Unduplication of Persons and Housing Units in the 2004 Census Test

Robin A. Pennington
Decennial Statistical Studies Division, U.S. Census Bureau, Washington, D.C. 20233

Abstract

The capture of respondent names in Census 2000 enabled for the first time a real-time assessment of duplication within the census. After preliminary analysis suggested the housing unit count on the production files was too high and duplication was the likely cause, an ad hoc operation was mounted to research and eliminate certain categories of duplicate housing units from final Census 2000 counts. The approximately 1.4 million housing units deleted as a result of the unduplication operation made this one of the largest coverage improvement activities in the census. Subsequent evaluation revealed the need for a comprehensive research effort to improve the unduplication operation for the 2010 census. This paper discusses the research undertaken in the 2004 census test, as well as plans for future testing, in order to have an integrated unduplication process in place by 2010. This process will consist of automation and followup fieldwork on cases of potential duplication, as identified by respondent data collected in the census.

Keywords: coverage improvement

1. Background: Census 2000 Unduplication

Analysis in advance of Census 2000 indicated that duplication of units on the housing unit list was contributing to an overcount of addresses, which naturally was leading to an overcount of people, as well. An unduplication operation conducted while final census results were being processed led to the elimination of over a million housing units that were considered to be housing unit duplicates. Over two million people were listed in those housing units but eliminated from final census counts. A confounding factor in any unduplication effort is census form misdeliveries, a subset of which is apartment mix-ups. In such cases, questionnaires are not delivered to the exact addresses given on the form, either because apartment designations are unclear or mail is not delivered to individual units. In these situations, a questionnaire may be returned for one unit that has the household roster for what is, in fact, a different unit. If the other unit's questionnaire is not returned, it will be included in the Nonresponse Followup (NRFU) workload. In the absence of intervention to resolve the address mix-up, the people who sent in the original questionnaire will be re-enumerated during NRFU in the correct apartment. Meanwhile, the household at the other address has never been enumerated. In this way, form misdelivery results in duplicated people, but these are duplicates that should not be summarily deleted from census counts because both housing units exist.

For the operation in Census 2000, a file of potential duplicates was created by combining the results of address-matching and census-enumerated person-matching. Decisions about which of the potential duplicate housing units to keep in the census and which to remove from final counts were based on programming logic to determine which cases represented form misdelivery situations as opposed to housing unit duplications. Some of the paths in this decision process were shown in informal post-operation analysis to be error-prone.

2. Plans for the 2010 Census

The Census Bureau believes that it will be necessary to perform unduplication again in the 2010 census but is planning a radically different design from what was implemented in Census 2000. The plan for 2010 is to have an operation that works in tandem with census production operations and is focused on all aspects of coverage improvement. This is an expansion of the coverage efforts that were used in the past. In Census 2000, for example, households with more than six people as indicated by the person count were telephoned in order to get data for the remaining persons in the household. The Coverage Followup (CFU) operation planned for 2010 will include those cases, as well as other coverage cases concerned with residency determination and duplicates.

Potential duplicates will be identified in the course of regular census processing. All people identified as duplicates in person-matching of census enumerations will be designated as

eligible for the operation. Thus, person-matching is expected to yield a file of linked people who represent all of the following situations – housing unit duplication, form misdelivery, movers, children in shared custody counted by both caretakers, college students enumerated at their college address and at their parents', households where friends or family reside part of the time, and people in the following situations: with vacation homes, with separate houses maintained for their jobs, listed in error at an address (such as an apartment building superintendent), and in group quarters such as nursing homes or prisons. The housing unit records containing the linked people will be eligible for selection for the operation.

3. The 2004 Census Test Unduplication

The 2004 Census Test Coverage Research Followup (CRFU) was a first attempt at operationalizing this complex coverage activity. The CRFU universe encompassed unduplication cases as well as other coverage cases, with some cases in both universes. Related papers give operational results and analysis of the coverage cases.

The CRFU questionnaire was designed for followup of both the coverage and the unduplication cases. With respect to resolving person duplicates, Title 13 of the United States Code prohibits us from being able to ask respondents about people listed on another form. Another problem with operationalizing an actual resolution of duplicated people is that many of the duplicates will not be within a short distance of one another, so we will not be able to have the same interviewer ask about both cases.

