Potential Duplicates in the Census – Methodology and Selection of Cases for Followup[1]

Leah Marshall
U.S. Census Bureau, 4600 Silver Hill Rd, Washington DC 20233

Abstract

As the U.S. population becomes more mobile and people begin living in more complex living situations, there is greater opportunity for an individual to be counted in the census more than once. The Census Bureau uses computer matching programs to identify potential duplicates, but these potential duplicates need to be contacted to determine at which residence the person should be counted according to Census residence rules. This gives rise to a larger issue of how to select and follow up with these potentially duplicated individuals. Duplicates are not always necessarily pairs – some people could even be listed on three or four forms, which complicates processing. Also, a person may be duplicated between two housing units or between a housing unit and a group facility – another complication of processing, as the Census Bureau counts people in these types of living quarters separately. People can sometimes also be duplicated not because of a complex living situation, but because of problems with the Census Bureau's Master Address File. This paper describes the methodologies for selection and followup of potential duplicates in the 2010 intercensal tests. In particular, this paper contains an examination of findings from the 2004 Census Test and how these findings shaped the methodology and selection of duplicates in the 2006 Census Test. Findings from the 2006 Census test and how results from the 2006 Census Test then shaped the direction of the methodology and selection of duplicates in the 2008 Census Dress Rehearsal will also be discussed.

Key Words: Case Selection, Census, Coverage, Duplication

1. Background

1.1 Types of Duplication

A duplicate occurs when a person is counted in the census more than once. Take for example, a college student who is included on his parents' census form and on his census form at college. Or, take for example a housing unit whose address can be represented in two different ways; the housing unit may appear on the Census Bureau's Master Address File twice, and the people and the housing unit may be counted in the Census twice. Considering these examples we can see that there are two primary reasons for duplication: person-level duplication, shown in the former, and housing-level duplication, shown in the latter.

Person-level duplication refers to duplication caused by a person's living situation. Some examples include: duplication caused by a vacation home, joint custody situations, college attendance, moving, or other part-time residency situations. The people are duplicated because they have been incorrectly included on more than one census questionnaire. Usually, this duplication occurs because of a misunderstanding of Census' residence rules for census day. The cause of this type of duplication lies inherently with the people and their possibly complex living situations. For this reason, person-level duplication issues are also sometimes referred to as living situation duplication issues.

Housing-level duplication refers to person duplication caused by issues with the address list. Since the cause of this type of duplication lies inherently with the address list and not with the people, housing-level duplication issues are also sometimes referred to as addressing issues. These duplication issues can be more difficult to envision than the

[1] *This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.*

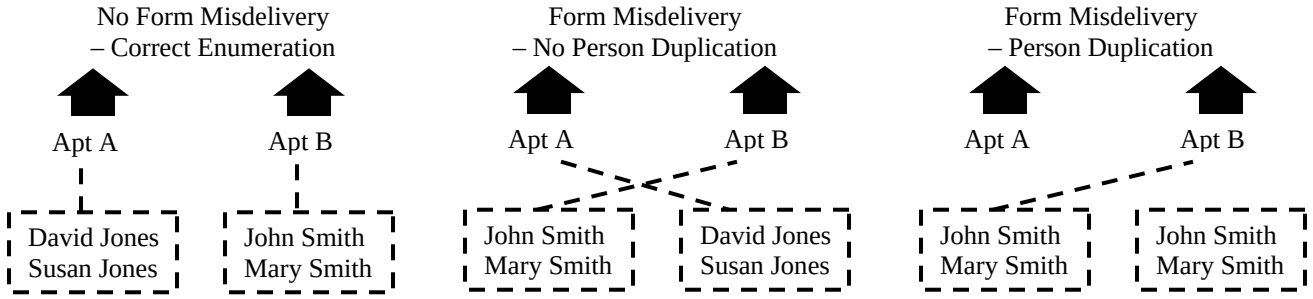
person-level issues previously mentioned. Below are some common examples of addressing issues that can result in person duplication.

