Clerical Review of Duplicate Enumerations in Census 2000

Rosemary Byrne, Michael Beaghen and Mary H. Mulry[1]
US Census Bureau, Washington DC 20233

Keywords: census coverage error, nonsampling error, Accuracy and Coverage Evaluation Survey

1 INTRODUCTION

The Clerical Review of Census Duplicates study (CRCD) examined the quality of the estimation of census duplicates used in the Accuracy and Coverage Evaluation (A.C.E.) Revision II estimates of coverage error in Census 2000.

As part of the Census 2000, the A.C.E. was an independent coverage measurement survey conducted to estimate the net undercount of the US population. The A.C.E. initially estimated a 1.18 percent undercount rate in the Census 2000 population count of 281,421,906. However, an evaluation of the A.C.E. results, which included a computerized search of the census, yielded an estimate of 2.9 million duplications not discovered by the A.C.E. methodology (Fay 2002). Based on these findings, the Census Bureau produced the A.C.E. Revision Preliminary estimates, which indicated the net undercount was 0.06 percent (Thompson, Waite, Fay 2001, Mule 2002). Recently the Census Bureau further refined its estimate of duplication and produced A.C.E. Revision II estimates that incorporate corrections for the duplications as well as other errors uncovered by evaluations. The revised estimate of the census undercount rate was -0.5 percent, an overcount (U.S. Census Bureau 2003).

The A.C.E. Revision II estimation used data collected in the A.C.E. and its evaluations. The A.C.E. comprised two samples, a population (P) sample to measure census omissions and an enumeration (E) sample to measure census erroneous enumerations. The P sample was obtained by independently listing housing units in a sample of block clusters and conducting person interviews in those housing units. The E sample consisted of the census enumerations in housing units in those sample blocks.

The A.C.E. Revision II also used the computer matching results from the Further Study of Person Duplication (FSPD) (Mule 2002), which was the focus of the CRCD evaluation. The FSPD estimated duplication of E-sample and P-sample cases to enumerations outside the search area around the A.C.E. sample blocks. The A.C.E. Revision II did not use the FSDP results within the search area, since the results of the A.C.E. matching operation within the search area were considered to be the most accurate (Childers 2001).

In the clerical review for CRCD, a highly trained and experienced matching team clerically examined the quality of the identification of census duplicates. The team examined the duplicates identified by the Further Study of Person Duplication used in the A.C.E. Revision II estimation and duplicates identified by the Census and Administrative Records Study (CARDS) (Bean and Bauder 2002), another evaluation of FSPD. The FSPD used a computer-matching algorithm to match the A.C.E. sample enumerations to all other census enumerations. The FSPD algorithm used a statistical matching methodology that assigned a probability of linked records being a match. Links with probabilities above thresholds were considered duplicates. The thresholds varied by the geographical distance between the pair, ranging from links between enumerations in the same block cluster to links in different states. Note that only links in households with two or more links could be identified as duplicates in the statistical matching. The statistical matching component of FSPD was augmented with an exact matching component based on name and birth date for the A.C.E. Revision II estimation. The exact matching component had the ability to identify duplicates in housing units where only one member was duplicated.

The Census and Administrative Records Study (CARDS) examined the effectiveness of the FSPD methodology by comparing the FSPD links to links identified using an administrative records database. The CARDS methodology required first an assignment of a Protected Identification Key (PIK) (based on Social Security Numbers) to each census and P-Sample record. The assignment of PIKs came from a previous study. Some PIKs were assigned using both personal and address information while others were assigned using only personal characteristics. CARDS designated each FSPD duplicate as confirmed (same PIK), denied (different PIKs), or undetermined (PIK could not be assigned to at least one record). In addition, CARDS identified duplicates that FSPD did not designate as

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

duplicates. For more information on the CARDS study, see Bean and Bauder (2002). When CARDS and FSPD exact matching links overlapped, they were included in the review. The clerical review considered only Census information for duplicates identified by CARDS and did not use information from the administrative records database.

In the CRCD, the clerical matchers reviewed duplicates between census enumerations in the A.C.E. sample blocks and census enumerations outside the search area used by the A.C.E. matching operation. The expert matching staff reviewed the whole households of these potential duplicate enumerations. Each linked pair was designated as either a confirmed duplicate, not the same person, or undetermined. With these results, we computed the accuracy rate for the computer identification of duplicates in the census and between the A.C.E. population sample and the census.

