Once and Only Once: Searching Near and Far For Person Duplications in the 2010 Census

Aaron Cantu and Susanne Johnson
U.S. Census Bureau[1]
4600 Silver Hill Road, Washington, DC 20233

Abstract

Census respondents often have their own ideas regarding where people should be counted, which can lead to people being enumerated in more than one place (duplicated). Problems in establishing a person's Census Day residence led to the 2000 coverage measurement program underestimating erroneous enumerations (many of which were duplicates). This paper will discuss methodology used by the 2010 Census Coverage Measurement (CCM) Program to improve measurement of Census Day residence and improve techniques to detect and resolve duplicate enumerations. CCM collected address information for other places people could have been counted on Census Day. We conducted computer and clerical searches for census duplicates and matches between the CCM and census near these additional addresses (on top of searching near the sample address). We also conducted a nationwide computer search that was less conservative than previous computer searches for census duplicates because CCM had the opportunity to clerically review these computer links and conduct field followup, when necessary, to resolve whether the links were truly duplicates and establish where the person should have been counted.

Key Words: duplication, 2010 Census, coverage measurement, coverage error, erroneous enumerations, record linkage

1 Background

Prior to 2000, programs evaluating census coverage had estimated net undercounts in the decennial census population. However, the original demographic analysis of the 2000 Census found a net overcount for the first time. This suggested duplication of people within the census. It was suspected that the coverage evaluation of the 2000 Census (then called the Accuracy and Coverage Evaluation [A.C.E.]) failed to measure a significant number of erroneous enumerations, many of which were duplicates. This suspicion was later confirmed with the Person Duplication Studies, which were the first studies to utilize a computerized matching operation *across the entire nation.* Since the 2000 A.C.E. only searched within pre-determined geographic areas around the evaluation sample block clusters[2], the Person Duplication Studies provided a significant insight into the measurement of duplicate enumerations in the census. That is, they found an estimated 6.6 million duplicates in the 2000 Census (Mule, 2012). However, using only the data originally collected by the A.C.E., there was still no way of knowing which one

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

[2] Census coverage evaluations are sample-based operations. The nation is divided into thousands of geographic "blocks", which are grouped into block clusters of one or more contiguous blocks. A sample of block clusters and a sample of housing units (also referred to as sample addresses) within those block clusters are then selected to be included in the census coverage evaluation program.

of the multiple records for a duplicated person represented where the person really should have been counted in the census.

The coverage evaluation for the 2010 Census, called the 2010 Census Coverage Measurement (CCM), was designed with the issue of census duplication in mind. It set out to not only find these census duplicates, but also to determine where the person should have really been counted. Since the CCM was an evaluation, its results did not affect the 2010 Census.

The 2010 CCM was a large, complex survey conducted independently of the 2010 Census, which was designed to produce coverage estimates for housing units and persons within housing units. This paper focuses on the methodology and results of the CCM Person Interview (PI), Person Matching, and Person Followup operations, and the effect these had toward identifying and resolving duplicate persons in the census. All results presented here are given from an operational standpoint and do not reflect final CCM estimates of person coverage. These results reflect unweighted data for operations conducted in the CCM sample areas within the United States (excluding Puerto Rico), thus no statistical significance testing was conducted and no inferences to the general population are intended.

2 What is a Census Duplicate?

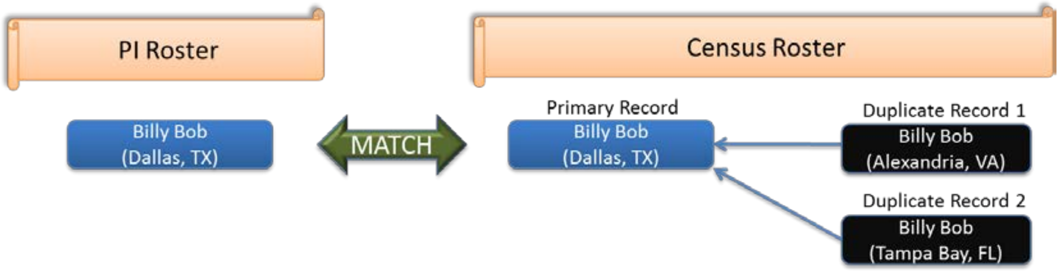
When people are represented in the census more than once, they are said to be duplicated. CCM defined the record that represents where the person should really be counted in the census as the *primary* record, and any other records that represent the same person were called *duplicate* records. The rules CCM followed when designating primary and duplicate census records are listed below.

1. If a person was enumerated in the census more than once, there could only be one primary record and any other record that refers to the same person must be a duplicate.
2. Only the primary census record can link to a person record from the CCM PI. That is, if it is decided that a PI record and a census record refer to the same person, the PI record must be linked to the primary census record (not to any of the duplicate census records for the same person).
3. The primary and duplicates were designated to reflect the true residence for the person. The record representing where the person should have been counted according to the 2010 Census Residence Rule (http://www.census.gov/population/www/cen2010/resid_rules/resid_rules.html) was designated as the primary. The person should not have been counted anywhere else in the census and thus any other census records referring to the same person were coded duplicates (and were considered erroneous enumerations).

In order to find these duplicates, the CCM PI collected a roster of people who lived at the sample address at the time of the interview and also whether they lived at that address on Census Day. This operation was done independently of census operations. The rosters were then compared and records representing the same person were linked together.

Figure 1 below shows a made-up example of a person named Billy Bob collected only once on the PI roster but collected at three addresses for the census. It illustrates the proper way to designate the primary and duplicate census record for a person that was enumerated multiple times in the census. The primary record for Billy Bob is at an address located in Dallas, TX. There are two duplicate records in the census for Billy Bob (one located in Alexandria, VA and one located in Tampa Bay, FL).