Housing Unit Duplicates – Two seemingly different addresses on the address list actually represent the same unit on the ground; the housing unit could appear on the Census Bureau's Master Address File twice, and the people could be duplicated.

Nonexistent Housing Units – A housing unit on the Census Bureau's Master Address File may no longer exist at the time of enumeration. For example, the housing unit could have been recently demolished, or a duplex could have been remodeled into a single-family house. The post-office may deliver a questionnaire for the nonexistent address to an address nearby; that nearby address gets multiple questionnaires, and the people could be duplicated.

Form Misdelivery – A census questionnaire may be delivered to the wrong housing unit, and neighbors may receive each other's census form. There are procedures for followup interviewers to correct these form misdelivery situations in the field, but if the interviewers do not or cannot correct the problem, person duplication will occur. See Figure 1 below for a detailed example.

Figure 1. Form Misdelivery Scenarios



- Apt A and Apt B both receive the correct questionnaire.
- The residents at Apt A and at Apt B return the questionnaires.
- The residents at Apt A and at Apt B are correctly enumerated.

- Apt A receives Apt B's questionnaire and vice versa.
- The residents at Apt A return the questionnaire labeled Apt B; the residents at Apt B return the questionnaire labeled Apt A.
- The residents at Apt A and at Apt B are correctly enumerated, albeit at the wrong addresses.

- Apt A receives Apt B's questionnaire and vice versa.
- The residents at Apt A do not return the questionnaire; the residents at Apt B return the questionnaire labeled Apt A.
- The Census Bureau conducts followup to get the missing questionnaire. The IDs on the returned forms indicate which units require followup. The form for Apt A has been returned, so the Census Bureau conducts followup with Apt B.
- The residents in Apt B are duplicated.

1.2 Duplication in Census 2000

In the midst of Census 2000 an unexpectedly high duplication rate was discovered. As such a late, ad-hoc unduplication operation was mounted with over 2.4 million housing units marked for potential deletion using name matching and/or address matching rules. Additional research into the types of duplicate situations then determined rules for whether the housing units should be deleted from final census counts or reinstated. As a result of the unduplication operation in Census 2000, one million housing units were reinstated and the remaining 1.4 million

were deleted from final Census counts. The plan for later censuses was to have a functional unduplication operation in place from the start. This paper examines some of the research and tests done throughout the decade in an effort to implement a successful unduplication operation.

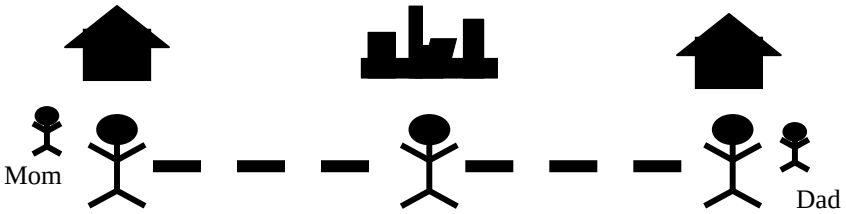
1.3 Overview of Unduplication Methods

Before examining the detailed methodologies for case selection, we first begin with an overview of the unduplication process used during the 2010 intercensal tests.

The first step of unduplication processing is to determine a list of potentially duplicated people. To do this, the Census Bureau matches the entire Census against itself through the use of computer matching programs, namely the Duplicate Person Identification software. This processing system and software compares persons listed on census returns. People are matched using various characteristics such as first name, last name, phone number, and date of birth, among others. People in housing units are matched to people in other housing units and to people in group quarters. There is no group quarters to group quarters matching done during this process.[2] The computer matching process involves multiple passes of the matching system in which the matching parameters and constraints are varied for each pass. The matches are scored and ranked and cutoffs are set. No duplicates are removed from the Census during computer matching, nor is a person's actual duplication status resolved during this phase of the process.

