Census and Administrative Records Duplication Study

Mary H. Mulry, Susanne L. Bean, D. Mark Bauder, Deborah Wagner, Thomas Mule, Rita J. Petroni[1]
U. S. Census Bureau, Washington, DC 20233
mary.h.mulry@census.gov susanne.l.bean@census.gov

KEYWORDS: Census 2000, undercount, overcount, Accuracy and Coverage Evaluation Survey

1. Introduction

The Census and Administrative Records Duplication Study (CARDS) used administrative records to examine the quality of the estimates of duplicate enumerations in Census 2000. The estimates of duplicate enumerations were incorporated in the revision of the estimates of coverage error in Census 2000 from the Accuracy and Coverage Evaluation (A.C.E.) Survey. These revised estimates, known as A.C.E. Revision II (U. S. Census Bureau 2003), showed a net undercount rate of -0.5 percent, an overcount, in the census count of 281,421,906. The estimates of duplicate enumerations used in the formulation of the A.C.E. Revision II estimates demonstrated that duplicate enumerations occurred in the census much more frequently than previously observed or suspected in Census 2000 or other censuses.

By the time a computerized search of the census provided evidence of a large number of duplicate census enumerations in October 2001(Thompson, Waite, Fay 2001), field tests for confirmation were not practical. However, the Statistical Administrative Records System (StARS), created with the Census Bureau's newly developed administrative records database methodology (Leggieri, Pistiner, Farber 2002, Judson 2000), provided the possibility of evaluating the estimates of duplicate enumerations without fieldwork. CARDS coincided with the preparation of the estimates of duplicate enumerations by the Further Study of Person Duplication (FSPD) (Mule 2002) for the A.C.E. Revision II. Generally, CARDS agreed with FSPD on the identification of duplicate enumerations in Census 2000. The estimate of the number of duplicates in the census using only the duplicates identified by administrative records was 6,653,171 while the FSPD methodology estimated 5,826,478. CARDS found more duplicates that were geographically distant and more group quarters duplicates while the FSPD process was better at finding duplicates that were geographically close.

In this paper, we describe the methodology for identifying duplicates in the census used by CARDS and FSPD. In addition, we compare the CARDS results with the estimates of duplicate enumerations from FSPD. More detailed results can be found in Bean and Bauder (2002).

2. Methodology

The A.C.E. Revision II estimation (U. S. Census Bureau 2003) used the dual system estimator with adjustments to account for duplicate census enumerations and other measurement errors that were detected by the Measurement Error Reinterview (Raglin and Kresja 2001) and the Matching Error Study (Bean 2001). A correction for correlation bias also was included. Two overlapping samples were used to produce the estimates, a sample of census enumerations (E-sample) for estimating erroneous enumerations and a sample of the population (P-sample) for estimating census omissions using the A.C.E. The two samples overlapped by using the same block clusters. The universe for the A.C.E. was people living in housing units and did not include those living in group quarters.

A.C.E. Revision II estimation used FSPD's computer match to estimate duplication of E-sample cases to enumerations outside the search area around the A.C.E. sample blocks where the A.C.E. matching operation considered enumerations correct. FSPD linked E- sample records to Census 2000 person records using the Hundred Percent Census Unedited File (HCUF). Although the processing for the original A.C.E. did not include people in housing units identified as

[1]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 to encourage discussion.

potential duplicates by the Housing Unit Duplication Operation (HUDO) (Miskura 2000), A.C.E. Revision II included links outside the A.C.E. search area to people in housing units reinstated in the census after HUDO and those deleted by HUDO. FSPD also estimated duplication within the A.C.E. search area, but A.C.E. Revision II used the A.C.E. clerical matching for these areas. CARDS used the results of a previous Census Bureau match between the HCUF and an administrative records file to estimate duplication in the census. Below are brief descriptions of the FSPD and CARDS linking processes. FSPD used similar methodology in matches for the P-sample to identify possible data collection error and obtained similar results. To save space, we restrict the discussion to the E-sample.

2.1 FSPD Linking

FSPD linking used two types of matching to create links and assign probabilities to those links. These types of matching are referred to as statistical matching and exact matching (Kostanich 2003). The statistical matching had two stages. The first stage was a statistical matching of source (E-sample) to target (census) records based on name (first name, last name, and middle initial) and age/date of birth (computed age, month of birth, and day of birth). After the first stage identified a person link between two housing units (HUs), the second stage performed a statistical match of people in those two HUs. The second stage matching was also based on name and age/date of birth, but used different parameters than those used in the first stage. For links in HUs with 2 or more links (2+ HUs), the statistical matching process assigned a Probability of No Trial Having Observed Outcome, called p. If the link had a probability p greater than a cutoff defined for the distance between the links, then it was considered a statistical duplicate and was assigned a final duplicate probability of 1.

