Person Duplication in Census 2000

Vincent Thomas Mule, Jr., Bureau of the Census, Washington DC 20233

Key words: record linkage, census coverage

Abstract:

This paper shows estimates of the amount of person duplication in Census 2000. These estimates were done for the October 2001 decision of the use of census data for non-redistricting purposes. We were concerned that perhaps the estimate of erroneous enumerations in the 2000 Accuracy and Coverage Evaluation (A.C.E.) was too low because the estimate of duplicate enumerations as measured by the A.C.E. was less than the estimate from the 1990 Post-Enumeration Survey (PES). Our matching work attempted to identify duplicate enumerations across the United States. As a benchmark ,we were able to compare our results to the A.C.E. results for the same geographic search areas. This work identified the amount of duplicate enumerations 1) outside of the geographic search area of A.C.E. and 2) between housing units and group quarters.

1. Introduction

We were concerned that perhaps the estimate of erroneous enumerations in the 2000 Accuracy and Coverage Evaluation (A.C.E.) was too low because the estimate of duplicate enumerations as measured by the A.C.E. was fewer than the estimate from the 1990 Post-Enumeration Survey (PES).

To estimate net coverage error, a coverage study needs to estimate the number of erroneous enumerations. One category of erroneous enumeration is persons duplicated in the census. The PES estimated more erroneous enumerations than the A.C.E. The PES estimated that 1.6 percent of the enumerations were duplicates (Hogan 1993). This is approximately 3.97 million duplicate enumerations (Childers 2001a). The A.C.E. estimated that 0.8 percent of the enumerations were duplicates. This is approximately 2 million duplicate enumerations (Feldpausch 2001a).

The PES estimated coverage for persons in housing units and non-institutional group quarters. Persons living in institutions, military personnel living in barracks or on ships and people living in homeless shelters were excluded in 1990 (Hogan 1993). The A.C.E. estimated coverage for persons in housing units. A.C.E. did not estimate coverage of persons in group quarters (Childers 2001b).

All of the enumerations in Census 2000 were not eligible for the A.C.E. For the United States, the Census Duplicate Housing Unit operation excluded 5.9 million person records from the Census. This operation later reinstated 2.3 million of these person records in the final census count. However, none of the reinstated or excluded records were part of the A.C.E. Hogan (2001) showed that the exclusion of this universe would not bias the estimate of the Dual System Estimate if the number of matches is reduced proportionately to the number of census correct enumerations. However, this could produce a lower estimate of erroneous enumerations, overall, and in particular duplicate enumerations.

The Census Duplicate Housing Unit operation initially identified housing units suspected as being included in error with a relatively high likelihood based on a set of person matching and address matching rules. Research focused on the ability of the person matching to identify duplicate housing units, rather than the duplicate person records serving as substitutions for other households. Algorithms were established for identifying instances where a duplicate household was more likely than not to reflect a substituted enumeration, rather than a duplication of housing units (Nash 2000). These cases were among the 2.3 million person records reinstated in the census count. If these cases had been available for matching, the A.C.E. potentially may have estimated these “substituted” enumerations as duplicate enumerations if they occurred within the search area.

The search area for duplicates in the 1990 PES was the block cluster and the ring(s) of blocks surrounding the cluster. For all non-matches or erroneous enumerations, the PES searched one or two rings of surrounding blocks depending on the type of geography. Also, the PES rematched persons in some clusters with high numbers of non-matches or erroneous enumerations. The PES extended the search area beyond two rings for some of these clusters.

The search area for the A.C.E. was primarily the block

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and encourage discussion.

cluster. Targeted Extended Search expanded the search area for a sample of units by one ring of surrounding blocks for certain cases believed to be geocoding error.

Our analysis classifies person records into the following categories based on the following types of units:

Table 1: Categories of Units in this Analysis

Category	Description
E-sample Eligible[1]	Persons enumerated in housing units that were eligible to be selected for the Enumeration sample (E sample) for the Accuracy and Coverage Evaluation.
Reinstated	Persons enumerated in housing units suspected to be potential duplicates by the Census Duplicate Housing Unit process. These housing units were ineligible for the E sample and the A.C.E. matching. The Duplicate Housing Unit process examined these cases and reinstated them into the census count.
Group Quarters	Persons enumerated in group quarters
Deleted	Persons enumerated in housing units suspected to be potential duplicates by the Census Duplicate Housing Unit process. These housing units were ineligible for the E sample and the A.C.E. matching. The Duplicate Housing Unit process examined these cases and did not include these in the census count.