The next step, and the focus of this paper, consists of case selection and case preparation – the methodology and selection of unduplication cases for followup. After computer matching creates a list of potentially duplicated people, the Census Bureau processes the list to determine which cases to followup with and the method of followup. It is important to note that at this stage of the process, duplicates are not always necessarily pairs. Although it is rare, some people could even be listed on three or four census forms. Take for example a college student whose parents are divorced. He could be listed on his mother's census form, his father's census form, and on his college census form (see Figure 2 below). We refer to these situations as large clusters. That is, a large cluster is a duplicate situation which involves more than two households. These and other situations are taken into account as part of case selection. More details will be discussed in section 3 of this paper.

Figure 2. Example of a Large Cluster of Duplicates



The final step of the unduplication process is to followup with the potential duplicates to see whether the people are in fact duplicated, and if so, where to count them. Because there are two types of duplication issues – housing-level issues and person-level issues – research this decade has focused on two approaches for followup. For obvious reasons, followup for person-level issues must focus on the people, while followup for the addressing issues should focus on the addresses.

For the person-level issues, a detailed interview is administered to determine the person's living situation – either joint custody, college, work, military, etc – and then where that person spends most of his or her time. The most straightforward way to determine whether a person is duplicated would be to simply ask the respondent. Because of privacy issues, however, the Census Bureau cannot reveal the "other half" of the duplicate to the respondent. So instead, the general approach for the person-level duplication issues has been to contact each household involved in the duplicate, that is, to contact each "half". We then ask the respondents more open-ended and leading questions to

[2] Also note that the 2004 Census Test was the first Census Test that included unduplication, and during that test there was no housing unit to group quarters matching.

find out about their living situations. For example, to find out if anyone in a particular household was in a college living situation, we would ask the respondent, "Did you or anyone in this household often stay somewhere else to attend college?" Then, if the answer is yes, in order to determine where those persons should be counted, we would ask "Where do you/they stay most of the time, at this address or at that other address?" If the responses result in only one location where the person should be counted, we have corrected the duplication.

For the housing-level duplication issues, administering such an interview will not resolve the duplication. Instead, it is necessary to send a Census worker into the field to observe the actual addresses in question. By fixing the address list, we will fix the duplication.

2. Limitations and Assumptions

The 2004 and 2006 Census Tests were site tests and therefore results were not drawn from a nationally representative sample. The 2004 Census Test was conducted solely in areas of Queens, New York and in areas of southwest Georgia. The 2006 Census Test was conducted in Travis County, Texas and in the Cheyenne River Indian Reservation and Off Reservation Trust Lands in South Dakota. The 2008 Census Dress Rehearsal was conducted in San Joaquin County, California, and select counties surrounding Fayetteville, North Carolina.

Relatively few person-level duplication situations have appeared in the site tests because these tests tend to cover a relatively small geographic area. For example, we would not generally expect a person's regular address and his vacation home address to be contained in the same county. Therefore, the focus of this paper is on methodology and distinguishing between person-level duplication issues and housing-level duplication issues, not on methodology to improve resolution of person-level duplication.

The interviews and followup strategies used for these duplication cases are conducted only at housing units. Followup with persons at group facilities such as college dorms, jails, and nursing homes is considered out of scope. For the detailed interview, census residence rules generally state that the person should be counted at the group facility and not at another housing unit, making followup with persons at group facilities unnecessary. For the addressing issues, we rely on other Census field operations with more experienced field staff to ensure that the housing unit and group facilities' address lists are correct.

Results and methodologies discussed in this paper are based upon the computer matching algorithms used at the time of each census test. While methods of selection and followup continued to change throughout the decade for these unduplication cases, so did the computer matching algorithms to identify the possible duplicates. The changes to the algorithms involve changes in the number of blocking passes, changes to the type of blocking passes, and constraints placed on the matches, among others. As mentioned above, cutoffs are set as part of the Duplicate Person Identification software. Since it would be inefficient to send cases that are not actual duplicates to a costly followup, these cutoffs are set so that we believe all the people above the cutoffs are in fact true duplicates. This part of the software, which would affect the methodology of followup, has not changed throughout the decade, and hence changes to the Duplicate Person Identification software are not discussed in this paper.

