Overview and Results of Further Study of Person Duplication for the A.C.E. Revision II

Vincent Thomas Mule Jr. and Deborah Fenstermaker

For purposes of Acuracy and Coverage Evaluation (A.C.E.) Revision II estimates, the Further Study of Person Duplication (FSPD) used matching and modeling techniques to link the Enumeration (E) and Population (P) samples to census enumerations. These links allowed the A.C.E. Revision II estimates to correct for measurement error in the original A.C.E. estimates. The matching algorithm used statistical matching to identify linked records. Statistical matching allowed for the matching variables not to be exact on both records being compared. Because linked records may not refer to the same individual even when the characteristics used to match the records are identical, modeling techniques were used to assign a measure of confidence, the duplicate probability, that the two records refer to the same individual. These duplicate probabilities were used in the A.C.E. Revision II estimates.

Mule (2001) reported results for initial attempts at measuring the extent of person duplication in Census 2000. This work was conducted by an inter-divisional group as part of the further research to inform the October, 2001 decision on adjusting census data products. (This study is referred to as the ESCAP II duplicate study in this document.) The ESCAP II duplicate study used conservative computer matching rules to minimize the number of false matches that could be introduced when doing a nation-wide search since there was no clerical review of the results. As a consequence of the matching rules, comparisons to benchmarks indicated that the ESCAP II duplicate estimates were a lower bound. Specifically, comparing the ESCAP II results within the A.C.E. sample area to the A.C.E. clerical matching results showed that only 37.8 percent of the census duplicates were identified. Fay (2001, 2002) estimated the matching efficiency at 75.7 percent when accounting for the census records out-of-scope for the A.C.E. duplicate search, the reinstated and deleted records from the Housing Unit Duplication Operation (HUDO), Nash (2000). See Mule (2001) for more detail on this ESCAP II work.

The ESCAP II census duplicate methodology satisfied the intended project goals and provided a valuable evaluation of the census by showing that person duplication existed. However, limitations of the methodology made it difficult to get a good handle on the magnitude of the person duplication in the census.

Overview of Duplicate Study Plan

The A.C.E. Revision II duplicate plan involved matching the full E and P samples to the census to establish potential duplicate links. Then, modeling techniques were used to identify the links most likely to be duplicate enumerations and to assign a measure of confidence that the links are duplicates. Key differences with the ESCAP II study include extending the use of statistical matching and developing models to assign a duplicate probability to the links. An advantage of duplicate probabilities over the Poisson model weights used in ESCAP II is that all duplicate links outside the A.C.E. search area could be reflected in the A.C.E. Revision II estimates. Fay (2001, 2002) used a subset of the ESCAP II duplicate links to produce a lower bound on the level of erroneous enumerations that the A.C.E. did not measure.

Estimates of census duplication were based on matching and modeling of the E-sample cases to the census. For purposes of A.C.E. Revision II estimation, the P sample was matched to the census as well to account for measurement error of residence status in that system but did not contribute to estimates of person duplication in the census. The P sample included all nonmovers, outmovers and inmovers.

The matching algorithm consisted of two stages. The first stage was a national match of persons using statistical matching, Winkler (1995). Statistical matching attempted to link records based on similar characteristics or close agreement of characteristics. Exact matching required exact agreement of characteristics. Statistical matching allowed two records to link in the presence of missing data and typographical or scanning errors. The Statistical Research Division matching software called Bigmatch, Yancey (2002), was used in the first stage.