FIGURE 1



3 In Pursuit of Duplicates

So, how do we identify duplicates in the census? Moreover, how do we determine which record represents where the person should really be counted? Below we summarize steps within the CCM project that accomplished this task.

1. Conducted PI at sample addresses.
2. Conducted computer matching: Searched for duplicates and matches between the PI rosters and the census enumerations.
3. Conducted Before Followup Clerical Matching (BFU): Reviewed computer matching results, search for additional matches and duplicates, and made updates as necessary.
4. Conducted Person Followup (PFU): Sent cases needing additional information to the field for followup.
5. Conducted After Followup Clerical Matching (AFU): Reviewed information from followup to determine match, duplicate, and enumeration status.

3.1 Person Interview

The process of finding duplicates began with the CCM PI. These interviews were conducted at sample addresses located within CCM sample block clusters. For this discussion, we are particularly interested in the questions listed below.

- Who lived at the sample address at the time of the interview (PI Day)?
- Who lived at the sample address on April 1, 2010 (Census Day)?
- Were there any other addresses where people lived or stayed around Census Day?
- If people had multiple addresses, when did they move or how often did they cycle between different addresses?

While these questions formed the basis for creating the PI roster, they also provided information for resolving census duplicates. Asking respondents about other addresses where the people may have lived or stayed around Census Day gave us an indication of additional locations where the people could have been enumerated in the census (in case they were counted in the census more than once). Further, the information about when they were at different addresses was used to determine each person’s Census Day residence status (i.e., where the person should have been counted according to the Census Day Residence Rule).

Now, we will continue to use our made-up example about Billy Bob to illustrate how we locate and resolve Billy Bob's duplicate records in the census. In the example, the CCM sample block cluster lies in Dallas, TX. During PI, an interviewer went to each sample address to conduct interviews. From the interview at Billy Bob's house, we learn the following information:

- Billy Bob was the only person living at the sample address on PI Day.
- Billy Bob lived here on Census Day. In fact, he had lived there alone the past 10 years.
- Billy Bob visited his son in Alexandria, VA for a couple of weeks in April 2010.

In this example, the PI roster for the sample address contains only one person, Billy Bob, and based on the interview, we know the following things about Billy Bob:

1. He lived at the sample address in Dallas, TX on Census Day and we *should* find him enumerated in the census at the Dallas address (his sample address).
2. He also stayed at another place around Census Day, so we *might* also find him enumerated in the census at an address in Alexandria, VA (his alternate address).

3.2 Computer Matching

After the PI was completed and the data was processed, the computer linked PI records to census records throughout the country and assigned match scores indicating the likelihood that the linked records referred to the same person. Using the match scores, match codes were assigned to classify people as Matches, Possible Matches[3], and Nonmatches, as well as Duplicates and Possible Duplicates. Links were identified in the sample, inmover[4], alternate or nationwide search areas, as defined below.

Match Search Areas:

a. When the census record was located in the block cluster containing the PI sample address, or the ring of blocks surrounding that block cluster (i.e., the "surrounding blocks"), then the link was in the sample search area.

b. Otherwise, if the census record was located in the block cluster or surrounding blocks of an inmover address reported for the linked PI record, then the link was in the inmover search area.

c. Otherwise, if the census record was located in the block cluster or surrounding blocks of an alternate address reported for the linked PI record, or reported for some other PI record in the household as an inmover or alternate address, then the link was in the alternate search area.

[3] A Match was a "strong" link in the sense that we were confident that the linked census and PI records represented the same person. A Possible Match was a "weak" link between two records that we believed may have represented the same person, but we were not certain. This definition is analogous for Duplicates and Possible Duplicates, except for the fact that the two linked records in question were both census records.

[4] An inmover address was a specific type of respondent-provided address. If a respondent moved into the sample address after Census Day but before PI day, that person was an *inmover* because he moved *into* the sample address after Census Day. The address he came from, where he lived on Census Day, was the *inmover address*.

d. If the link was beyond the sample search area and it was not associated with an inmover nor an alternate address, then the link was in the nationwide search area.

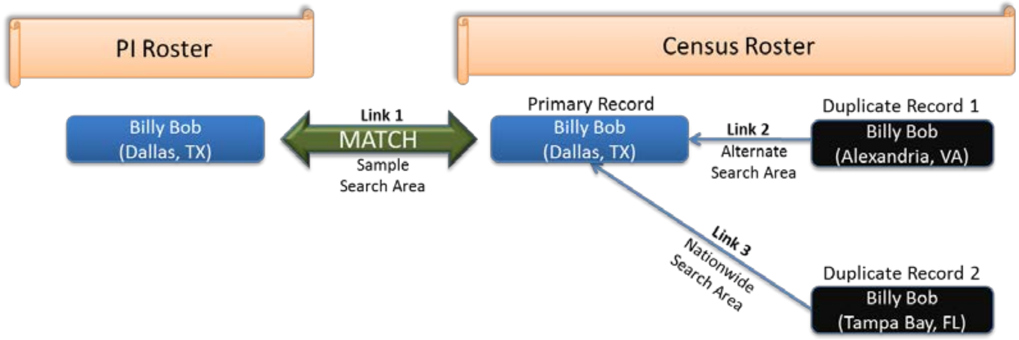
In addition, the computer searched for census duplicates throughout the country, so similar search areas were also identified for census duplicate links, as defined below.