3. Methodologies and Results

3.1 2004 Census Test

The 2004 Census Test was the first opportunity to test our planned unduplication effort. The universe of Census housing units was matched against itself to obtain a list of people who were thought to be potential duplicates. The planned followup method for these potential duplicates was the Coverage Research Followup operation, which also encompassed other types of cases thought to have coverage issues.

The major issue when heading into the 2004 Census Test was to determine a priori which duplication issues were a result of person-level duplication situations and which were a result of housing-level duplication situations. As mentioned before, person-level issues are best resolved by talking to the persons involved in the duplicate and asking detailed questions about their living situations; this type of interview can take place over the phone or in

person. Housing-level duplication issues, on the other hand, are best resolved in the field by observing the housing units and the addresses in question. Because this means potentially different methods of followup for each type of duplication issue, the distinction needs to be made ahead of time.

One theory behind how to determine this distinction ahead of time was through the use of a "whole/partial household" indicator. This indicator is assigned to each duplicate link pair after the links are made. It is designed to distinguish how many of the people in a household are linked. The three different values for this indicator are defined below.

Whole-to-Whole Household Match – A match between two Census returns in which all of the people in one household (whole match) match to all of the people in the other household (whole match).

Whole-to-Partial Household Match – A match between two Census returns in which all of the people in one household (whole match) match to some of the people in the other household (partial match).

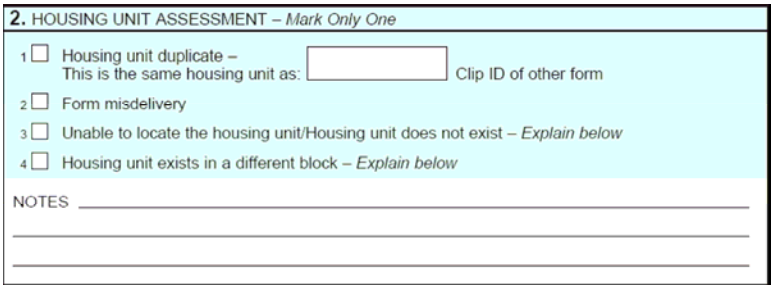
Partial-to-Partial Household Match – A match between two Census returns in which some of the people in one household (partial match) match to some of the people in the other household (partial match).

In the 2004 Census Test, the distinction between person-level duplication issues and housing-level duplication issues was determined through the use of this "whole/partial household" indicator. It was thought that the whole-to-whole matches represented housing-level duplication issues – if an address is duplicated, all the people in a household would be duplicated. The partial-to-partial matches were thought to represent person-level duplication issues – most living situations (such as being away for college, work, or joint custody) only involve one or two people in the household and not the entire household. The whole-to-partial matches were then thought to represent some combination therein.

Therefore, when determining how to followup with cases, the whole-to-whole and whole-to-partial household matches were sent, as pairs, straight for a field interview, whereas the partial-to-partial matches did not require a field component and were sent first for a telephone interview. Those interviews that were not completed over the telephone recycled for an attempt at a field interview.

The field interview consisted of two parts. The first part of the interviewer form included a housing unit assessment box where the interviewer could indicate whether the duplicate represented a housing unit duplicate, a form misdelivery situation, or any other notes he or she had about the address (see Figure 3 below). The second part of the interview consisted of detailed questions about potential living situations that members of the household might be in. As mentioned before, because of privacy issues the Census Bureau does not reveal the "other half" of the duplicate to the respondent. Rather, the interviewer asks open-ended, leading questions to determine any potential living situations the respondents might have and where their household members stay most of the time. The telephone interview consisted only of this section of the interview, asking detailed questions about potential living situations.

Figure 3. 2004 Coverage Research Followup Form – Housing Unit Assessment Box



Because 2004 was the first planned unduplication effort, large clusters were excluded from the main processing stream for the Coverage Research Followup operation. They were instead included as part of a separate, informal followup phone call performed by clerks at the Census Bureau's National Processing Center.