The clerical review study also examined FSPD links to census enumerations in housing units that the Housing Unit Duplication Operation (HUDO) (Nash 2000) reviewed as possible duplicates but reinstated in the census, and to census records in housing units that HUDO deleted when these links were also identified by CARDS. The enumerations in housing units both reinstated and deleted by HUDO were not eligible for the E-sample, and are not part of this analysis. Additionally, the clerical review did not review duplicate links made to group quarters. In this paper only results of the review of duplicate links between E-sample census enumerations and E-sample eligible census enumerations outside the A.C.E. search area are discussed. The methodology and results for the P-sample are similar. For complete CRCD results, including those for the P sample and for duplicate links between E-sample census enumerations and the enumerations in housing units reinstated and deleted by HUDO, see Byrne et al (2002).

2 METHODOLOGY

2.1 Design of the Clerical Review Operation

Using a specially designed computer based clerical matching application, the expert matchers reviewed the census information for all household members, including those not designated as duplicates. The data include name, sex, age, date of birth, relationship to householder, race and Hispanic origin. The expert matchers entered a code indicating whether a pair was a duplicate along with "why" codes that indicated the reason for declaring the pair a duplicate or denying the duplication. They also included notes if applicable. For the household members that were not designated as having a duplicate by FSPD or CARDS, the expert matchers entered a code indicating whether an additional duplicate was found. If there was a 'better' duplicate in the census household other than the one designated by FSPD or CARDS, the expert matchers recorded a code showing the duplicate was rearranged.

2.2 Sample Selection

The expert matchers reviewed housing units with two or more duplicate links identified by the FSDP and duplicates identified by CARDS, another evaluation of FSPD. The review consisted of households with duplicates in the Evaluation Sample clusters, a subsample of A.C.E. clusters (Davis & Raglin 2001). The review included duplicate pairs that FSPD linked but did not declare to be duplicates because the probability of being a duplicate was below the threshold. For the E sample, the review was restricted to duplicates between enumerations in the E sample and census enumerations outside the A.C.E. search area (Childers 2001).

The review was restricted to households where FSPD found that more than one member was duplicated, although households with only CARDS duplicates were reviewed when they had only one link. The additional cases from CARDS did not include links to enumerations in group quarters. We restricted the additional cases to links between households because we believed that few additional duplicates would be found between a household and a group quarters residence. The clerical workload included a total of 18,713 links in 11,935 housing units. From the E sample there were 10,248 links in 6,412 housing units while 8,465 links in 5,523 housing units were from the P sample.

2.3 Review of Duplicates

The expert matchers determined whether the sets of two enumerations referred to the same person. The expert matchers assigned a "why" code that indicated the reason for declaring the pair a duplicate, denying the duplication, or not being able to decide. When the expert matcher could not decide, the case was considered unresolved. This occurred most often when one or both of the linked records contained an insufficient amount of information to consider the pair a duplicate. In addition to reviewing the linked pairs, the expert matchers also reviewed household members not linked by FSPD to determine if they too had duplicates.

The clerical review was generally restricted to households with at least two duplicates in another

housing unit. The study did not evaluate duplicates identified in households with only one duplicate, that is, those duplicates identified by the exact matching. However, it did review exact matching links for households where CARDS identified a single duplicate in another housing unit and FSPD statistical matching found none.

The study could only find missed duplicates within households where duplicate links were identified by the statistical matching component of FSPD and/or CARDS.

3 RESULTS

Tables 1, 2 and **3** show the results of the clerical review coding for E-sample links (except those linking to deleted and reinstated units) sent for review. The columns in **Tables 1** and **2** represent the CARDS status. CARDS duplicates were those with a CARDS status of "confirmed", and all those listed in **Table 3**, the "CARDS only" table. A CARDS status of "denied" means CARDS concluded that the two enumerations were different. A CARDS status of "undetermined" means that CARDS could not assign a unique identification number to the one or both of the enumerations in an FSPD link and therefore could not assess the duplicate status. This may happen when CARDS was unable to find a match, or when it found multiple matches for a given record.