Duplicate Search Areas:

a. If both census records were located in the sample block cluster or its surrounding blocks, then the link was in the sample search area.
b. Inmover and alternate addresses weren’t collected for census records during the original PI, however if a census duplicate link was located in the block cluster (or surrounding blocks) of an inmover or alternate address reported for a PI record that is linked to someone in the census record's household, then we say the link was in an inmover or alternate search area.
c. If the link was beyond the sample search area and it was not associated with an inmover or an alternate address, then the link was in the nationwide search area

Let’s look at what the computer pulled together for our Billy Bob example in Figure 2.

FIGURE 2 – Computer Matching Results



There were several records linked together, both within the sample block cluster and nationwide. There are three links to consider:

- Link 1 – This is a link within the sample block cluster (i.e., within the sample search area). The computer found a census Billy Bob at the address in Dallas, TX to link to the PI Billy Bob.
- Link 2 – This is a link beyond the sample search area (i.e., beyond the sample block cluster and its surrounding blocks). During its nationwide search for census duplicates, the computer found another Billy Bob in the census in Alexandria, VA that may refer to the same person as the record in Dallas. This record was within the block cluster of the PI Billy Bob’s alternate address, thus this link is within an alternate search area.
- Link 3 – This is another link beyond the sample search area. The computer found a third Billy Bob in the census in Tampa Bay, FL that may refer to the same person as the record in Dallas. Since this record is beyond the sample search area and is not associated with an inmover or alternate address, it is considered a link within the nationwide search area.

Computer matching made the census Billy Bob at the sample address in Dallas the primary census record and coded it as a match to the PI Billy Bob. The other two census Billy Bobs were coded as duplicates.

3.3 Before Followup Clerical Matching

There were two phases of clerical matching with a field followup operation between them: BFU Clerical Matching and AFU Clerical Matching.

During BFU, clerical matching staff reviewed the results of computer matching and all the information provided by the PI respondents to determine who lived at the sample address on both Census Day and PI Day. They also looked for any inmover or alternate addresses the respondents may have provided. Clerical matching staff also searched for additional matches and duplicates. Comparing PI respondent data to the data of the linked census record, matchers could determine if the link and match codes the computer made were correct, or update any codes if necessary. In other words, matchers were confirming whether the computer-linked records indeed referred to the same person. If the matchers determined that the records did not refer to the same person, then they unlinked the records (i.e., the computer link was not confirmed). Moreover, if duplicate records were involved, they were also confirming whether the computer correctly identified which record should be the primary and which record(s) should be the duplicate(s). Recall that the primary census record represented where the person should have been counted in the census. If matchers did not have enough information to determine with confidence whether or not the linked records referred to the same person, then the case was sent out to the field for a followup interview to obtain more information.

Continuing the Billy Bob example (refer back to Figure 2), we will now clerically review the links made by the computer:

- Link 1 is a Match between PI Billy Bob and census Billy Bob at the sample address in Dallas. They appear to refer to the same person, so they are left a Match.

- Link 2 is between two census records: Billy Bob at our sample address in Dallas, and a Billy Bob in Alexandria (beyond the sample search area). Do these two records really refer to the same person? And if so, which record represents where Billy Bob should really be counted in the census (the primary) and which record is the erroneous enumeration (the duplicate)?

  Since PI Billy Bob told us that he sometimes visits Alexandria to see his son, and when we looked in the census we found a Billy Bob with similar demographics in Alexandria, we can be fairly confident that these two census records do indeed refer to the same person. The criteria here is that since PI Billy Bob and census Billy Bob in Dallas match and refer to the same person, anything we know about PI Billy Bob we may also assume about census Billy Bob in Dallas.

  Also, since Billy Bob told us that he only visited Alexandria briefly, and has really lived at the Dallas address for the last ten years (including Census Day), we know that Billy Bob should be counted in the census at his Dallas address. Hence, at this point, we can say the computer correctly made Billy Bob in Dallas the primary census record, and Billy Bob in Alexandria the duplicate record. In other words, the census correctly enumerated Billy Bob in Dallas, TX and erroneously enumerated Billy Bob a second time in Alexandria, VA.

Furthermore, since Billy Bob told us about the alternate address in Alexandria, this serves as a confirmation for this nationwide computer link that is in an alternate search area.

- Link 3 is also between two census records: Billy Bob in Dallas again, and Billy Bob in Tampa Bay. We ask the same questions here: Do these two records refer to the same person? And if so, which one represents where he should be counted in the census (the primary) and which one is the erroneous enumeration (the duplicate)? These two records could refer to the same person but we have no other information to help us confirm the nationwide computer link. We leave Billy Bob at the sample address in Dallas as the primary record since we believe that is where he should have been counted based on the information provided by PI Billy Bob, and we keep Billy Bob in Tampa Bay as a duplicate. Then both records will be sent out to the field for a PFU interview to determine whether the records actually do refer to the same person, and if so, where the person should have been counted.

3.4 Person Followup Interview

When matchers did not have enough information to confirm or unlink a nationwide computer link with confidence, the case was sent to the field for a followup interview (there were also different types of cases that required followup, which we will not go into detail about in this discussion). When there was no inmover or alternate address available for the matchers to be able to confirm the nationwide link, the records were sent to PFU in order to determine if the two records in different locations actually referred to the same person. One interviewer went to the sample address and another went to the nationwide address to conduct PFU interviews. The expectation was that if the same person lived or stayed at both of these addresses, one of the respondents would mention the other location as another address where the person lived or stayed. However, in order to preserve people's privacy, interviewers were not allowed to specifically ask the respondent about (or even mention) the other location where we linked the person, as census confidentiality rules prohibit interviewers from asking about the linked address.