Following the Coverage Research Followup interviews, the interviews were re-linked into pairs of duplicates. Since the 2004 Census Test was the first test that included unduplication, the goal of this research phase was to assess the true outcomes of the duplicates and not necessarily to assess what outcome could or would be attained in an automated environment. Therefore, experienced clerks reviewed each set of duplicates to determine a final outcome. The clerks determined whether the situation was a person-level duplication situation or whether it was a housing-level duplication situation, and they also determined the specifics of each – whether a form misdelivery or a housing unit duplicate, or whether someone was away for college or work or another reason, and at which address the respondents should be counted. The outcomes of the clerical review were not incorporated into final census test results.

The results of the clerical review were not quite as we had expected. It turned out that whole-to-partial matches acted more like whole-to-whole matches, but partial-to-partial matches over a short distance also acted like housing-level situations. Our results conclusively showed that whole-to-whole matches within the same geographic block were more likely to be caused by addressing issues. Our results also indicated that whole-to-partial and partial-to-partial matches within the same geographic block were likely to be caused by addressing issues, but the results were not as conclusive.

The housing unit assessment box did not perform as well as expected, either. The majority of the interview was very person-focused – the housing unit assessment box took up only a small part of the lengthy questionnaire. Hence, enumerators would begin the interview by finding the people and not necessarily the addresses. Their instructions were to go to the address indicated on the form, use the housing unit assessment box as necessary, and continue with the rest of the interview. Many form misdelivery situations and housing unit duplicate situations occur in multi-unit structures (apartment buildings, condominiums, etc). Oftentimes the enumerator would go to the street address without regard to unit designation, look at the names on the questionnaires, and mark down "duplicate" because the people were duplicated. Training enumerators to separate the concepts of persons and addresses in such a person-focused field operation ended up being more difficult than envisioned.

3.2 2006 Census Test

The 2006 Census Test was our major opportunity to expand upon what was learned in the 2004 Census Test and design a system that could be used in a full-scale census. The entire universe of persons in Census housing units was again matched against itself. This time it was also matched against the universe of Census group quarters to obtain a list of people who were thought to be potential duplicates. The followup method for these potential duplicates in 2006 was the Coverage Followup operation, which again encompassed other types of cases also thought to have coverage issues. The outcomes of the 2006 Coverage Followup operation were incorporated into final 2006 Census Test resuls.

This time, instead of sending all duplicate cases for the full-length interview, only the cases thought to be caused by person-level duplication issues were sent for a full-length interview. Again, they were sent first for a telephone interview, and then recycled to the field if we could not contact the respondent by phone. The duplicates thought to be caused by housing-level duplication issues were sent for an in-person field visit called a Housing Unit Verification. The Housing Unit Verification operation was developed to mimic an address listing operation. The enumerators had each pair (or large cluster) of addresses on one sheet, and their job was to physically visit the location of each address and either "verify" the address, "delete" the address, or mark the address as a "duplicate" of another address on the list.

Since the results of 2004 indicated that the whole-to-whole matches within the same geographic block were more likely to be addressing issues, all whole-to-whole matches within the same geographic block were sent for a Housing Unit Verification visit. Since the results of 2004 were not as conclusive for the whole-to-partial matches or the partial-to-partial matches within the same geographic block, the whole-to-partial matches within the same block

and the partial-to-partial matches within the same block were sampled so that some were sent for the full-length Coverage Followup interview and some were sent for the Housing Unit Verification visit.

The goal of the 2006 unduplication processing was to effectively simulate and test a process that could be used in a full-scale decennial census. There would not be enough time in a full-scale decennial census for analysts to review every single duplicate and make a determination on each. Therefore, the final coding of each situation was turned into an automated procedure for the 2006 Census Test. In a full-scale decennial census, there is also not enough time to re-link the duplicate pairs or large clusters after interviewing, and therefore the pairs and large clusters were separated for processing up front. This also allowed followup and processing of large clusters. For those duplicates sent for a full-length coverage followup interview, each individual "half" of the interview was used to determine whether or not a person should be counted at that address regardless of whether the duplicate was a pair or a large cluster. For those duplicates sent for a Housing Unit Verification, the address code (either verify, delete, or duplicate) on each address was used to determine the resolution of that particular address, again regardless of whether the duplicate was a pair or a large cluster.

