Census Unduplication Research Plan for 2010

Damon R. Smith and David C. Whitford
U.S. Bureau of the Census, 4301 Suitland Road, Room 2410/2, Suitland MD 20746

KEY WORDS: Census unduplication, Census 2010

1. Introduction

One of the important findings from Census 2000 was that duplication of person data has an significant impact on census results. Post-Census studies uncovered significant duplication of both housing units and persons. In Census 2000, an effort to reduce duplication was launched in an operation known as the Housing Unit Unduplication Operation. This operation identified and removed what was thought to be apparent duplicates from the census inventory of addresses. This process, however, was a computer-only effort. Its results pointed out the need for a more rigorous unduplication program to be built into census operations for the coming census.

Looking ahead to Census 2010, the United States Census Bureau decided that continued research would be needed to produce more effective and efficient duplicate identification and removal strategies to help solve duplication in the census. The Census Bureau is looking to broaden the scope of its approach to this problem by identifying duplicates early in census operations. Implementation of a program that identifies duplication after completed census questionnaires are returned may increase the likelihood of counting persons and housing units only once. The 2010 Research and Development Planning Group for Coverage has chartered a work group to support the development of strategies and requirements for identifying and removing duplication in the census. The Unduplication Work Group has developed proposals for research aimed at gaining knowledge for various aspects of identifying and removing duplication. The objective of the Census Unduplication Research Plan for 2010 is to identify and remove duplication in real time during the next decennial census. One of the proposals of the group is to determine the feasibility of a real-time person unduplication effort as census questionnaires are returned. This effort is being implemented and tested for the 2004 Census Test. The unduplication effort will entail a computer matching program to identify potential duplicate persons and a followup effort to discern residence status of the person(s) in question. This information will then be used to help the US Census Bureau confirm if the person is a duplicate.

In this paper we will summarize the major evaluations of duplication from Census 2000 and present the proposed research projects that may lead to a reduction of duplication in the census of 2010.

2. Census 2000

Housing Unit Duplication in Census 2000

The Census Bureau's analysis of the Decennial Master Address File, the address list for Census 2000, indicated some regions of the country where the number of housing units was in excess of the predicted housing unit count. A team was created to perform research and field work to identify attributes of areas in which housing unit coverage was too high and identify the causes of such overcoverage. As a result of the research and continuing concerns about housing unit overcoverage due to duplication, it was decided that a process needed to be implemented to identify and remove (from the census) duplicate housing units that still remained on the decennial file after all data collection activities had been completed. This operation was known as the Housing Unit Unduplication Operation (Nash, 2000).

The goal of the Housing Unit Unduplication Operation was to ensure that a housing unit was accounted for only once and not duplicated in the data files. This operation was conducted prior to the creation of the final census inventory. The Census Bureau developed an automated process that identified housing units with a relatively high likelihood of being a duplicate based on a set of person matching and address matching rules. The research focused on the ability of the person matching to identify duplicate housing units. The process was conducted in two phases.

The first phase involved developing algorithms to identify census IDs that were likely to be duplicates. These algorithms were developed in two approaches. The first approach identified pairs or clusters of census IDs that were likely duplicates because the addresses were the same or substantially equivalent. These algorithms were run on the Master Address File. The second approach identified pairs or clusters of census IDs that were likely duplicates because the households (using person data) were the same or substantially equivalent.

The second phase of the Housing Unit Unduplication Operations identified which of the approximately two million census IDs flagged for potential deletion in phase one would be reinstated, and thus included in the final census counts.

After substantial research, rules were developed to classify potentially duplicated census IDs as either reinstated or deleted. The result from applying the rules was to reinstate approximately one million census IDs (42 percent) and to delete approximately 1.4 million census IDs (58 percent).

