Chapter 5: Further Study of Person Duplication in Census 2000

Debbie Fenstermaker and Thomas Mule 12/31/02

Evaluations of the March 2001 A.C.E. coverage estimates indicated the A.C.E. failed to detect a large number of erroneous census enumerations. One type of these census erroneous enumerations was duplicate census enumerations; census enumerations included in the census two or more times. The A.C.E. was not specifically designed to detect duplicate census enumerations beyond the search area. However, the expectation was that the A.C.E. would detect that these E-sample enumerations had another residence and that roughly half the time this other place was the usual residence. Feldpausch (2001) showed this expectation was not met.

For purposes of A.C.E. Revision II estimates, this study used matching and modeling techniques to identify duplicate links between the full E and P samples to census enumerations. 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.

This chapter lays out the matching and modeling methods. This study did not identify which enumeration was in the correct location. A component of the A.C.E. Revision II estimation methodology was the determination of the probability that the sample case was in the correct location given that it had a link to a census enumeration outside the A.C.E. search area. This impacts the status of correct enumeration in the E sample and the status of residence in the P sample; see Chapter 6.

5.1 Background

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, Nash (2000).

The ESCAP II matching was a two step process. First the sample of census records were matched to the full census on first name, last name, month of birth, day of birth and computed age. Age was allowed to vary by one year. We accounted for middle initials and suffixes being scanned into the first name field but otherwise, the other characteristics had to be exact matches at this stage. This first-stage match established a link between households. In the second stage, all person records in the linked households from the first stage were statistically matched using first name, middle initial, last name, month of birth, day of birth, and computed age. The matching parameters used in the statistical matching were borrowed from other Census 2000 matching operations. Mule (2001) describes this matching algorithm in more detail.

To reduce the impact of false matches, particularly with respect to persons with common names and the same month and day of birth, model weights were applied to each set of linked records as a measure of confidence that the linked records were indeed duplicates. Due to schedule constraints, a national, Poisson model was used in lieu of a probability model.

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.

5.2 Overview of Duplicate Study Plan

Like the ESCAP II study, 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, but did not contribute to estimates of person duplication in the census. The A.C.E. Revision II estimation methodology adjusted the A.C.E. correct enumeration rate for the E-sample cases with links outside the A.C.E. search area. Further, the A.C.E. Revision II estimation methodology adjusted the A.C.E. match rate for the P-sample cases which linked to census cases outside the search area.

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.

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. 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. 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. Blocking criteria were sort keys and were used to increase the computer processing efficiency by searching for links where they were most likely to be found. For instance, if we wanted to search only for duplicates when the first and last names agree, then both the sample and the census files would have been sorted by first name and last name, the blocking criteria. Then, all possible pairs within each first name/last name combination would have been searched for duplicate links. True matches can be missed by using blocking criteria. We used four sets of blocking criteria. Multiple sets of blocking criteria minimized the number of missed matches.

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 does not go to the second stage. The first stage established a link between 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.

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.

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 geographical distance of the links. Further, duplicate links were modeled separately by how common the last name was as well as separately for Hispanic names.

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 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.

5.3 Matching Algorithm

Efforts to increase matching efficiency over the ESCAP II method included carrying out statistical matching of persons at the first stage and the use of more discriminating matching parameters at the second stage.

5.3.1 Inputs

Both the full E and P samples were matched to the census records. The E sample records reflected any updates made by the clerical staff during the A.C.E. matching operation when the census characteristics were incorrectly transcribed or scanned. The P sample included all nonmovers, outmovers and inmovers. The same matching algorithm was used for the full E and P samples.

The census files consisted of data-defined person records for both the household and group quarters populations. Both the reinstated records and the deleted records from the Housing Unit Duplication Operation (Nash 2000) were included in the matching so that these links could be reflected in the A.C.E. Revision II estimates.

5.3.2 First Stage: Person-Level Matching

The first stage was a statistical match of the full E and P samples to the census. This was a national match where each full sample case was compared with census records across the nation to assess how well the matching variables agreed.

The matching variables were first name, last name, middle initial, month of birth, day of birth, and computed age. The matching variables and parameters are given in Table 5.1. The agreement weight and the disagreement weight are the matching parameters of each variable. We used standard matching parameters at the first stage. The relationship of the agreement and disagreement parameters translated into the match score for each variable. For example, the full agreement value for first name was 2.1972; whereas, the full disagreement match score was -2.1972. The sum of the variable match scores was the total match score. When the match score was 9.4006, this indicated full agreement of all variables. A match score of -9.4006, on the other hand, indicated full disagreement.