The results of the 2006 test showed that more potential duplicates were successfully deleted, or in other words, the duplicate situation was successfully resolved, by sending the whole-to-partial and partial-to-partial matches to the Housing Unit Verification rather than to the full-length Coverage Followup interview. Therefore we recommended that for future unduplication efforts, all duplicates within the same geographic block, regardless of whether they are whole-to-whole, whole-to-partial, or partial-to-partial matches should be sent to an address check operation similar to the Housing Unit Verification operation.

Through observations of the 2006 Census Test, and successful results of the 2006 Housing Unit Verification operation, we were able to determine that enumerators did in fact understand their jobs better for the Housing Unit Verification cases. We should therefore continue sending duplication cases resulting from address issues to an address listing operation.

3.3 2008 Dress Rehearsal

The plans for the 2008 Dress Rehearsal included sending the within-block duplicates to a listing check operation, this time called Field Verification. Field Verification is an existing operation which is designed to take one final look at the entire address list and verify or delete certain addresses before final Census tabulations. By sending the housing-level duplication issues to this operation, the lister can better make a determination about the address problems. He or she can see the entire list of units in a building and make a more accurate decision on whether a given unit was duplicated.

Since the 2006 Census Test validated our hypothesis that duplicates within the same geographic block were more likely to be caused by addressing issues, we decided to expand this search for housing-level duplication issues to geographic blocks that surrounded one another. We believe that errors with the geographic block-codes for the addresses could be the cause of some housing unit duplication, and we believe that the rates for which person-level duplication issues occur in surrounding blocks will also be relatively low. For example, we believe that college students and people with another place for work do not usually have that second address right down the street from the first. While shared custody cases could certainly occur within surrounding blocks, we believe these rates to be relatively low as compared to the rates of geocoding errors, or at least worth testing. The plans for the 2008 Dress Rehearsal therefore included sampling the duplicates located in surrounding geographic blocks to see if these duplicates were more likely caused by person-level issues or housing-level issues. A sample of duplicates in surrounding blocks were to be sent for the full-length interview, and a sample were to be sent for the Field Verification. Due to descoping of the 2008 Dress Rehearsal, the Field Verification operation was removed. Therefore, all surrounding block links were again sent for the full-length interview in order to test our hypothesis from the one side still available. Although we will not be able to tell how well these surrounding block cases would have done with an on-the-ground Field Verification check, we will be able to distinguish in our analysis surrounding block cases from other types of cases receiving the full-length interview. As a result, we will at least be able to evaluate how well surrounding block cases do in a full-length interview as compared to duplicates across further geographic distances.

Acknowledgements

The author wishes to thank and acknowledge Elizabeth Krejsa and Robin Pennington for their continued guidance as well as their contributions to this research, analysis, and presentation. The author also wishes to thank Edward Banz for his contributions to the 2006 analysis.

References

Krejsa, Elizabeth A., Linse, Kyra, Kostanich, Martine, Heimel, Sarah, Marshall, Leah, Banz, Edward, and King, Ryan (2007). "2006 Census Test Evaluation #2: Coverage Improvement, Final Report," U.S. Census Bureau Internal Memorandum from Whitford to Vitrano, DSSD 2006 Census Test Memorandum Series #F-05, September 24.

Marshall, Leah (2007). "Analysis Plan for Unduplication Cases in the 2006 Census Test Evaluation of Coverage Improvement, " U.S. Census Bureau Internal Memorandum from Marshall to Pennington, DSSD 2006 Census Test Memorandum Series #I-12, June 18.

Pennington, Robin A. (2005a). "2004 Census Test Evaluation 8: Evaluation of Person Duplication, Final Report," U.S. Census Bureau Internal Memorandum from Vitrano, 2010 Census Test Memorandum Series, Chapter 2004 Census Test, Memo # 43, September 29.

Pennington, Robin A. (2005b). "Unduplication of Persons and Housing Units in the 2004 Census Test". In *2005 JSM Proceedings*, Social Statistics Section. Alexandria, VA: American Statistical Association.