Let's refer back to our example to see how this works. Recall that Link 3, the nationwide link between census Billy Bob in Dallas (at our sample address) and census Billy Bob in Tampa Bay (our nationwide address), is the case that requires followup because our matchers do not have enough information to confirm whether the records actually referred to the same person.

Two interviewers are sent out; one to Billy Bob's address in Dallas and the other to the address in Tampa Bay. The results of the interviews are as follows:

Interview at Sample Address (Dallas)

- No new information obtained.

Interview at Nationwide Address (Tampa Bay)

- Elderly lady informs us that she and Billy Bob, her husband, had lived here together since the 1960s and nowhere else.

- But she also mentions he passed away in 2008.
- After Billy Bob passed, she felt like he was still with her so she put him on the 2010 Census form.

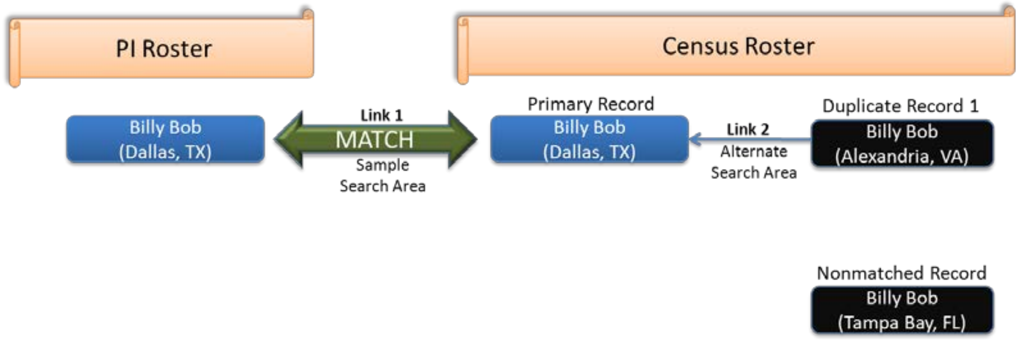
Now that the interviews are complete, it's time to move on to the AFU phase so the matchers can clerically review this new information and see if there's anything there that can help them resolve the case.

3.5 After Followup Clerical Matching

AFU was the last phase of the person matching operations. Matchers reviewed data gained from the followup interviews, including any new respondent-provided addresses, and updated match codes and other information as needed.

To continue the example, the interviews are complete and the data is ready for review. In Dallas, we find that Billy Bob answered our followup questions, but did not provide any new information that we can use to help solve the case. However, the Tampa Bay interview reveals that Billy Bob at that address actually passed away a few years ago and should have never been put on the 2010 Census form. With this in mind, the clerical matchers make their final adjustments and the results of AFU are shown in Figure 3.

FIGURE 3



The followup interview did not supply any new information to suggest changing anything with the records involved in Link 1 and Link 2, so matchers left them alone. Link 1 remains a match between the PI and census Billy Bob in Dallas, who also retains his status as the primary census record. Billy Bob in Alexandria remains an erroneous enumeration, coded as a duplicate to the primary Billy Bob in Dallas. However, matchers realized in Link 3 that the Billy Bob in Tampa Bay and the Billy Bob in Dallas do not represent the same person. So matchers completely unlinked the two records and the Billy Bob in Tampa Bay is now a Nonmatch (i.e., the nationwide computer link was not confirmed). Because the CCM does not affect the census, this record remains in the census, but in the CCM files the record is assigned a match code of Nonmatch and is not used for the final CCM coverage estimates.

4 2010 Census Coverage Measurement Person Matching Results

In this section, unweighted results from the 2010 CCM person matching operations are presented. As a reminder, all results presented here are given from an operational standpoint and do not reflect the final CCM estimates of person coverage. No inferences to the general population are intended.

4.1 E-sample Matching Results

Table 1 below shows the match codes assigned to E-sample records upon completion of each of the person matching phases: computer matching (prior to clerical review), BFU (prior to field followup), and AFU (final Person Matching and Followup operations result). E-sample records are census enumerations in housing units selected for the CCM sample. There were a total of 383,537 E-sample records for 2010.

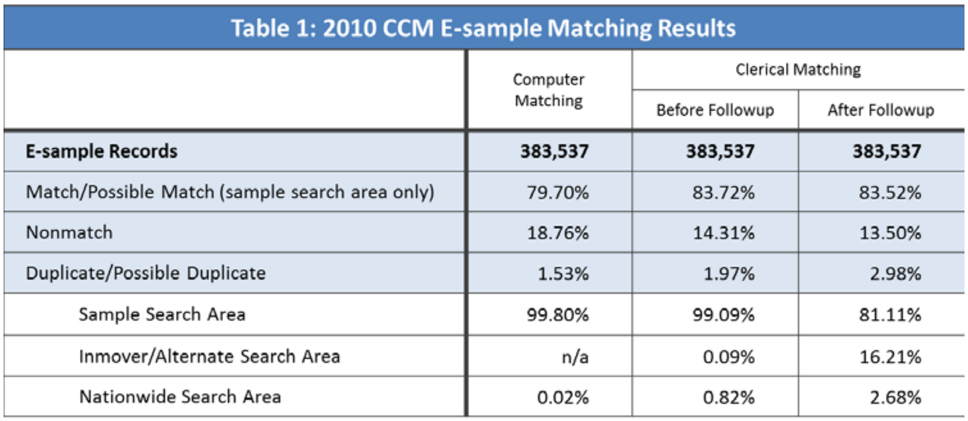


Table 1: 2010 CCM E-sample Matching Results

	Computer Matching	Clerical Matching	
		Before Followup	After Followup
E-sample Records	**383,537**	**383,537**	**383,537**
Match/Possible Match (sample search area only)	79.70%	83.72%	83.52%
Nonmatch	18.76%	14.31%	13.50%
Duplicate/Possible Duplicate	1.53%	1.97%	2.98%
Sample Search Area	99.80%	99.09%	81.11%
Inmover/Alternate Search Area	n/a	0.09%	16.21%
Nationwide Search Area	0.02%	0.82%	2.68%