[1] Does not include Remote Alaska

2. METHODS

This paper focuses on matching census person records to determine estimates of person duplication. We implemented three steps in this analysis:

- created files for computer matching
- conducted two stages of computer matching
- produced estimates of person duplication

2.1 Matching Files

We created the **Source** and the **Target** files:

- The **Source file** contained the data-defined persons in E-sample eligible and reinstated housing units in the **11,303 A.C.E. sample block clusters**.
- The Target file contained the data-defined records in 1) housing units and group quarters in the census enumeration and 2) housing units deleted from the census by the Census Duplicate Housing Unit operation. The **Target File** contained **all of these records from the entire nation.**

2.2 Stages of Computer Matching

We implemented **two stages** of computer matching. Our approach used an exact matching procedure during the first stage. This stringent approach would require records to have the same values for specified characteristics to be linked together as potential duplicates.

The second stage built on the results of the first stage. By matching persons in the first stage, we identified person duplication between two units. For the second stage, we statistically matched the persons in just these two units by using the Survey Research Division matcher. The statistical matching compares the agreement of several characteristics. We determined that two records were duplicates based on the overall agreement of those characteristics.

Because of the time constraints for this project, we were unable to clerically review the duplicate links identified by the computer matching.

2.2.1 First-Stage Matching

We used an exact matching approach to link duplicate records. We compared **each record on the Source file** to **every record on the Target file**.

For this exact matching, we required agreement of **all** of the following variables:

- First Name
- Last Name
- Month of Birth
- Day of Birth

To be eligible for first-stage matching, **we required each record on both files to have non-blank values for all four fields**.

While we required exact correspondence for the characteristics, we did add the following enhancements to improve the matching:

- Flip-flopped the first and last name during matching. This allowed "John Jones" to link to "Jones John".
- Removed "Jr","Sr" and "III" from the first and last name fields.
- Checked to see if the middle initial was scanned into the first or last name field. This allowed us to link "Mary L. Smith" with "Mary Smithl" or "Maryl Smith".
- Required computed age to be within one year if reported by both records.

2.2.2 Second-Stage Matching

For the second stage, we used statistically-based matching with the Fellegi-Sunter algorithm as implemented by the Statistical Research Division at the Census Bureau**.** The strength of this approach is that it allowed us to link "Timothy" and "Tim" together. We are also able to account for data capture errors ("Steve" can be linked with "Steue"). One concern is that statistically-based matching has the potential for yielding substantially more incorrect matches than exact matching if it is applied widely. Our process of requiring an exact match during the first stage between the units minimizes this potential.

We examined the agreement of the following characteristics**:**

- First Name
- Middle Initial
- Last Name
- Month of Birth
- Day of Birth
- Computed Age (which accounts for the reporting of year of birth and age fields)

2.3 Producing Estimates of Duplication

The matching files contained only the information needed to link records from the Source file to records on the Target file as duplicates. The analysis files contained each link of a Source person record to a Target person record. We appended the person, unit and block characteristics to the Source and Target person record of each link. Also, we assigned the A.C.E. sampling weights so weighted estimates of person duplication could be generated. For each link, we assigned sampling weights and duplication factors.

For variance estimates, we used a simple jackknife methodology on the final A.C.E. cluster design. These variance estimates should be slight underestimates of the variances if they reflected the full A.C.E. two-phase cluster sampling plan. Since all of the person records in E-sample eligible or the reinstated housing units in a cluster are on the Source file, we used the cluster-level weight of the Source person.

We assigned **two factors to each link**. The first factor was an unbiased probability of duplication or multiplicity factor for the link. The second factor was a model weight which expresses the confidence in the link representing true duplication.

We assigned a **model weight** to each link in **three parts**. For the first part, we determined how many duplicate links were identified between the two units. The more links we establsihed between units, the more confident we were in the links.

We determined **two sets of links** where we were **confident in the links** because of the multiple links between the units. We assigned a model weight of 1 to these cases.

- All persons in the housing unit on the Source file link to the same housing unit on the Target file.
- Two or more persons in the housing unit on the Source file link to the same housing unit on the Target file within the same state.

We determined **two sets of links** that we **removed from the analysis.** These links were identified by the **second-stage matching (statistical matching)**.

- person links from housing units to group quarters. The statistical matching created too many false matches between relatives in the housing unit to other occupants of the group quarters. Example: "Margaret Brown's" sister Melanie was matched to Melanie Smith in the group quarters.
- person links between housing units in different states where the entire household was not duplicated. We were concerned about false matches when the geographic distance increased. We used state boundaries as a proxy for geographic distance.