Six characteristics common to both files, called matching variables, were used to link records in the full E and P sample with records in the census. The matching variables were first name, last name, middle

initial, month of birth, day of birth, and computed age. Matching parameters were associated with each matching variable that measure the degree to which the matching variables agree between the two records, ranging from Full Agreement to Full Disagreement. The measurement of the degree to which each matching variable agreed was called the variable match score. The overall match score for the linked records was the sum of the variable match scores. Full agreement of at least four characteristics was required to be considered a duplicate link. Because this study was a computer process without the benefit of a clerical review, this limitation of the statistical matching was necessary to minimize linking records having similar characteristics but were different people. This was particularly a concern with looking for duplicate enumerations across the entire country. Second, the total match score had to be 4.7 or greater. This minimum score was about half the total score for full agreement of all matching variables. The need to use statistical matching at the first stage was apparent after the limited success of the ESCAP II exact matching in identifying the A.C.E. duplicates in the A.C.E. sample areas. The statistical matching yielded better identification of the A.C.E. duplicates, but to identify all of the A.C.E. duplicates would have required fewer characteristics to be exact matches, thus opening the door to high numbers of false links.

The search for duplicate links between the full E and P samples and the census was limited to those pairs that agree on certain identifiers or blocking criteria. True matches can be missed by using blocking criteria so we used four sets of blocking criteria to minimize the number of missed matches.

The blocking criteria were:

1. First name, Last name
2. First name, First initial of last name, Age groupings (0 - 9, 10 - 19, 20 - 29, etc.)
3. Last name, First initial of first name, Age groupings (0 - 9, 10 - 19, 20 - 29, etc.)
4. First initial of first name, First initial of last name, Month of birth, Day of birth

At the first stage of matching it was possible for one sample case to link to multiple census records. All of these links were retained for the second stage of matching.

The second stage of matching was limited to matching persons within households. If an E- or P- sample case linked to a census record in a group quarter, the case did not go to the second stage. The first stage established a link between persons in two housing units. The second stage was a statistical match of all the household members in the sample housing unit to all of the household members in the census housing unit. The second-stage matching variables were the same as the first stage; however, the matching parameters differed. Using a subset of the first-stage links, the second-stage matching parameters were derived using the Expectation-Maximization (EM) algorithm; see Winkler (1995). A key difference between the first and second stage parameters was that there was considerably less emphasis on needing the last name to agree in the second stage. This intuitively makes sense since this matching was within a household.

The Statistical Research Division Record Linkage software, Winkler (1999), was used for the second stage. Only one set of blocking criteria was used at the second-stage, the household. The sample records were allowed to link to only one census record within the household. As a consequence, this limited our ability to pick up within-household duplicate links. Each link had an overall match score based on the second-stage matching.

Occasionally, first and last name was captured in reverse order on the data files. The first name was in the last name field and the last name was in the first name field. When the data was in reverse-order on one file but not the other, it was difficult to identify these duplicate links. To attempt to identify these cases, the first and last name fields were reversed and then matched to the census files a second time. The duplicate links from both runs, name in the usual order and in reverse order, were input to the modeling.

The set of linked records from the second-stage matching and the links to group quarter enumerations from the first stage consisted of both duplicate enumerations and person records with common characteristics. Using two modeling approaches, the probability that the linked records were duplicates was estimated. One approach used the results of the statistical matching and relied on the strength of multiple links within the household to indicate person duplication. The second relied on an exact match of the census to itself and the distribution of births, names and population size to indicate if the individual link was a duplicate. These two approaches were referred to as the statistical match modeling and the exact match modeling, respectively. These two approaches were combined to assign to each sample case with a link to a census enumeration an estimated probability of being a duplicate.

The statistical match modeling was used when two or more duplicate links were found between housing units in the second stage. After the second-stage matching, each duplicate link between a sample household and census household had an overall match score. So, for each sample household, a set of match scores was observed. For any resulting set of match scores, a probability of not observing this set of match scores was estimated. The higher this probability, the more likely that the set of linked records in the household were duplicates.

The estimate of the probability of not observing this set of match scores assumed independence of the individual match scores within each household. This assumption was based on using the EM algorithm to determine the second-stage matching parameters. The probability of observing the individual match scores was estimated from the empirical distribution of individual match scores resulting from the second-stage matching. Further, this measure accounted for the number of times that a unique sample household was matched to different census households within a given level of geography. The probability of not observing this set of match scores was translated into 1/0 "statistical match" duplicate probability based on critical values which varied by level of geography.