Table 5.1: First-Stage Matching Parameters

Matching Variables	Type of Comparison	Matching Parameters		Match Score	
		Agreement Weight (m)	Disagreement Weight (u)	Agreement ln(m/u)	Disagreement ln(1-m/1-u)
First Name	String (uo)	0.9	0.1	2.1972	-2.1972
Last Name	String (uo)	0.9	0.1	2.1972	-2.1972
Middle Initial	Exact	0.7	0.3	0.8473	-0.8473
Month of Birth	Exact	0.8	0.2	1.3863	-1.3863
Day of Birth	Exact	0.8	0.2	1.3863	-1.3863
Computed Age	Age (p)	0.8	0.2	1.3863	-1.3863
Total				9.4006	-9.4006

The type of comparison indicated the statistical matching method for comparing the variables. For example, the string comparitor was used for first name and last name. This method dealt with typographical error in names. For example, “Tim” and “Tum” can get a positive agreement score. An exact match algorithm would have treated these as a disagreement. For age, the age values could have been off ± one year and still received a full agreement score on computed age.

The Statistical Research Division matching software called Bigmatch, Yancey (2002), was used in the first stage. This software allowed a sample record to link to more than one census record. We wanted this capability since it was possible for there to be more than two enumerations of the same person in the census.

Four blocking criteria were used. Blocking restricted the comparisons of records to only those that exactly agreed on certain values. Most records that did not agree on the values below are probably not duplicates. 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

All possible links within each blocking criteria were compared. For each comparison, the variable match score and the total match score was computed. The first-stage matching decision rule required these scores. First, a match must have had at least four of the match variables in full agreement. This means that four of the variables had to have to have a match score equal to the agreement match score in Table 5.1. The one exception was the middle initial. When the middle initial was blank, it was considered to be in full agreement in this study since the middle initial was often missing on the sample and census records. In this case, the middle initial score was zero. 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.

First-Stage Match Rule

Table 5.2 shows the distribution of A.C.E. links within cluster that were identified by the resulting number of matching variables in full agreement. There were a total of 10,559 duplicate links identified by the A.C.E. clerical staff that agreed on the first letter of the first and last name. The table shows the number of these A.C.E. duplicates identified as the number of matching variables in full agreement decreased. The table also shows the number of total links that were identified. The percent A.C.E. links in each row of the table decreases as the number of matching variables in full agreement decreases.

Table 5.2: Distribution of Links Within A.C.E. Clusters by Full Agreement

Number of Variables in Full Agreement	A.C.E. Links			Total Links	Percent of A.C.E. Links in Row
	Count	Percent	Cumulative Percent		
6	2348	22.2%	22.2%	2451	95.8%
5	2895	27.4%	49.7%	3983	72.7%
4	1983	18.8%	68.4%	6520	30.4%
3	2211	20.9%	89.4%	40891	5.4%
2	954	9.0%	98.4%	180324	0.5%
1	164	1.6%	<100%	601370	<0.1%
0	4	<0.1%	100%	350987	<0.1%
Total	10,559	100%	100%	1,186,526	0.9%

- Percentages may not add due to rounding.

By requiring at least four matching variables to be in full agreement, 68.4 percent of these A.C.E. duplicates were identified. On the other hand, when only four of the six variables fully agreed, only 30.4 percent of the total links identified by this criteria were A.C.E. Revision II duplicates. Note that it was tempting to require only three variables to be in full agreement since this increased the number of A.C.E. duplicates by 20 percent. However, the number of false matches greatly increased.

Table 5.3 shows introducing a minimum total score greatly increased the density of A.C.E. links identified. Note that some A.C.E. duplicate links were dropped by using this criteria. This was a consequence of applying rules that reduce the false link rate.

Table 5.3: Distribution of A.C.E. and Total Links within A.C.E. Clusters
(Only Include Links with Score >= 4.7)

Number of Variables in Full Agreement	A.C.E. Links	Total Links	Percent of A.C.E. Links in Row
6	2348	2451	95.8%
5	2868	3763	76.2%
4	1680	2670	62.9%
3	0	0	n/a
2	0	0	n/a
1	0	0	n/a
0	0	0	n/a
Total	6896	8884	77.6%

5.3.3 Second Stage: Household-Level Matching

The second stage of the matching was restricted to the household population. The person links from the first stage established a link between two housing units. The second stage was a statistical match of the household members from the two housing units. A sample household was included in the second stage multiple times if the sample household had persons with links to multiple census households in the first stage. This was the same approach used for the ESCAP II work.