Our approach to resolving all the cases included in CRFU was a thorough re-enumeration of both households, with additional steps implemented to resolve duplicates. There were sections in which housing unit issues could be noted, questions in which alternative addresses were collected, and finally major sections for doing a complete, dependent re-enumeration of the household. To the extent that it was possible to pair cases together for the followup, the pairs were clipped and worked together. This reduced respondent burden, as well as resulted in some efficiencies in the field when the housing units were duplicates.

Followup work was completed both with telephone interviewing and with fieldwork, so it was necessary to determine which cases should be sent where for followup. The philosophy for determining which unduplication cases were sent for a telephone interview and which were sent directly to field was based on the belief that housing unit problems are better resolved by assessing ground truth in the field, while a telephone interview and re-enumeration is sufficient for the resolution of person-level problems. Whole household duplication was assumed to be associated with housing unit problems. Therefore, cases with whole household to whole household or with whole household to partial household duplication were sent directly to the field, while partial household to partial household cases were sent to telephone resolution.

Whole household to partial household matches are cases where some persons included on one form are missing from another. Possible explanations for this occurrence are that the housing unit was duplicated or there was a form misdelivery and, in conjunction, somebody was left off one of the enumerations. Another explanation is that the collected data on a person on one of the forms was not of high enough quality for matching. While it is also possible that such a situation could result from movers or other person-level problems, it was believed that most of these cases would be housing-level problems and should be sent to the field for resolution. Analysis of 2004 data has an objective of further classifying sets of these cases that can be presumed to be housing-level problems only.

For the 2004 research stage, there was a clerical review after the followup operation, which was applied to all unduplication cases for a two-fold outcome. The primary outcome was an indication of the source of the duplication, taking into account any housing unit assessment outcome assigned by the interviewer, alternative addresses collected and all the enumeration data. The second part of the outcome was an indicator of whether the duplicate was resolved by the interviews; that is, whether a unique household could be determined for the duplicated person(s).

We do not expect to have time to follow up on every case identified as a duplicate in the 2010 decennial census. There will not be a clerical review to resolve cases in 2010, either. We are

assessing how to minimize the follow up and the workload, as well as how to automate certain paths of the process.

It should be noted that this back-end operation to fix identified duplicates is not the only effort by the Census Bureau that impacts duplication. There is a monumental effort to align all streets in the Bureau's mapping database with Geographic Positioning System (GPS) coordinates. With this map improvement, it becomes possible to append GPS coordinates to housing unit records during field operations. Additionally, address list development operations in preparation for the 2010 decennial census should be sequential and dependent, which was planned but not accomplished in Census 2000. Address-matching algorithms may be more refined than those used in Census 2000. We anticipate lower levels of housing unit duplication as a result of these improvements. The Census Bureau is also exploring means of reducing form misdelivery situations in areas where apartment designations are likely to lead to delivery problems. However, as society becomes more complex, we can only hope that questionnaire design and residence rule refinements ameliorate the number of person-level duplications.

4. Limitations

There are a number of limitations emanating from the fact that the 2004 Census Test was a site test, and, in addition, many of the usual census operations were not performed. For example, we expect relatively few cases of person-level duplication within a test site. Also, there was no enumeration of people who live at something other than a housing unit, such as in a nursing home in the 2004 Census Test. Thus, we should not look to these test results for any kind of indicator of counts of duplicates or of relative proportions of particular duplication problems.

Another limitation of a site test is that person-level duplicates that exist within a site may be complex residency situations. The two housing units where the duplicated person was counted are within a short distance of one another, and residency may be erratic and inconsistent.

The complexity of the CRFU operation and questionnaire is a limitation. Anecdotal data showed there to be some difficulty on the part of the interviewers when mixing housing unit and person enumeration concepts in the same operation. In addition, variations from stated procedures that may make sense to the field representative trying to get an interview could result in the wrong path being followed on the questionnaire. One means of skirting some of the difficulties is by designing a separate operation for the sole purpose of determining the correct number of housing units for those cases that require housing-level resolution rather than re-enumeration. Such a design would require an *a priori* determination of which units are housing-level problems, as opposed to person-level.

One additional complication with matching records is that sometimes links can bring together more than two housing unit records. For example, if one of the forms involved in a form misdelivery situation also contains duplicated children who are in shared custody, the result of the matching would be three units linked through the unit in common. These situations cannot be resolved in exactly the same way as two linked records. We were unable to include these in the 2004 operation because there was not a way to operationalize resolving the duplicates, as opposed to re-enumerating the household. However, for research purposes, some cases were selected for a clerical followup, which yielded mostly qualitative results. We believe certain categories of these could be worked in CFU as it is currently planned for the future. These results are not presented here.