The Match/Possible Match row contains E-sample records that matched, or possibly matched, to a PI record (which was by definition located within the CCM sample). The Nonmatch row contains E-sample records that did not match to a PI record and were not duplicates of another census record. If an E-sample record was duplicated in the census and it was determined that the person should have been counted in the sample block cluster, then the primary E-sample record is included in the Match/Possible Match row (if also linked to a PI record) or the Nonmatch row (if no link to the PI was found). Finally, the Duplicate/Possible Duplicate row contains E-sample records that were erroneously enumerated in the sample block cluster because they were duplicates of other census enumerations. The last three rows under the Duplicate/Possible Duplicate row indicate where the duplicate record was located in relation to the primary record.

In our Billy Bob example, the census Billy Bob located in Dallas would be an E-sample record and since he was coded as a Match to the PI Billy Bob throughout all phases of matching, he would be included in the Match/Possible Match row in the Computer Matching, Before Followup Matching and After Followup Matching columns. The other two census Billy Bob records in Alexandria and in Tampa Bay are not located in the CCM sample blocks clusters, and therefore are non E-sample records and are not included in Table 1.

Now that we have a better understanding of what is included in the table, we will discuss some of the key findings. Upon completion of the Person Matching and Followup operations, 83.52 percent of the 383,537 E-sample records were Matches or Possible Matches, 13.50 percent were Nonmatches, and 2.98 percent were Duplicates or Possible Duplicates. Of those, 81.11% were duplicates of another census record in the sample search area (i.e., both the primary and the duplicate record were in the sample block cluster or its surrounding blocks). However, there were also times when the E-sample record was a Duplicate or Possible Duplicate of another census enumeration beyond the surrounding blocks (i.e., a non E-sample record). In those cases, the non E-sample record was determined to be the primary record, indicating that the person who was enumerated multiple times in the census should not have been counted in the sample; the correct location was beyond the surrounding blocks and the E-sample record was an erroneous enumeration. Upon completion of AFU, 16.21 percent of the Duplicates and Possible Duplicates were identified in an inmover or alternate search area (i.e., found near an address provided by a PI or PFU respondent) and 2.68 percent were identified at some other nationwide location.

Table 1 shows that there were fewer E-sample Nonmatches after final clerical review than following computer matching or BFU Clerical Matching, with the larger change between computer matching and BFU (14.31 percent E-sample Nonmatches following BFU compared to 18.76 percent following computer matching, a difference of 4.45 percentage points). Recall that during BFU, not only did the clerical matching staff review the results of computer matching, but they also searched for additional matches and duplicates.

There were also more E-sample Duplicates/Possible Duplicates as a result of the AFU clerical review, with the larger change between BFU and AFU (2.98 percent following AFU compared to 1.97 percent following BFU, a difference of 1.01 percentage points). In a census duplicate link, the E-sample person was usually considered the primary (and thus coded a Match, Possible Match, or Nonmatch) unless there was further information based on a link to a PI person or further field followup to determine that the person should have actually been counted outside the sample cluster. Therefore, it is not surprising that there were relatively more E-sample Duplicates/Possible Duplicates due to the additional information collected during PFU (including other addresses where a person *could* have been counted) that the clerical matching staff reviewed to determine where the person should really have been counted (and in some cases, determine that the sample address was not the correct location). Thus, any E-sample record that was considered a primary of a nationwide record (a duplicate census record located beyond the surrounding blocks), was sent to followup. If it was determined that the two records actually did refer to the same person but the person should not have been counted at the sample address, then the E-sample person became the duplicate (instead of being the primary with a match code of Match, Possible Match, or Nonmatch during BFU).

4.2 Non E-Sample Matching Results

Table 2 below provides results for linked non E-sample people after each of the person matching phases. Linked non E-sample people are census enumerations beyond the sample search area (i.e., records that are located beyond the CCM sample block clusters and surrounding blocks), which are linked either to a PI record or to another census record (as either the primary or the duplicate). Unlike Table 1, which consisted of E-sample records, the universe of linked non E-sample records does change as we progress through the various matching operations because Table 2 only includes non E-sample records that are linked to either a PI record or another census record. There are many more non E-sample records in the census that are not linked to anyone, but these are excluded from the results presented here.