The exact match modeling relied on an exact match of the census to itself. The methodology took into account the overall distribution of births, frequency of names and population size in a specific geographic area. Duplicate probabilities were computed separately by geographic distance of the links. Further, duplicate links were modeled separately by how common the last name was as well as separately for Hispanic names. See Fay (2003) and Fay (2002b) for more information.

The two approaches were combined to assign an estimated probability that the linked records were duplicates. The duplicate probability for the links to group quarters in the first stage and one-person household links were from the exact match modeling. For all other links, the duplicate probability was the larger of the two model estimates. For non-exact matches, this was always from the statistical match modeling. For exact matches, adjustments were made to account for the integration of these two methods.

Based on the results of this matching and modeling, an overall estimate of census duplicates was derived from the E-sample links. Further, these results provide for each full E- and P-sample person who linked to an enumeration outside the A.C.E. search area the probability that they were in fact the same person. These probabilities were used in the A.C.E. Revision II estimates.

Results of the E-sample Matching

Table 1 shows the results of the E-sample matching. Some highlights of these results:

- Our study estimated 5.8 million duplicates in the Census.
- Our study estimated there would have been 8.7 million duplicates in the census if the HUDO had not been implemented. We included this estimate because the A.C.E. Revision II estimation used links to cases reinstated and deleted by the HUDO.
- For the E-sample Eligible universe, our improvements in computer matching identified 61.9 percent more duplicates within the cluster as compared to our ESCAP II results. We identified 1.2 million duplicates within the cluster as compared to 725,000 duplicates identified by the ESCAP II analysis.
- Overall, we identified 3.5 million duplicates in the E-sample Eligible universe. This was 580,000 more duplicates than was found in the ESCAP II analysis. Most of these were found within the cluster or one ring of surrounding blocks.
- Within the E-sample eligible universe, we identified 2.1 million duplicates outside the surrounding blocks. This total is comparable to the 2.1 million duplicates outside the surrounding blocks identified for the ESCAP II analysis. While we have roughly the same aggregate total, we believe we have done this by more accurately determining the duplication status of each case.
- While our study and the ESCAP II analysis estimated roughly the same aggregate total outside the surrounding blocks, the distribution of duplicates by geography has changed. Our study estimated more duplicates in the same county and fewer in a different state than the ESCAP II analysis. We believe this result is based on the improvements in the matching and modeling which we were able to implement in this analysis.

- Our estimate of housing unit to group quarter duplication is similar to our ESCAP II estimate.

- The estimates of duplicates to the Reinstated and Deleted universes are consistent with the results from the ESCAP II analysis. We were expecting this. The reinstated and deleted cases were identified during the computer matching of the HUDO. Since computer matching of person records was used in the HUDO operation, we were expecting our computer matching process this time to identify the same duplicates as in the ESCAP II analysis.

Table 1: Overall Estimates of Person Duplication

Geography		Type of Record in Census: E-sample Eligible	GQ	Reinstate	Total (Records in Census)	Duplicates to Records Deleted During HUDO	Total (Records in Census + Records Deleted During HUDO)
Within Cluster		1,173,344 (46,173)	76,381 (15,736)	1,058,548 (48,295)	2,308,273 (74,924)	1,967,199 (94,454)	4,275,472 (129,245)
Surrounding Block		259,805 (21,718)	25,373 (9,701)	24,751 (6,971)	309,929 (24,734)	678,355 (57,469)	988,284 (65,896)
Outside Surrounding Block	Same County	1,011,920 (24,292)	231,774 (39,795)	482,015 (27,797)	1,725,709 (55,097)	208,246 (20,789)	1,933,956 (59,590)
	Different County, Same State	563,270 (18,873)	190,417 (9,488)	88,331 (12,567)	842,018 (25,154)	35,111 (7,262)	877,129 (26,615)
	Different State	527,796 (23,744)	91,793 (7,093)	20,959 (17,316)	640,548 (31,433)	16,184 (4,902)	656,732 (33,930)
Total		3,536,136 (68,045)	615,738 (46,003)	1,674,604 (60,317)	5,826,477 (110,721)	2,905,096 (116,541)	8,731,572 (177,071)