Person Duplication in Census 2000

The "Further Study of Person Duplication" (Mule, 2002) was used as key input to estimate and identify duplication in order to make corrections to the original Accuracy and Coverage Evaluation (A.C.E.) and, subsequently, produce the A.C.E. Revision II estimates. Using a computer matching algorithm, the study performed a national match of A.C.E. Enumeration and Population sample records to census enumerations on the Hundred Percent Census Unedited file. The study used matching and modeling techniques to identify links between the full Enumeration and Population samples of the Accuracy and Coverage Evaluation to census enumerations outside the search area. The algorithm used a statistical matching methodology that assigned a probability of linked records being a match. Links with probabilities above specified thresholds were considered duplicates. This was a computer-only effort.

The study used Big Match software for the national statistical match. Big Match software allows matching a moderate size file of 100 million records against large administrative lists having upwards of four billion records.

This analysis was done based on evaluations of the March 2001 Accuracy and Coverage Evaluation coverage estimates, which indicated the Accuracy and Coverage Evaluation failed to detect a large number of erroneous enumerations. One type of these erroneous census enumerations were duplicates, which were census enumerations included in the census two or more times. The Accuracy and Coverage Evaluation was not specifically designed to detect duplicate census enumerations beyond the Accuracy and Coverage Evaluation's search area.

3. Research Plans Gearing Toward Census 2010

Plans for Census 2010 are well underway and research is ongoing into how to identify duplication in the census and how to deal with it. An unduplication work group was formed with a mission to research and gain knowledge about various aspects of identifying and removing duplication in future censuses. The group has produced proposals that address duplication of housing units, persons, and improvements in computer matching research. Our first attempt testing unduplication strategies during real time will be for the Coverage Research Followup operation to be employed in the 2004 Census Test.

The goal of the Coverage Research Followup is to gather information regarding the presentation of the residence rules instructions, the usefulness of the coverage questions, and the types of possible duplicates that should be contacted in order to resolve residency status (Urrutia, 2003). In the 2004 Census Test we will not actually remove any duplicates from the test census. Ultimately, identifying duplicated persons and removing them from the census in real-time will reduce the frequency of erroneous enumerations and improve coverage of housing units and persons.

The coverage questions are aimed at identifying potential people missed (undercount) after the size of the household has been determined and/or identifying potential people counted in error (overcount) after the characteristics of the household have been collected. The 2004 Census Test will help determine the best wording of the residence rules instructions and coverage questions for the 2010 Census. The follow-up interview of potentially overcounted people will attempt to identify why respondents included the people who were not census defined members of the household. The follow-up interview of potentially undercounted people will also attempt to identify why respondents did not include one or more members of the household. There will be follow-up interviews conducted by telephone and/or personal visit.

Research from Census 2000 showed that duplicates resulted from many situations. For example, duplicate

households could result from form misdeliveries and/or apartment mix-ups. Other times respondents had moved and were enumerated in both locations, their old and new residence. In addition, there were a number of duplicate addresses on the Master Address File because addresses resulted from numerous listing operations, and exact matching was used to compare the addresses.

For the 2004 Census test, potential person duplicates will be identified by computer matching and follow-up interviews will be conducted to resolve the potential duplicates. The unduplication operation will first computer match census data capture files against themselves to find potential duplicates based on demographics. This process is very similar to what was done as an evaluation in 2000 but it will be done in "real time" during and immediately after census forms are data-captured. The potential duplicates will then be followed up in the field by telephone and/or personal visit interview. There are two types of duplicates: whole household duplicates and partial household duplicates. Whole household duplicates will be followed up by a personal visit interview. A whole household duplicate occurs when all persons at the household are duplicated. Partial household duplicates will be followed up by a telephone interview and in some cases may be sent to the field if the case was not resolved by phone. A partial household duplicate occurs when at least one person, but not all persons in the household are duplicated. The goal of both types of follow-up interviews is to gather additional information to resolve the status of the potential duplicate and to ensure the person or persons in the household are accounted for once at one location.