The exact matching required agreement on first name, last name, month of birth, and day of birth among census records from HUs and from group quarters. When the exact matching linked two records and the statistical matching had not already assigned a final duplicate probability of 1, the process assigned a final duplicate probability between 0 and 1. The links with final duplicate probabilities assigned from exact matching were links whose probability p did not meet a statistical matching cutoff, links to group quarters, and links where only one person linked between the HUs.

2.2 CARDS Linking

CARDS linking had two basic steps. First, Protected Identification Keys (PIKs) were assigned to HCUF records, and thereby E-sample records, by matching census and A.C.E. files to administrative records in the StARS 2000 database. Then, links were created between records which were assigned the same PIK.

The StARS 2000 database incorporates data from seven administrative record files[2]: (1) Internal Revenue Service Individual Master File (1040), (2) IRS Information Returns File (W-2 / 1099), (3) Department of Housing and Urban Development Tenant Rental Assistance Certification System File, (4) Department of Housing and Urban Development's Multifamily Tenant Characteristics System File, (5) Center for Medicare and Medicaid Services Medicare Enrollment Database File, (6) Indian Health Services Patient Registration System File,(7) Selective Service System Registration File.

In addition, the Census Bureau created "Geokey Numident" from the Social Security Administration's Numerical Identification File (Numident)[2]. The Numident was edited, and for confidentiality reasons a PIK was created for each Social Security Number. Then a geokey variable was added to represent each address from the IRS 1040 and 1099 files from StARS 2000 for each person.

[2]The Census Bureau obtains administrative data for its StARS database as authorized by Title 13 U.S.C., section 6 and supported by provisions of the Privacy Act of 1974. Under Title 13, the Census Bureau is required to protect the confidentiality of all the information it receives directly from respondents or indirectly from administrative agencies and is permitted only to use that information for statistical purposes.

In work prior to CARDS, the Census Bureau had performed a two phase computer match to link Geokey Numident records with HCUF records in order to assign PIKs. In the Geokey Search phase, matching between the files was done based on name, date of birth, and geokey. Additional links were created in the Name Search phase where matching was based on name and date of birth only. Via this match, PIKs were found for HCUF people and added to HCUF person records. We called the resulting file the HCUF Research File, which served as the source of PIKs with E-Sample records. Note that some person records on the HCUF and thereby the E-sample file had no PIK assigned. This could happen in two ways. If the HCUF record was not linked with any PIK, none could be assigned. In addition, when one HCUF record was linked with more than one PIK (which is likely to have occurred when linking people with common names and characteristics), no PIK was assigned to the HCUF record. We believe that in many cases, this aspect of the CARDS process avoided linking different people whose person characteristics were similar. However, we do not know how many false links remained.

Links were created between source (E-sample) and target (census) records with the same PIK. The CARDS process did not assign probabilities, thus each link is considered a duplicate. We compared links identified by CARDS to those identified by FSPD to determine which links were found by both studies and which were only found by CARDS. If the source and target person had the same PIK and FSPD also identified the link, we classified the CARDS link as found by both CARDS and FSPD. If the source and target person had the same PIK but FSPD did not find the link, we called it a CARDS only link.

3. Results

To examine the quality of the estimates of duplicate enumerations from FSPD, we have computed estimates of duplication based on CARDS to compare to FSPD estimates. Standard errors were calculated using a simple jackknife method. We also looked at some characteristics of the CARDS links in an attempt to explain some differences between the estimates.

Table 1 shows weighted frequencies of CARDS E-Sample duplicate links by geographical categories and type of census record while Table 2 shows the same results from FSPD.