The rows in each table show the distribution (weighted) of the clerical coding separately for each possible FSDP outcome: duplicate, not a duplicate, and unresolved.

In **Tables 1, 2** and **3** we analyze the results by considering the three possible FSDP statuses; those identified as duplicates by the statistical matching component of FSPD (Table 1), those linked by FSPD's statistical matching but not declared duplicates because their score was below the threshold (Table 2), and those not identified by FSPD's statistical matching, i.e., CARDS only (Table 3).

Table 1 shows the cases considered to be duplicates by FSPD. Ignoring the CARDS status, the clerical coding confirmed 94.9% of the FSPD duplicates. Of the 3.8% of the FSPD duplicates considered not to be a duplicate by the clerical coding, 75% were also denied in the CARDS study. Overall, both CARDS and CRCD agreed that 73.4% (922,325 out of 1.25 million) of the duplicate links found by FSPD's statistical matching were duplicates.

Note that **Table 1** shows an evaluation of 1.25 million of the 3.4 million duplicates found outside the A.C.E. surrounding blocks by FSPD and used in the A.C.E. Revision II estimation (Mule, 2002). The estimate of 3.4 million includes duplicates to enumerations in E-sample eligible housing units, group quarters, and housing units reinstated and deleted by HUDO.

Table 2 shows the clerical coding for the cases linked by FSPD but considered to be below the threshold to be considered a duplicate. Here we see that disregarding the CARDS status, 93.8% of these links were also not considered to be a duplicate by the expert matchers. The expert matchers did determine that 4.6% of the links not considered to be duplicates by FSPD were indeed duplicates. About half of these (90,092) were also identified as duplicates by CARDS.

Overall, both CARDS and the clerical review study determined that 81.7% (3.2 million out of 3.9 million) of the links FSPD found but did not declare duplicates were not duplicates.

Table 3 shows the clerical coding results for cases identified by CARDS, but not by the statistical matching part of FSPD. This table includes duplicates identified by the exact matching part of FSPD but not identified in the statistical matching part of FSPD. They were among the CARDS cases that happened to be linked in the exact matching and do not represent a probability sample of the exact matching duplicates.

For cases identified by CARDS but not by FSPD's statistical matching component, the confirmed duplication rate from the clerical matching is much lower, 55.3%. Of these, 61% (695,968 out of 1.14 million) were also identified by the exact matching.

Note that about 175,398 (20%) of the 871,366 links also identified by exact matching were either unresolved or not considered to be duplicates by the clerical review. However, no conclusions can be drawn about the quality of exact matching duplicates not also found by CARDS based on these results since these were not sampled for the clerical review study. Additionally, in the A.C.E. Revision II estimation, the exact matching cases received a probability of being a duplicate, which was usually less than 1. Therefore, when the probability is applied to these cases, their weighted sum is less than 871,366.

Not appearing in any of the tables is the number of additional duplicates found. The expert matchers found additional duplicates because they reviewed all members of the household containing the duplicate links. Only 46 unweighted household members, or

0.2%, who were not previously identified as duplicates by FSPD or CARDS were determined to be duplicates.

4 SUMMARY

The Clerical Review of Census Duplicates study represented the first time that skilled expert matchers clerically reviewed in a systematic way a sizeable sample of duplicate links found by the automated duplicate searches, the FSPD and the CARDS. This review yielded several important insights. We found that the links identified by the statistical component of the FSPD appear to have both a high level of genuine duplication and a low level of erroneously identified duplication. The level of erroneous duplication in the FSPD does not threaten the integrity of the A.C.E. Revision II estimates.

Additionally, we found that when the clerical reviewers confirmed the CARDS links, these links likely represent genuine duplication missed by the FSPD. This information can be used in the future to refine the FSPD algorithm. However, the clerical reviewers disagreed with a large proportion of CARDS-only links. Since CARDS used a previous assignment of PIKs to census enumerations, we recommend tailoring the methodology for the PIK assignment to identification of duplicates based on further analysis of the CRCD and CARDS results.

Lastly, few additional duplicates were found by the clerical matchers, indicating a good degree of accuracy in the FSPD algorithm for within household duplicate identification. In conclusion, the Clerical Review of Census Duplicates study confirmed the validity of using the FSPD results for the A.C.E. Revision II estimates.