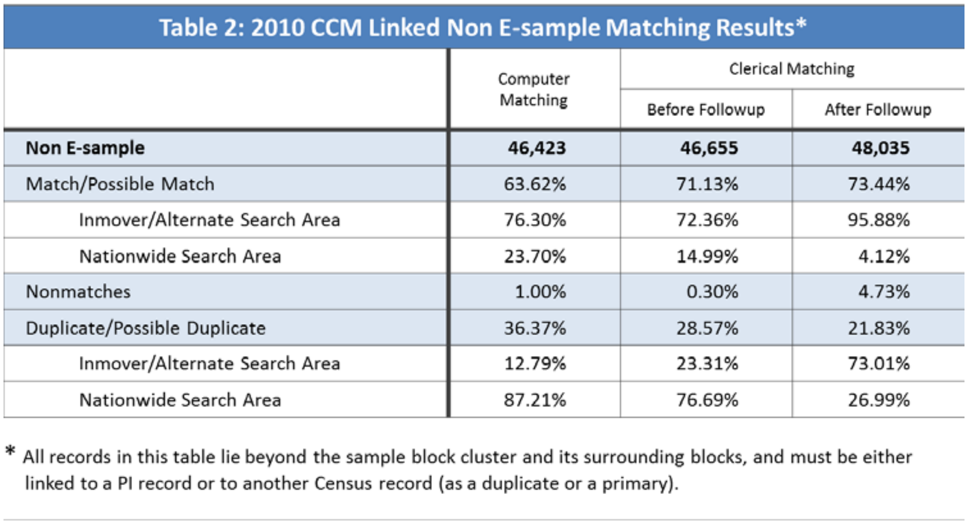


Table 2: 2010 CCM Linked Non E-sample Matching Results*

	Computer Matching	Clerical Matching	
		Before Followup	After Followup
Non E-sample	**46,423**	**46,655**	**48,035**
Match/Possible Match	63.62%	71.13%	73.44%
Inmover/Alternate Search Area	76.30%	72.36%	95.88%
Nationwide Search Area	23.70%	14.99%	4.12%
Nonmatches	1.00%	0.30%	4.73%
Duplicate/Possible Duplicate	36.37%	28.57%	21.83%
Inmover/Alternate Search Area	12.79%	23.31%	73.01%
Nationwide Search Area	87.21%	76.69%	26.99%

* All records in this table lie beyond the sample block cluster and its surrounding blocks, and must be either linked to a PI record or to another Census record (as a duplicate or a primary).

The Match/Possible Match row includes census people, located beyond the CCM sample clusters and surrounding blocks, which were linked to a PI record. Census records that were Matches or Possible Matches could also have been considered a primary of other census duplicate records. The Nonmatch row includes census records beyond the surrounding blocks that were not linked to PI records but were considered the primary record of another census duplicate record. The Duplicate/Possible Duplicate row includes census records, located beyond the surrounding blocks, which were duplicates or possible duplicates of other census enumerations. The two rows below the Match/Possible Match row and the two rows below the Duplicate/Possible Duplicate row indicate whether the non E-sample record was located near an inmover or alternate address provided by a PI or PFU respondent or whether the non E-sample record was located at some other nationwide location.

Upon completion of the Person Matching and Followup operations, 73.44 percent of the 48,035 linked non E-sample records were Matched or Possibly Matched to a PI record, 4.73 percent were Nonmatches (they were primaries of other census duplicate records), and 21.83 percent were Duplicates or Possible Duplicates. Of the census records beyond the surrounding blocks that were linked to PI people, 95.88 percent were found in an inmover/alternate address search area, and 4.12 percent were found in areas that the respondents did not tell us about. Of the census records beyond the surrounding blocks that were Duplicates or Possible Duplicates, 73.01 percent were found in an inmover/alternate search area, and 26.99 percent were found in areas that the respondents did not tell us about.

Note that as we progressed through the matching operations, there were relatively more non E-sample Duplicates/Possible Duplicates in an inmover/alternate search area than in some other nationwide location, with the biggest change coming between BFU and AFU (73.01 percent of the non E-sample Duplicates/Possible Duplicates were located near an inmover or alternate address following AFU compared to 23.31 percent following BFU, a difference of 49.70 percentage points). Remember, following BFU, if there was no inmover or alternate address to confirm that the two records in a nationwide link actually did refer to the same person, then the records were sent to PFU to determine if they actually were the same person and if so, where the person should have been counted. During AFU, if the clerical matching staff could determine from the followup information that the nationwide address was an inmover or alternate location for the person, then the nationwide link was confirmed and was then considered located in an inmover/alternate search area. Thus, it is not surprising that relatively more of the Duplicates/Possible Duplicates were located in an inmover/alternate search area following AFU.

To close out our Billy Bob example, we will discuss where the remaining two census records (the ones that were not in our Dallas sample block cluster) would be included in Table 2. Billy Bob in Alexandria was coded as a duplicate throughout all phases of matching, so he would be included in the Duplicate/Possible Duplicate row in the computer matching, BFU, and AFU columns. Furthermore, he was located in an alternate address provided by the PI respondent, so under the Duplicates/Possible Duplicates he would also be represented in the Inmover/Alternate Address Search Area row in each of the three columns. Billy Bob in Tampa Bay was coded as a duplicate during the computer matching and BFU matching phases, and unlinked during AFU matching. Therefore, he would have been included in the Duplicate/Possible Duplicate row in the Computer Matching and BFU columns. Furthermore, since there was no indication from any respondent that Billy Bob lived or stayed in Tampa Bay, then the Tampa Bay location would simply be in a nationwide search area. Thus Billy Bob in Tampa Bay would have also been included in the Nationwide Search Area row under the Duplicates/Possible Duplicates for the Computer Matching and BFU columns. During AFU, we determined that the Billy Bob in Tampa Bay was not the same as the Billy Bob in Alexandria, so he was unlinked and would not be included in any of the rows of the AFU column. (Note that he would not be included in the Nonmatches row because while that row does indeed contain nonmatched records, those records must also be a primary to some other record. In other words, they must have a duplicate attached. Since Billy Bob in Tampa Bay does not have any duplicates attached to him, he would not be in the Nonmatches row.)

4.3 Disposition of Nationwide Matches and Duplicates

As discussed above, computer matching linked PI people to census people throughout the country and also searched for census duplicates throughout the country. Any match or duplicate from the nationwide computer matching that was determined to be within an inmover or alternate search area did not need to be sent to PFU (because there was already information from PI that the records referred to the same person). The other nationwide links were sent to PFU to determine if the two records in different locations actually referred to the same person and if so, where the person should have been counted. After field followup and clerical review, if it was determined that the nationwide address corresponded to a respondent-provided address (i.e., was within an inmover or alternate search area), then that served as confirmation that the two records actually did refer to the same person.

