council (2009). "Coverage Measurement in the 2010 Census." Panel on Correlation Bias and Coverage Measurement in the 2010 Decennial Census, Robert M. Bell and Michael L. Cohen (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.