Note: This first part assigned all of the second-stage links. The next two parts of the modeling apply to the remaining first-stage links.

For the second part, our processing identified the following instances where we believe the link does not represent duplication in the census.

- For links outside the cluster, the Source and Target reported different middle initials or computed ages. We allowed these links to be created in the first-stage matching to attempt to find additional links during the second-stage matching. Since we were unable to find additional links during the second stage, we removed links that had conflicting middle initials or where the computed ages differed by one year.
- Duplicate links between "Jane Doe's" and "John Doe's". These are fictitious enumerations or field imputations by the enumerator and not duplicates.
- Duplicate links with first names whose birth day is the feast day of their patron saint. We have anecdotal evidence that some people report the feast day of their patron saint as their date of birth. An example is a link between two persons named "Jose" who were born on March 19th. March 19th is the feast day of St Joseph
- Duplicate links between Nonresponse Follow-Up (NRFU) training examples. These links are fictitious enumerations and not duplicates

For the remaining links in the third part, we have exact matches on first name, last name, month of birth and day of birth. We used a Poisson distribution approach to account for the chance that these records were linked together because of common characteristics. Our model weight compared the actual number of days with two or more births to the expected value using a Poisson distribution.

3. LIMITATIONS

- This type of analysis has not been conducted nationally before; therefore we do not have data available for comparisons outside of the A.C.E. search areas.
- We only conducted automated matching due mostly to time constraints; there was no clerical matching or field work to resolve unknown matches. Likewise, a conservative automated matching algorithm was used to ensure that we can be confident in our identification of duplicates.
- All duplicates identified by A.C.E. were clerically identified. Clerks were able to use more characteristics and look at the scanned census forms to determine duplicates. Because of our approach, our estimate of E-sample to E-sample duplication within the cluster compared to the A.C.E. estimate will be a conservative underestimate of the duplication within this universe.

4. RESULTS

4.1 Comparing A.C.E. to PES

The A.C.E. measured fewer duplicate enumerations because of design differences between the A.C.E. and the PES. Table 2 shows the results of our duplication analysis within the cluster and surrounding blocks for various universes.

Table 2 Highlights:

- Our estimate of duplication for E-sample Eligible to E-sample Eligible within the cluster (724,687) was 37.8 percent of the duplication for this universe identified by A.C.E.
- We identified a small number of duplicates within the cluster that were identified by our matching but were not found by A.C.E. (41,046 of the 724,687). This is approximately 2 percent of the A.C.E. total estimate of duplication
- Our computer matching estimate of duplication for E-sample Eligible to Reinstated universe was very close to the clerical estimate of duplication for this universe from the Planning and Research Evaluation Division (PRED) evaluation of Reinstated persons (Raglin 2001).

Table 3 shows the A.C.E. estimate of duplication. A.C.E. searched for duplicates amongst the E-sample eligible to E-sample eligible universes. Table 4 shows the results of using a methodology more similar to the PES. This result is approximately 1.2 million higher than the A.C.E. estimate. These estimates extend the search area for all units to one ring of surrounding blocks. These estimates include searching for duplication to the reinstated housing units and group quarters. These housing units and the non-institutional group quarters would have been in-scope for the PES.

Table 2: Person Duplication Within Cluster and Surrounding Blocks

	Within Cluster		Surrounding Block	
Universe	Estimate	Standard Error	Estimate	Standard Error
E-sample Eligible to E-sample Eligible	724,687	30,145	146,880	9,683
E-sample Eligible to Reinstated	1,049,699	41,703	24,029	6,637
Reinstated to Reinstated	15,386	4,040	1,532	542
E-sample Eligible to Group Quarter	103,168	27,820	46,736	25,595
Reinstated to Group Quarters	95	95	0	0
E-sample Eligible to Deleted	1,941,732	78,312	682,909	44,690
Reinstated to Deleted	8,767	2,796	640	334

Table 3: A.C.E. Estimate of Person Duplication

Universe	Cluster and Surrounding Block Estimate
A.C.E. Estimate of E-sample Eligible to E-sample Eligible	2,014,675

Source: Feldpausch (2001a)

Table 4: Estimate of Person Duplication Using a Methodology Similar to the PES on 2000 Census

Universe	Cluster and Surrounding Block Estimate
A.C.E. Estimate plus E-sample Eligible to Reinstate, E-sample Eligible to GQs	3,238,307

Table 5 shows the results of using a methodology similar to the PES on the 2000 Census counts prior to the Duplicate Housing Unit operation. This result is approximately 3.8 million more duplicates than the A.C.E. estimate. This universe is not entirely comparable to the PES. Census 2000 used multiple sources of addresses when compiling the Master Address File (Nash 2000). These results show what the estimate of duplication would have been if the Duplicate Housing Unit operation was not done.