Table 1. CARDS Weighted Estimate of E-sample Duplicates by Geography and Census Record Type

Geography	Census Record Type				Total
	E-Sample Eligible	GQ	Reinstate	Delete	
Within Cluster	998,239 (35,162)	107,305 (21,452)	920,405 (42,888)	1,681,962 (82,499)	3,707,911 (113,548)
Surrounding Block	202,741 (15,516)	31,355 (11,686)	22,870 (5,926)	588,300 (48,878)	845,266 (55,656)
Same County	1,145,036 (24,177)	334,983 (47,946)	420,917 (24,624)	187,804 (18,520)	2,088,740 (64,559)
Diff. County, Same State	693,540 (20,531)	307,014 (13,610)	79,986 (10,708)	35,618 (6,734)	1,116,159 (29,646)
Different State	1,183,055 (30,328)	183,917 (10,500)	21,808 (3,276)	32,472 (4,350)	1,421,251 (34,133)
Total	4,222,611 (68,660)	964,574 (57,701)	1,465,986 (52,042)	2,526,156 (102,200)	9,179,326 (169,735)

Note: This table is weighted by the product of the A.C.E. sampling weight and the multiplicity factor (Mule 2002, Appendix D). Standard errors are in parentheses.

Table 2. FSPD Weighted Estimate of E-sample Duplicates by Geography and Census Record Type

Geography	Census Record Type				Total
	E-Sample Eligible	GQ	Reinstate	Delete	
Within Cluster	1,173,344 (46,173)	76,381 (15,736)	1,058,548 (48,295)	1,967,199 (94,454)	4,275,472 (129,245)
Surrounding Block	259,805 (21,718)	25,373 (9,701)	24,751 (6,971)	678,355 (57,469)	988,284 (65,896)
Same County	1,011,920 (24,292)	231,774 (39,795)	482,015 (27,797)	208,246 (20,789)	1,933,956 (59,590)
Diff. County, Same State	563,270 (18,873)	190,417 (9,488)	88,331 (12,567)	35,111 (7,262)	877,129 (26,615)
Different State	527,796 (23,744)	91,793 (7,093)	20,959 (17,316)*	16,184 (4,902)	656,732 (33,930)
Total	3,536,136 (68,045)	615,738 (46,003)	1,674,604 (60,317)	2,905,096 (116,541)	8,731,572 (177,071)

Note: This table is weighted by the product of the A.C.E. sampling weight, the multiplicity factor, and the final probability of duplication. Standard errors are in parentheses. *The high standard error is due to clustering. (Mule 2002)

The geographical categories are: (1) within the A.C.E. block cluster, (2) outside of the A.C.E. block cluster, but within surrounding blocks, (3) outside of surrounding blocks, but within same county, (4) outside of surrounding blocks and county, but within same state, and (5) outside of surrounding blocks, in a different state. The types of census record are: (1) E-Sample eligible enumerations in housing units, (2) enumerations in Group Quarters, (3) enumerations in housing units that HUDO reinstated, and (4) enumerations in housing units that HUDO deleted.

CARDS identified approximately 6.65 million census duplicates, of which about 4.2 million were between E-sample eligible census records. CARDS found about 2.5 million links between the E-sample and records deleted by HUDO. Within the cluster, CARDS found fewer than one million links between E-sample eligible records. Therefore, CARDS was not as efficient as the A.C.E. clerical person matchers who found about 1.9 million duplicates for this group (Mule 2002, p.7).

FSPD identified approximately 5.83 million census duplicates, which is approximately 0.82 million fewer than CARDS found. FSPD also found fewer duplicates between E-sample eligible census records than CARDS (3.5 million versus 4.2 million). However, FSPD found about 2.9 million links between the E-sample and records deleted by HUDO, which is more than CARDS found. Within the cluster, FSPD found about 1.2 million duplicates and was more efficient than CARDS but not as effective as the A.C.E. clerical person matching.

Two other differences stood out between the CARDS and FSPD E-sample results. CARDS identified more duplicates to group quarters and to census records in different states. A reason that CARDS could have identified more duplicates to group quarters is that, in FSPD, links to group quarters were assigned final duplicate probabilities using the exact matching process. Because the FSPD exact matching process did not use information from other links within the household, the criteria to link records together were more strict. A more exact match on person data was required. CARDS criteria may have been less strict.

In an attempt to explain some of the differences between the FSPD and CARDS estimates within and between states, we examined the CARDS links by household (HH) composition, which looks at size of the sample HH and HH duplication status (the number of links between the HHs relative to the size of the source HH). Table 3 shows the distribution by

whether the link was to the same state (the first four categories of geography in Tables 1 and 2) or to a different state, since the latter category is where CARDS tended to find more duplication.