We note that there was an additional test of probabilistic address matching programs to identify housing unit duplicates in the 2004 Census Test, the Record Linkage Followup (RLFU). To the extent that there were housing unit duplicates in the test sites, we would expect these to show up in both the address-matching and person-matching universes. For the purposes of minimizing respondent burden, such cases were selected for the RLFU operation. The results were used in the assessment of unduplication cases, as well. The RLFU results will not be presented here. We note only that the number of cases selected for unduplication followup understates the number of cases identified as duplicates by person-matching as a result of selecting these units for RLFU. Operational results for unduplication cases will understate the count of housing unit duplicates. In addition, to the extent that there were site differences in proportions of duplicates identified by probabilistic address matching, the distinction

will not appear in, and in fact could be obscured by, presentation of only our results.

5. Results

Within the operation and analysis, it is necessary to examine both the housing-level and the person-level data. Fieldwork is accounted for at the housing unit record-level, but outcomes were assigned in clerical review to all people linked as duplicates.

Across the waves and coverage operations, there were 7759 housing unit records selected for the unduplication operation. This count includes those cases that were selected both as a coverage case and as an unduplication case but does not include those cases that were selected by RLFU. The Georgia test site included 51,250 housing units, and the Queens, New York site included 151,239 housing units. Thus, the unduplication operation identified about 4.8 percent of records in the Georgia site and about 3.5 percent of records in the Queens, NY site.

In table 1, we have the breakdown of the initial mode selected for unduplication cases, according to the operational system output. We see that almost 85 percent of unduplication cases were initially selected for field followup. That is because we selected cases with whole household duplication for field followup, and person-level duplication is less likely to occur in a test site.

In table 2, we show the data from the clerical review stage, which yielded outcomes for cause of the duplication and whether a unique location for the duplicated person could be determined from the interviews. These outcomes were critical for the 2004 research to determine how to automate unduplication using the interview data.

Table 1: Initial Mode for Housing Unit Records

Initial Mode	Total	IM % of workload
Phone	1191	15.3
Field	6568	84.7
Total	7759	100.0

Source: Operational Output File

Table 2: Clerical Review Codes for Linked Persons

Clerical Review Codes	Person Counts	# Resolved	% Resolved
HU duplicates	2785		
Misdelivery	4644	4286	92.3
Shared custody	145	54	37.2
Movers	219	147	67.1
Group Quarters	2	0	0.0
Friends/relatives	27	21	77.8
Student	6	3	50.0
Vacation home	9	5	55.6
Work residence	8	7	87.5
Listed in error	518	476	91.9
Uncoded	2533		
No duplication	59		
Total person links	10955		

Source: Clerical Output File

One should look at this data for the percent of types of cases resolved, rather than relative percents of outcomes, due to site test limitations. Cases of form misdelivery and people listed in error have high rates of resolution from the followup. Movers and children in joint custody have lower rates of resolution from this followup. The numbers of people with vacation homes, people staying with friends and family, people with second homes for work, people in group quarters, and college students are too low to use for this assessment. While many of these rates could potentially be improved, shared custody stands out as an example of a situation where even an extended followup does not often yield resolution, at least in a test site. Certainly these cases are of paramount concern.

The determination of which housing unit to keep when there are housing unit duplicates will likely involve geographic considerations. This component was not in scope for the 2004 research. Also note that some of the linked people were found to be not true duplicates. The threshold for determining a potential link to be a duplicate at the point of creation of the file was set quite high; nevertheless, some cases that were not actual duplicates were selected for followup.

The category of uncoded cases appears here as a large piece of the workload. We were careful not to overstate the rate at which cases could have been resolved with automated scoring of the questionnaires. Information had to come from the questionnaires, not from inspection of uniqueness of the names of the linked people or any other consideration that could be gleaned only through human intervention. Additional research needs to be completed on these categories to determine if there are changes to the questionnaire or to training or procedures that could result in a higher rate of codable cases.

There were five selected cases that crossed site boundaries but that clerical review showed as not being duplicates or as uncoded. These were assigned a lower matching threshold but were selected for the operation in order to test cross-state unduplication processing and followup. Research on matching continues.