Table 5: Estimate of Person Duplication Using a Methodology Similar to the PES on a 2000 Census Count Prior to the Duplicate Housing Unit Operation

Universe	Cluster and Surrounding Block Estimate
A.C.E. Estimate plus E-sample Eligible to Reinstate, E-sample Eligible to GQs, E-sample Eligible to Deleted	5,862,916

4.2 Total Estimates of Duplication From Our Analysis

Table 6 shows the estimates of duplication from our analysis for various universes. This table presents total results and results for outside the surrounding blocks. The table has two sets of estimates for the Census housing unit to Census housing unit universe. The first set includes all duplicates (Total). The second set does not include duplicate links to reinstated units. The Duplicate Housing Unit operation developed algorithms for identifying instances where a duplicate household was more likely than not to reflect a substituted enumeration, rather than a duplication of housing units (Nash 2000). Because of this, we presented both sets of estimates.

5. Conclusions

The Executive Steering Committee on Accuracy and Coverage Evaluation Policy II (ESCAP II) asked us to do additional research on the 2000 Accuracy and Coverage Evaluation (A.C.E.) estimate of duplication

Table 6: Total Estimate of Person Duplication from Our Analysis

Universe	Estimate	Standard Error
Census Housing Units to Census Housing Units		
Total	4,625,019	77,941
Outside Surrounding Blocks	2,662,806	44,389
Not including duplicate links to reinstated units	2,960,675	47,786
Outside Surrounding Blocks	2,089,107	33,210
Census Housing Units to Group Quarters	660,189	65,119
Census Housing Units to Deleted Housing Units	2,911,016	95,665

An inter-divisional group conducted computer matching to determine the extent of duplicate census enumerations. This analysis of duplicates is limited to the extent that there was no clerical matching and that these results are generally conservative. We were concerned that perhaps the estimate of erroneous enumerations in the A.C.E. was too low because the estimate of duplicate enumerations as measured by the A.C.E. was less than the estimate from the 1990 Post-Enumeration Survey (PES). Our matching work identified duplicate enumerations that were outside of the scope of the A.C.E. This included duplicate enumerations identified outside of the geographic search area and enumerations in housing units and group quarters outside of the A.C.E. universe.

Our analysis found an additional 1.2 million duplicate enumerations in units that were out-of-scope for the A.C.E. but would have been in-scope for the PES. The A.C.E. estimate of duplication was different from the PES estimate because the two surveys searched for duplicate enumerations in different universes of units. Accounting for these differences produced an estimate of duplicate enumerations that was much closer to the PES estimate.

In summary, the A.C.E. measure of duplicate enumerations within the search area was less than the PES estimate primarily due to design differences; therefore, it is not a concern. This paper also shows that patterns of duplicate enumerations are intuitive and not unexpected. This paper does not say anything about how A.C.E. treated the duplicate enumerations found in this study. This is a subject of further analysis in Feldpausch (2001b).

References

Childers, D., "1990 E-Sample Documentation" Internal Census Bureau memorandum, Washington, D.C., 2001(a).

_____., "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series S-DT-1, U.S. Census Bureau, Washington, D.C., 2001(b).

Feldpausch, R., "E-sample Erroneous Enumeration Analysis," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report 5, U.S. Census Bureau, Washington, D.C., 2001(a).

_____., "Census Person Duplication and Corresponding A.C.E. Enumeration Status," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report 6, U.S. Census Bureau, Washington, D.C., 2001(b).

Hogan, H., "The 1990 Post-Enumeration Survey: Operations and Results" Journal of the American Statistical Association, September 1993, Volume 88 Number 423.

_____., "Accuracy and Coverage Evaluation Survey: Effect of Excluding 'Late Census Adds'," DSSD Census 2000 Procedures and Operations Memorandum Series Q-43, U.S. Census Bureau, Washington, D.C., 2001.

Nash, F., "Overview of the Duplicate Housing Unit Operations," Internal Census Bureau memorandum, Census 2000 Informational Memorandum Number 78, U.S. Census Bureau, Washington, D.C., 2000.

Mule, T.., "Person Duplication in Census 2000," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report 20, U.S. Census Bureau, Washington, D.C., 2001.

Raglin, D., "Effect of Excluding Reinstated Census People from the A.C.E. Person Process," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II Report Number 13, 2001.