Table 3 below presents the final disposition of nationwide computer matches and duplicates (i.e., non E-sample records beyond the surrounding block that the computer linked to a PI record or another census record as a primary or duplicate). The left column represents the computer matching match code of the non E-sample records and in which search area the computer found the link. The right three columns categorize the final disposition for records in each of these rows after the Person Matching and Followup operations were completed. The final disposition categories are:

- **Confirmed -** The non E-sample record was confirmed to refer to the same person as a PI record or another census record.
- **Not Confirmed -** The non E-sample record was unlinked because it does not refer to the same person as the PI or census record that it had been linked to.
- **Undetermined -** It is unknown whether or not the non E-sample record refers to the same person as the PI or census record that it was linked to.

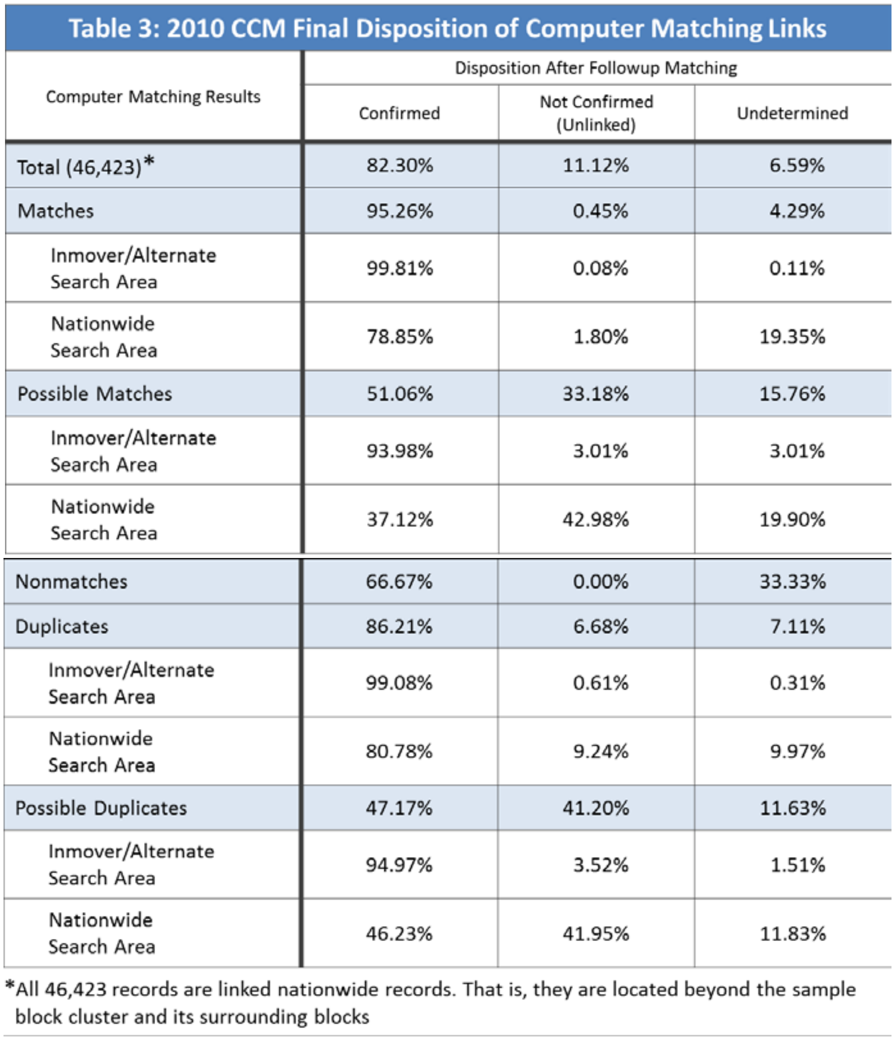


Table 3: 2010 CCM Final Disposition of Computer Matching Links

Computer Matching Results	Disposition After Followup Matching		
	Confirmed	Not Confirmed (Unlinked)	Undetermined
Total (46,423)*	82.30%	11.12%	6.59%
Matches	95.26%	0.45%	4.29%
Inmover/Alternate Search Area	99.81%	0.08%	0.11%
Nationwide Search Area	78.85%	1.80%	19.35%
Possible Matches	51.06%	33.18%	15.76%
Inmover/Alternate Search Area	93.98%	3.01%	3.01%
Nationwide Search Area	37.12%	42.98%	19.90%
Nonmatches	66.67%	0.00%	33.33%
Duplicates	86.21%	6.68%	7.11%
Inmover/Alternate Search Area	99.08%	0.61%	0.31%
Nationwide Search Area	80.78%	9.24%	9.97%
Possible Duplicates	47.17%	41.20%	11.63%
Inmover/Alternate Search Area	94.97%	3.52%	1.51%
Nationwide Search Area	46.23%	41.95%	11.83%

*All 46,423 records are linked nationwide records. That is, they are located beyond the sample block cluster and its surrounding blocks

Computer matching found a total of 46,423 non E-sample records that linked to either a PI record or another census record. Of those original nationwide links from computer matching, 82.30 percent were confirmed (i.e., information about an inmover or alternate location provided by the PI and/or PFU respondents confirmed that the two records in different locations referred to the same person). After final clerical review, 11.12 percent of the original nationwide links from computer matching were unlinked and not confirmed (i.e., it was determined that the records did not refer to the same person). For 6.59 percent of the original nationwide links, the CCM Person Matching and Followup operations were unable to determine whether the two records linked by the computer actually referred to the same person.