Source: Mule (2002). Standard errors in parentheses

Decision Not to Adjust For Efficiency of Identifying Census Duplicates

We estimated efficiency by using the duplicates detected by the A.C.E. clerks as a benchmark. We estimated two efficiency measures. The first was the estimate using only the links to cases in the A.C.E. universe as was done by Mule (2001). Using this approach, the overall efficiency was 64.7 percent. Mule (2001) estimated an efficiency of 37.8 percent within the cluster for the ESCAP II analysis.

The second estimate used the cases in the A.C.E. universe and duplicates to the cases detected in the HUDO as was done by Fay (2002). Using this approach, we estimated an overall efficiency of 86.9 percent. Fay (2002) estimated an efficiency of 75.7 percent in his ESCAP II analysis.

Table 2 shows the results of applying both adjustments within the cluster. Both methods showed that we were more efficient in identifying duplicates when there were two or more duplicates between the housing units. For this group, we were able to effectively utilize statistical matching techniques to identify these duplicates. When there was only one duplicate between the units, we had to rely on exact matching methods which limits the number of duplicates that we could detect.

To apply these adjustments based on duplicates within the cluster to duplicates detected outside the cluster requires the assumption that for the specified subgroups, the mechanism that is causing the duplicates within the cluster is similar for the duplicates outside the cluster. This assumption is debatable because duplicates within the cluster can be caused by misdelivery of forms or families living close together. As the geographic distance increases, the duplicates are more likely to be movers or children in joint-custody situations. Also there may be other variables like age or

Table 2: Efficiency Estimates Within Cluster

HH Size	Number of Links	Including Links to Reinstates and Deletes: No — ACE Within Cluster	FSPD Estimate	FSPD Efficient (%)	Including Links to Reinstates and Deletes: Yes — Denominator[1]	FSPD Estimate	FSPD Efficient (%)
1 Person to 1 Person	Only 1	204,604 (10,055)	16,756 (2,882)	8.19 (1.35)	393,295 (14,075)	205,447 (10,299)	52.24 (1.79)
1 Person to 2+	Only 1	139,038 (8,146)	36,271 (3,744)	26.09 (2.48)	143,009 (8,254)	40,243 (3,957)	28.14 (2.52)
2+ to 2+	Whole HH	952,280 (39,696)	747,682 (31,772)	78.51 (2.37)	3,362,979 (92,348)	3,158,382 (90,191)	93.92 (0.77)
	Partial (2+)	329,631 (21,963)	318,557 (19,666)	96.64 (5.67)	741,709 (39,919)	730,636 (39,003)	98.51 (2.55)
	Only 1	148,869 (8,579)	29,288 (3,356)	19.67 (2.13)	150,722 (8,620)	31,141 (3,494)	20.66 (2.18)
Total		1,774,421 (53,349)	1,148,555 (43,185)	64.73 (1.27)	4,791,715 (119,603)	4,165,848 (114,677)	86.94 (0.58)

Source: Mule (2002).

[1] The denominator of the Fay alternative is the A.C.E. estimate plus the FSPD estimate of duplicates to reinstates and deletes. Standard errors in parentheses.

the type of response (Both Mail returns, One Mail/One Non-Mail, or Both Non-Mail) which can show differential efficiency. Including these variables could produce different adjustments. Based on concerns about the assumptions required, we decided not to adjust the estimates for efficiency.