A second path of analysis that has major operational implications is how to designate cases for field or telephone followup. We would like to know *a priori* which cases require a housing-level determination. For this research, we tallied outcomes by level of household duplication. The results on this were promising. Our categories of outcomes here are Housing unit Duplicates (HU resolved – HU dup), form misdeliveries (HU resolved – form misdelivery), person-level duplicates (P resolved), uncoded cases, and cases that were determined not to be actual duplicates. As shown in table 3, there is a decided shift away from housing-level outcomes when we move from whole-whole (WW) household duplication, through whole-partial (WP) household, to partial-partial (PP) household duplication. There is still a high rate of housing-level outcomes within PP matches, however, at least in the New York site. This may be because we are restricted to a test site, or there may be additional criteria that distinguish the housing-level cases from those at the person-level.

Table 3: Persons by Outcome, State, and Level of Household Match

Level of Duplication by outcome	**GA column %**			**NY column %**		
	WW	**WP**	**PP**	**WW**	**WP**	**PP**
HU resolved - HU Dup	39.4	24.6	4.5	24.6	23.6	16.1
HU Resolved – Form Misdelivery	28.4	20.4	7.5	48.6	37.6	30.8
P Resolved	4.8	17.9	15.6	3.7	12.5	7.7
Uncoded	27.4	36.2	71.0	23.1	26.0	41.1
No Duplication	0.0	0.9	1.4	0.0	0.3	4.3
Number of People	3826	804	887	11,142	3238	2013

Source: Unduplication Evaluation File

The variation across sites in percentages of the different outcomes by type of household match isparticularly interesting. We see that the New York site had a much larger percent of duplicates resulting from form misdelivery than the Georgia site did. On the other hand, the New York site has a lower rate of housing unit duplicates, which is a possible effect of having the address list updated in advance of the delivery of questionnaires in the New York site but not the Georgia site. The higher percentage of person-level duplicates in the Georgia site could be expected given its greater geographic area. In any event, the level of household duplication shows itself to be connected to types of outcomes for both sites.

Further analysis will include tabulations with different variables that we hope can be used for *a priori* determination of which cases can be sent to a housing unit-level operation, as distinct from those sent for a re-enumeration. For example, logic dictates that links within the same block are probably housing unit-level problems, and preliminary analysis validates this theory. We will also attempt to assign priorities for cases for fieldwork, as we expect to be constrained in 2010 to resolving fewer cases than we can identify.

In conclusion, we see both promising and disappointing results from our work on the 2004 CRFU. Whole to whole household matches were found to result from housing unit-level problems at a very high rate, while person-level problems appeared more frequently in partial to partial household matches. While this is how we selected cases for the different modes of followup in 2004, results from the test were reassuring. We have also seen that the questionnaire that was used in 2004 was successful at determining a unique household for duplicated people most of the time in the situations in which the cause of the duplication was form misdelivery or a person listed in error. However, shared custody cases have proven difficult to resolve, even with an extensive reinterview and a series of questions aimed at determining where people spend most of their time. It is possible that other situations also are unlikely to be resolved in such a followup, although the test site did not provide enough cases to test.

We find much in these results to use for planning of future person and housing unit unduplication, as well as confirmation that the problem of duplication is difficult to solve. The unduplication operation is just the last step in a series of measures that we believe will reduce duplication from the outset. We endeavor to plan efficient means of resolving those duplicates that occur despite our other efforts to eliminate them.

Acknowledgements

The author wishes to thank Leah Marshall for her contributions to this research, analysis, and presentation.

References

Fay, Robert E., “Probabilistic Models for Detecting Census Duplication at the Person and Household Levels”, 2003 Proceedings of the American Statistical Association, Survey Research Methods [CD-ROM], Alexandria, VA

Pennington, R.A., Vitrano, F., “A Final Assessment of the Census 2000 Master Address File”, 2003 Proceedings of the American Statistical Association, Survey Research Methods [CD-ROM], Alexandria, VA

Knight, L., Behler, J., Vitrano, F., “Operational Assessment of the 2004 Coverage Research Followup”, 2005 Proceedings of the American Statistical Association, *to appear*

Krejsa, E., “Results of the Coverage Research Followup in the 2004 Census Test”, 2005 Proceedings of the American Statistical Association, *to appear*

Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on operational issues are those of the author and not necessarily those of the U.S. Census Bureau.