In the end, a majority of the Matches and Duplicates from the nationwide computer matching were confirmed (95.26 percent of the Matches beyond the surrounding blocks were confirmed and 86.21 percent of the Duplicates beyond the surrounding blocks were confirmed). Note that when the computer linked a non E-sample record in an inmover/alternate search area, whether as a Match or a Duplicate, clerical matchers confirmed the record over 99 percent of the time. Among the weaker links from the nationwide computer matching (i.e., the Possible Matches and Possible Duplicates), there were relatively fewer links confirmed. Among the non E-sample records beyond the surrounding blocks that were Possible Matches of PI records but were not found in an inmover or alternate search area following computer matching, 37.12 percent were confirmed, 42.98 percent were not confirmed (unlinked), and 19.90 percent remained undetermined. Among the non E-sample records beyond the surrounding blocks that were Possible Duplicates but were not found in an inmover or alternate search area following computer matching, 46.23 percent were confirmed, 41.95 percent were not confirmed (unlinked), and 11.83 percent remained undetermined.

4.4 Enumeration Status For E-Sample Records

Remember that in addition to identifying duplicates, the CCM wanted to determine where each person should have been counted. So in addition to match and duplicate status, an enumeration status was determined for each E-sample record. The enumeration status indicated whether an E-sample record should have been counted in the census based on the 2010 Census Residence Rule. Table 4 below presents the enumeration status for each E-sample record based on the codes assigned during computer matching, BFU, and AFU.

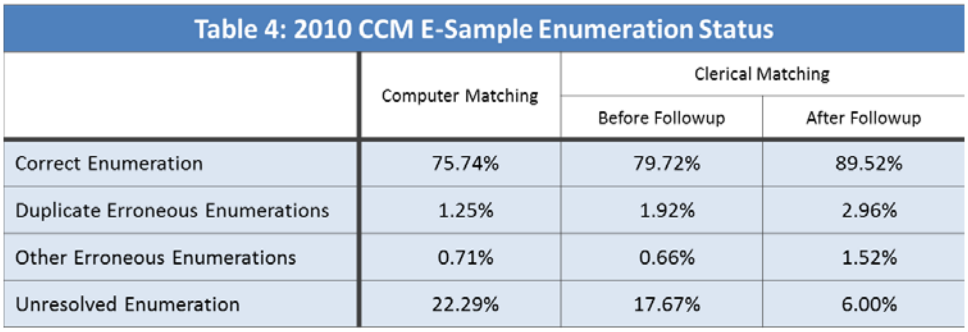


Table 4: 2010 CCM E-Sample Enumeration Status

	Computer Matching	Clerical Matching	
		Before Followup	After Followup
Correct Enumeration	75.74%	79.72%	89.52%
Duplicate Erroneous Enumerations	1.25%	1.92%	2.96%
Other Erroneous Enumerations	0.71%	0.66%	1.52%
Unresolved Enumeration	22.29%	17.67%	6.00%

Upon completion of the Person Matching and Followup operations, 89.52 percent of the E-sample records were correct enumerations, 2.96 percent were erroneous enumerations due to duplication, 1.52 percent were other erroneous enumerations, and 6.00 percent went unresolved. If an E-sample record referred to a person that was counted more than once and CCM determined that the E-sample record represented the location where the person should have been counted, then the record was a correct enumeration. Otherwise, if the person should have been counted somewhere else then the record was coded a duplicate, which was one type of erroneous enumeration.

Note that there were fewer unresolved cases as result of the final clerical review (6.00 percent were unresolved enumerations following AFU compared to 22.29 percent following computer matching (prior to any clerical review) and 17.67 percent following BFU (prior to field followup).

5 Conclusion

Computer and clerical matching operations within the sample search areas have been conducted in the past but for the first time, the 2010 census coverage evaluation program also conducted a computerized nationwide search with a clerical review and followup to identify census duplicates and determine where these people should have been counted. The data collected in the 2010 CCM PI and PFU operations provided information needed to both identify records for people that were enumerated multiple times in the census (near and far) and make decisions about which of these records represented where the person should have been counted (the primary) and which were the erroneous enumerations (the duplicates). From examining the results of computer and clerical matching, we can see that the computer did very well when linking records together, especially when inmover or alternate addresses were obtained from the PI respondents confirming the links found beyond the sample search area. When information from PI respondents could be used to confirm the links from the nationwide computer matching, we were able to avoid sending the case to the field for followup. Reducing the field followup workload saves money and has the benefit of determining enumeration status based on the results of data collected closer to Census Day (by using information collected during the PI interview instead of the later PFU interview). It is therefore recommended to continue the computerized nationwide search for matches and duplicates and targeted searches around respondent-provided inmover and alternate addresses. Further, it is recommended that additional research be conducted to refine computer matching and automated coding techniques to further improve results for future census coverage evaluations.

References

Johnson, Susanne, Patricia Sanchez, Anne Wakim, and Kopen Henderson (2012). “Assessment for the 2010 Census Coverage Measurement Person Matching and Followup Operations,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-I-24, U.S. Census Bureau.

Kostanich, Donna, David Whitford, and William Bell (2004b). “Plans for Measuring Coverage of the 2010 U.S. Census,” American Statistical Association Joint Statistical Meetings, 2004 Proceedings of the Section on Survey Research Methods.

Mule, Thomas (2012). “2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-01, U.S. Census Bureau.

Singh, Rajendra P. (2005). “2010 Census Coverage Measurement – Updated Plans,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-A06, November 29, 2005.

U.S. Census Bureau (2001). “Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses.”