Table 3. CARDS Weighted Estimate of CARDS Only E-sample Links by Household Composition and Geography

Household Composition		Geography			
HH Size	HH Duplication Status	Same State		Different State	
		% CARDS Only	*Total*	*% CARDS Only*	*Total*
1	*All*	36.0% (1.1)	727,889 (23,908)	54.6% (2.3)	132,379 (7,296)
2+	*All*	2.8% (0.3)	3,052,411 (100,883)	8.7% (1.1)	232,581 (18,014)
	Partial - 2+ links	10.3% (0.5)	2,139,959 (64,818)	39.2% (2.3)	202,463 (11,155)
	Partial - Only 1 link	34.1% (0.7)	1,837,816 (35,799)	64.7% (1.0)	853,828 (19,821)
	Total	13.3% (0.3)	7,030,186 (151,162)	50.6% (1.0)	1,288,872 (31,980)
Total		15.4% (0.3)	7,758,075 (159,182)	51.0% (1.0)	1,421,251 (34,133)

Approximately 51 percent of the CARDS links to different states were CARDS only, compared with 15.4 percent of CARDS links to the other geographical distances. We noticed that when more than one person linked between the HUs, the percentage of the CARDS links that were found only by CARDS was lower than for all links combined (in other words, there was more overlap between CARDS and FSPD). This general trend held both for links to different states and for links within a state. However, more CARDS links to different states were CARDS only.

When we examined the links in HHs with more than two people but only one link (called single links), we found they comprised a larger proportion of the links between states than the links within a state (60.1 percent (853,828/1,41,241) vs 23.7 percent (1,837,816/7,758,075)). Furthermore, 64.7percent of the single links to different states were found by CARDS only while the percentage dropped to 34.1 percent of the single links within the same state. The single links were a major source of the difference between FSPD and CARDS for links in different states.

Since many of the FSPD links to different states were single links, many of these links were assigned final duplication probabilities in FSPD by the exact matching process. Due to the large geographic distance, many of these links may have been assigned probabilities less than one. However, in CARDS all links were treated as duplicates (as if they all have a final duplicate probability of one). So even if there had been a lot of overlap between FSPD and CARDS links to different states, the FSPD estimates could have been substantially lower.

Recall that the matching process assigned PIKs in two phases: a Geokey Search phase (address and person information) and a Name Search phase (person information only). We are more confident of links created in the Geokey Search phase, because this phase requires similar address data as well as person data. Thus, Tables 4 and 5 show the CARDS links by whether the PIKs were assigned to the source and/or target record in the Geokey Search phase or not.

Table 4. CARDS Weighted Estimate of CARDS E-sample Links by Match Phase. Links within a State Only.

PIKs Assigned in Geokey Search Phase	Type of CARDS Link					
	CARDS Links	CARDS Only Links	% of CARDS Links That Are CARDS Only	% of Total CARDS only Links	CARDS Links Also in FSPD	% of Total CARDS Links Also in FSPD
Both Source & Target	5,815,854 (134,973)	805,416 (23,770)	13.8% (0.4)	67.4% (0.9)	5,010,438 (125,535)	76.4% (0.6)
Only Source or Target	1,369,758 (35,636)	318,126 (12,317)	23.2% (0.8)	26.6% (0.9)	1,051,632 (32,338)	16.0% (0.5)
Neither Source nor Target	572,463 (25,383)	72,240 (5,502)	12.6% (0.9)	6.0% (0.4)	500,224 (23,521)	7.6% (0.3)
Total	7,758,075 (159,182)	1,195,782 (29,173)	15.4% (0.4)	100.0%	6,562,294 (146,443)	100.0%

Table 5. CARDS Weighted Estimate of CARDS E-sample Links by Match Phase. Links Between States Only.

PIKs Assigned in Geokey Search Phase	Type of CARDS Link					
	CARDS Links	CARDS Only Links	% of CARDS Links That Are CARDS Only	% of Total CARDS only Links	CARDS Links Also in FSPD	% of Total CARDS Links Also in FSPD
Both Source & Target	199,937 (10,740)	58,436 (4,499)	29.2% (1.9)	8.1% (0.6)	141,502 (9,055)	20.3% (1.0)
Only Source or Target	1,092,517 (27,185)	586,635 (15,415)	53.7% (1.1)	81.0% (0.8)	505,882 (19,935)	72.6% (1.1)
Neither Source nor Target	128,797 (6,693)	79,290 (4,948)	61.6% (2.5)	11.0% (0.6)	49,506 (4,275)	7.1% (0.6)
Total	1,421,251 (34,133)	724,362 (17,831)	51.0% (1.0)	100.0%	696,889 (25,540)	100.0%

The categories for the two phases are: (1) Both the source and target PIK assigned in Geokey Search, (2) Only source or target PIK assigned in Geokey Search, and (3) Neither source or target PIK assigned in Geokey Search.