Another way of approaching duplication in the census is to address the issue before the mail out of census forms. This can be accomplished by eliminating duplicate housing units from the mailing list. The Census Bureau has a long history of using computer matching methods to accomplish this, but there is a need to improve these methods by trying different approaches to this matching effort. For the 2004 Census test, research for identifying housing unit duplication will be done by comparing exact and probabilistic computer matching results. This evaluation addresses whether duplication can be reduced at the time of the initial Master Address File extract and subsequent census operations by using improved address linkage methods. This evaluation will use four different unduplication methods, comparing the results of each, and reconciling the potential duplicate addresses through field work during the 2004 Census test.

Other unduplication research includes:

Research has been proposed that will use administrative records as an additional source of evaluation data to help improve duplicate detection and resolution methods. Duplicates, once found, must be assessed as to which record to "keep". In the 2004 census test we are making this determination by a field follow-up operation. However, other options exist: Various criteria could be used from administrative records including completeness, recency, source of record, geographic location, etc. In this proposal we research how to determine the best method for selecting "keep" records, including the possibility that administrative records might be used to help in the selection. Two subprojects are proposed: A proposed Post 2004 Census test research to extract census person records and determine whether the administrative record person and geographic information can be used as matching fields in subsequent unduplication matching operations, and to determine when a record outside the block is to be declared a "link" with the census captured record. This research project will allow the Census Bureau to calibrate and improve unduplication methods for future tests. Another proposed project will assess the ability of Administrative Record data to distinguish between accurate and inaccurate enumeration status when the same person has been captured outside the housing unit's immediate geographic area.

Research will be undertaken to explore a new set of commercially available computer matching tools for record matching and duplicate detection. The research will test unduplication software for preprocessing of record linkage person data, compare the new software and methods with existing software and methods, and examine the potential for embedding a person unduplication step into 2010 processing.

4. Conclusion

In summary, the Census Bureau has begun a multifaceted program to attack the duplication problem realized in Census 2000. After the census uncovered the extent of this problem, several evaluations were conducted to help us understand the causes of duplicated persons and housing units. As a result, an Unduplication Work Group was chartered to recommend a research plan that will identify and reduce (eliminate) duplication in the census.

The research plan includes a real time unduplication effort to use computer matching and a field follow-up operation in the 2004 Census Test to find duplicate persons and ascertain their census day residence

during census processing.

Additionally, research will be undertaken to keep duplicate housing units off of the census address list for the 2004 census test. That is, we will be comparing two forms of computer matching efforts used to unduplicate housing units included on the census list. Improvement of our address lists is a key to our unduplication efforts.

Lastly, administrative records provide promise as an efficient means of determining duplicates and deciding which of them should be kept in the census and which should be discarded. We have started research to see if any of these techniques are appropriate for census efforts.

References

Hogan, Howard (2000), "Specifications for Eliminating Duplicate Records on the Hundred Percent Census Unedited File," Decennial Statistical Studies Division Census 2000 Procedures and Operations Memorandum Series # D-10, U.S. Census Bureau, Washington, D.C., November 7, 2000.

Killion, Ruth Ann (2003), "Study Plan for Comparing Exact and Probabilistic Matching Results," Draft May 2, 2003.

Mule, Thomas (2001), "Person Duplication in Census 2000," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report 20, U.S. Census Bureau, Washington, D.C., October 11, 2001.
Mule, Thomas (2002), "Accuracy and Coverage Evaluation Revision II Results: Further Study of Person Duplication," Decennial Statistical Studies Division Accuracy and Coverage Evaluation Revision II Memorandum Series #PP-51, December 31, 2002.

Nash, Fay F. (2000), "Overview of the Duplicate Housing Unit Operations," November 7, 2000.

Urrutia, Maria (2003), "2004 Census Test Project Plan for the Coverage Research Followup," Draft March 19, 2003.

Whitford, David (2002), "Proposals and Recommendations for Unduplication Research for 2010 Census and 2004 Census Test," September 24, 2002.