Results of the P-sample Matching

Table 3 shows the results of the P-sample matching. We identified a large number of P-sample nonmovers who were enumerated at another residence outside the one ring of surrounding blocks. These cases raise the question as to whether some of these people were truly residents of the cluster on April 1, 2000. Our results show that approximately half of these cases were nonmatches within the cluster. The A.C.E. Revision II developed a methodology to account for measurement error in the residence status of these cases in the revised estimates.

References

Fay, Robert E. (2001), "ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee for A.C.E. Policy II, Report 9, Preliminary Version, October 26, 2001.

Fay, Robert E. (2002), "ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee for A.C.E. Policy II, Report 9, Revised Version, March 27, 2002.

Fay, Robert E. (2002b), "Probabilistic Models for Detecting Census Person Duplication," Proceedings of the Survey Research Methods Section, American Statistical Association.

Table 3: Overall Results of Matching the Nonmover Residents

		Type of Record							
		E-sample Eligible		GQ		Reinstate		Deleted	
		Match Status of P sample		Match Status of P sample		Match Status of P sample		Match Status of P sample	
		Nonmatch	Match	Nonmatch	Match	Nonmatch	Match	Nonmatch	Match
Geography									
Within Cluster		416,280 (17,506)	199,026,173 (2,078,493)	0 (0)	92,379 (22,905)	473,167 (57,598)	912,493 (45,194)	242,867 (33,394)	2,050,732 (117,371)
Surrounding Block		512,407 (40,315)	8,886,048 (547,289)	5,158 (2,874)	4,118 (1,668)	50,725 (13,974)	61,334 (14,600)	26,104 (7,477)	323,939 (30,050)
Outside Surrounding Blocks	Same County	2,059,658 (116,361)	1,194,385 (34,618)	39,927 (8,720)	127,393 (25,135)	12,843 (3,963)	195,517 (17,458)	56,759 (24,401)	96,294 (13,639)
	Different County, Same State	403,823 (28,067)	651,502 (23,513)	29,868 (4,155)	86,527 (6,467)	3,791 (1,732)	39,092 (7,308)	7,676 (3,455)	10,575 (2,928)
	Different State	268,031 (19,922)	843,350 (24,656)	15,480 (2,312)	102,439 (6,299)	3,851 (2,348)	3,272 (839)	2,871 (1,017)	10,071 (2,574)
Total		3,660,200 (132,526)	210,601,459 (2,192,069)	90,433 (10,535)	412,855 (35,536)	544,376 (59,711)	1,211,708 (51,211)	336,277 (43,085)	2,491,612 (124,742)

Source: Mule (2002). These estimates include the residence probability. For this table, a case was considered a match if the probability of being match was greater than zero. We used the residence probability and match probability from the March, 2001 estimates. Standard errors in parentheses.

References (Cont.)

Fay, Robert E. (2003), "Probabilistic Models for Detecting Census Duplication at the Person and Household Levels," Proceedings of the Survey Research Methods Section, American Statistical Association

Mule, Thomas (2001), "ESCAP II: Person Duplication in Census 2000," Executive Steering Committee for A.C.E. Policy II, Report 20, October 11, 2001.

Mule, Thomas (2002), "Further Study of Person Duplication Statistical Matching and Modeling Methodology," A.C.E. REVISION II MEMORANDUM SERIES PP-51, December 31, 2002.

Nash, Fay (2000), "Overview of the Duplicate Housing Unit Operations," Census 2000 Information Memorandum Number 78, November 7, 2000.

Winkler, William (1995), "Matching and Record Linkage," *Business Survey Methods*, ed. B. G. Cox et. al. (New York: J. Wiley, 1995), pp. 355-384.

Winkler, William (1999), "Documentation for Record Linkage Software," U.S. Census Bureau, SRD.

Yancey, William (2002), "BigMatch: A program for Extracting Probable Matches from a Large File for Record Linkage," U.S. Census Bureau, SRD.