5 REFERENCES

Bean, S. and Bauder, D.M. (2002). "Census and Administrative Records Duplication Study," DSSD Revised A.C.E. Estimates Memorandum Series #PP-44. Census Bureau, Washington, DC.

Byrne, R., Beaghen, M., and Mulry, M. (2002). "Clerical Review of Census Duplicates" DSSD A.C.E. Revision II Memorandum Series #PP-43. Census Bureau, Washington, DC.

Childers, D. (2001). "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, Revised. Census Bureau, Washington, DC.

Davis, M. & Raglin, D. (2001). "Creation of Master Data Variance Files for Coverage Measurement Evaluations," Planning, Research and Evaluation Division TXE/2010 Memorandum Series: CM-GES-S-01-R. Census Bureau, Washington, DC.

Fay, R. (2002). "Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee on A.C.E. Policy II Report 9. U.S. Census Bureau, Washington, D.C.

Kostanich, D. (2003). "A.C.E. Revision II: Design and Methodology," DSSD A.C.E. Revision II Memorandum Series #PP-30

Mule, T. (2001). "Person Duplication in Census 2000," Executive Steering Committee on A.C.E. Policy II Report 20. U.S. Census Bureau, Washington, D.C.

Mule, T. (2002). "Further Study of Person Duplicates". DSSD Revised A.C.E. Estimates Memorandum Series #PP-51. Census Bureau, Washington, DC.

Mulry, M. (2002). "Chapter 7: Assessing the Estimates," in "Revised ACE: Design and Methodology." Revised A.C.E. Estimates Memorandum Series #PP- 30. Census Bureau, Washington, DC.

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

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 Estimates." March 2003. U.S. Census Bureau, Washington, D.C.

6 TABLES

Table 1 E-sample Duplication by Study, FSPD Statistical Matching Duplicate - Weighted, Standard Errors in Parentheses

Clerical Review Status	Identified in FSPD's Statistical Matching as a Duplicate			**Total**
	CARDS status			
	Confirmed	Denied	Undetermined	
Duplicate	922,325 (59,472)	7,737 (2,101)	262,702 (23,708)	1,192,765 (71,834) 94.9% (1.0)
Not a Duplicate	3,536 (2,378)	35,654 (9,456)	8,121 (3,278)	47,311 (11,223) 3.8% (0.9)
Unresolved	10,841 (4,311)	0.0 (0.0)	5,496 (2,853)	16,336 (6,602) 1.3% (0.5)
Total	936,702 (59,639) 74.5% (1.6)	43,391 (9,679) 3.5% (0.8)	276,320 (24,290) 22.0% (1.5)	1,256,413 (73,671) 100%

Table 2 E-sample Duplication by Study, FSPD linked but not a duplicate - Weighted, Standard Errors in Parentheses

Clerical Review Status	**Identified in FSPD's Statistical Matching, but Not a Duplicate**			**Total**
	CARDS status			
	Confirmed	Denied	Undetermined	
Duplicate	90,092 (14,930)	18,239 (5,895)	76,603 (11,086)	184,934 (21,891) 4.6% (0.5)
Not a Duplicate	22,145 (4,911)	3,248,663 (143,023)	459,892 (30,880)	3,730,701 (153,928) 93.8% (0.6)
Unresolved	5,514 (2,214)	30,504 (9,226)	25,890 (7,341)	61,908 (12,331) 1.6% (0.3)
Total	117,752 (15,824) 3.0% (0.4)	3,297,406 (143,797) 82.9% (0.9)	562,385 (34,514) 14.1% (0.8)	3,977,543 (157,888) 100%

Table 3 E-sample Duplication by Study, CARDs Only - Weighted, Standard Errors in Parentheses

Clerical Review Status	**Status**		**Total**
	Also identified by exact matching	CARDS only	
Duplicate	695,968 (36,984)	445,703 (30,309)	1,141,672 (51,642) 55.3% (1.6)
Not a duplicate	72,647 (9,549)	564,881 (32,568)	637,528 (35,301) 30.9% (1.4)
Unresolved	102,751 (12,495)	184,071 (16,765)	286,822 (21,165) 13.9% (0.9)
Total	871,366 (40,427) 42.2% (1.4)	1,194,656 (52,033) 57.8% (1.4)	2,066,022 (71,515) 100%