When viewing links within and between states separately the phase in which the CARDS linking was done had some relation to the percentage of CARDS links that are also in FSPD, but more so for links between states. When both links were within state and found in the Geokey Search,13.8 percent of the CARDS links are CARDS only links, compared to about 20 percent (318,126+72,240)/(1,369,758+572,463)) for the other CARDS links. However, for links to different states, 29.2 percent of CARDS links where both records were matched in the Geokey Search were CARDS only links, compared to about 55 percent ((586,635+79,290)/(1,092,517+128,797))for the other CARDS links.

However, Columns 5 and 7 of Tables 4 and 5 show that the distributions of the links by whether PIKs were assigned in the Geokey Search were not dramatically different for those found only by CARDS and those found by both CARDS

and FSPD when tabulated separately by within state and between state. So, problems among CARDS-only links where both PIKs were not assigned in the Geokey Search may be present in such links found by both CARDS and FSPD.

4. Summary

CARDS demonstrated that administrative records can be valuable aids for detecting census duplicates. CARDS style processes have the potential to identify duplicates that statistical and exact matching methods have difficulty detecting – for example, people enumerated with different names, and people whose enumerations have reporting errors. We have seen that CARDS data has been useful in confirmation and denial of FSPD links, and has the potential for finding additional duplicates. But we also have seen reasons here and in a clerical review of a sample of the duplicates outside the surrounding blocks (Byrne, Beaghen, and Mulry 2002), to question some of the CARDS links that were not also found by FSPD. This limited our ability to draw significant conclusions about duplicates missed by FSPD. More research is needed to design methods to adequately address cases in which different people coincidentally have similar person characteristics.

We do believe that administrative records have great potential to be of value for research on preventing and detecting duplicate enumerations for Census 2010. The CARDS process used the results of an HCUF-Numident match with a different purpose. The goal of that match was to associate Census 2000 race data with Numident records. The match strategy and thresholds were developed with the goal of matching the HCUF as completely as possible, while maintaining a reasonably low false match rate over the whole of the HCUF. However, potential census duplicates are a small and special subset of the HCUF, and the effectiveness may not have been as high for such a subset as it was for the entire file. We believe that a CARDS style process that is developed with the sole purpose of detecting census duplicates, and that uses lessons learned from this study, the clerical review, and further research, can produce more complete and accurate results.

5. References

Bean, S. L. (2001) "ESCAP II: Accuracy and Coverage Evaluation Matching Error." Executive Steering Committee For A.C.E. Policy II, Report No. 7., October 12, 2001. U. S. Census Bureau, Washington, DC.

Bean, S. L. and Bauder, D. M. (2002) "Census and Administrative Records Duplication Study," DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-44. Census Bureau, Washington, DC.

Byrne, R., Beaghen, M., and Mulry, M. H. (2002) "Clerical Review of Census Duplicates." DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 43. U. S. Census Bureau, Washington, DC.

Judson, D. H. (2000). "The Statistical Administrative Records System: System Design and Challenges". Paper presented at the NISS/Telcordia Data Quality Conference, November, 2000.

Kostanich, D (2003) "A.C.E. Revision II Design and Methodology." DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 30. U. S. Census Bureau, Washington, DC.

Leggieri, C., Pistiner, A., and Farber, J. (2002) "Methods for Conducting an Administrative Records Experiment in Census 2000." *2002 ASA Proceedings*, American Statistical Association, Alexandria, VA.

Miskura, S. M. (2000) "Results of Reinstatement Rules for the Housing Unit Duplication Operations," Memorandum for P. J. Waite, Decennial Management Division, U. S. Census Bureau, November 21, 2000.

Mule, T. (2002) "Further Study of Person Duplication in Census 2000." DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 51. U. S. Census Bureau, Washington, DC.

Raglin, D. A. and Krejsa, E. A. (2001) "ESCAP II: Evaluation Results for Changes in A.C.E. Enumeration Status." Executive Steering Committee For A.C.E. Policy II, Report No. 16. U. S. Census Bureau, Washington, DC.

Thompson, J., Waite, P., Fay, R., (2001), “Basis of ‘Revised Early Approximations’ of Undercounts Released Oct. 17, 2001.” Executive Steering Committee for A.C.E. Policy II, Report 9a. U. S. Census Bureau, Washington, DC.

U. S. Census Bureau (2003) “Technical Assessment of A.C.E. Revision II” March 12,2003. U. S. Census Bureau.