The matching variables were the same as the first stage: first name, last name, middle initial, month of birth, day of birth, and age. Table 5.4 gives the matching parameters. The data in this table has similar meaning as that for the first stage in Table 5.1. Using a subset of the first-stage links, the second-stage matching parameters were derived using the EM algorithm as described in Winkler (1995). We anticipated that these parameters would be more discriminating that the set used for the ESCAP II study.

Table 5.4: Second-Stage Matching Parameters

Matching Variables	Type of Comparison	Matching Parameters		Match Score	
		Agreement Weight (m)	Disagreement Weight (u)	Agreement ln(m/u)	Disagreement ln((1-m)/(1-u))
First Name	String (uo)	0.9500	0.0125	4.3307	-2.9832
Last Name	String (uo)	0.9600	0.5700	0.5213	-2.3749
Middle Initial	Exact	0.0840	0.0220	1.3398	-0.0655
Month of Birth	Exact	0.6000	0.0600	2.3026	-0.8544
Day of Birth	Exact	0.3000	0.0200	2.7081	-0.3365
Computed Age	Age (p)	0.9750	0.1325	1.9959	-3.5467
Total				13.1984	-4.1948

Since the first stage established a link between two housing units, first name had more discriminating power than last name in the second stage. When first name fully agreed, it contributed 4.3307 toward the total score while last name only contributed 0.5213 when it was in full agreement. Further, month of birth and day of birth were more powerful than age. This was expected since adults in a housing unit often have similar ages but not the same month and day of birth.

The Statistical Research Division Record Linkage software, Winkler (1999), was used for the second stage. Each sample record was linked to only one census record within the household, one-to-one matching. There was no additional blocking criteria beyond household; all possible links within household were be compared. Each link had a total match score ranging from -4.1948 to 13.1984. This second-stage match score was used for the modeling. All links with a second-stage match score greater than 0.3419 were retained as input to the modeling.

5.3.4 Reverse Name 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 since the variable match scores for first and last name disagreed for both the first and second stage. 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. When both methods identified the same duplicate link, the higher of the two match scores was retained and used in the modeling.

5.4 Modeling Links

Since the goal of this study was to provide duplicate information to be used in A.C.E. Revision II estimates, it was important to provide a measure of confidence that two linked records were duplicates that can be incorporated into the estimation methodology. Consequently, modeling efforts focused on methods for estimating a probability that the two linked records were duplicate enumerations. An advantage of duplicate probabilities over the Poisson model weights used in ESCAP II was 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.

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 is 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 yield an estimated duplicate probability for the linked records from the statistical matching of the E and P samples to the census.

5.4.1 Statistical Match Probability

The statistical match modeling was used when there were two or more duplicate links resulting from 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 housing unit to census housing unit match, a set of match scores was observed. For any resulting set of match scores, a probability of not observing this set of match scores, Pr(NT), was estimated for each link within the sample household. The higher this probability, the more likely that the set of linked records in the household were duplicates.

Since a sample housing unit could have been matched to more than one census housing unit during the second stage, there were multiple sets of duplicate links and match scores for each sample housing unit. Each set of duplicate links for a sample housing unit was assigned a separate Pr(NT) since the match scores differ for each matching attempt. Further, the Pr(NT) for each set of duplicate links for a sample housing unit varied because of the geographic distance of the duplicate links. From the Appendix, the Pr(NT) was estimated by

$$\Pr(NT) = \left[1 - \prod_{d=1}^{p} \Pr\left(X_d \geq x_d\right)\right]^n$$

where

$\Pr(X_d \geq x_d)$ was the probability of getting a match score X_d that is greater or equal to x_d,
p was the number of duplicate links in the sample household, and
n is the number of census housing units the sample household was matched with in the second stage within a geographic area.

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 entire 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 geographical levels were block, tract, same county (outside tract), same state (outside county), different state.

For the E sample, this analysis was done at the E-sample household level. For the P sample, a household consisted of any combination of nonmovers, outmovers, and inmovers. To account for this, the duplicate links were analyzed separately by mover status when looking at patterns of match scores.

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. Table 5.5 shows the minimum value of Pr(NT) for assigning a statistical match duplicate probability of 1 for E and P samples.

Table 5.5: Minimum Value for Assigning Statistical Match Probability

Geographic Distance of Linked Records	Minimum Pr(NT)	
	E Sample	P Sample
Same Block	0.00	0.25
Same Tract (different block)	0.70	0.35
Same County (different tract)	0.97	0.60
Same State (different county)	0.97	0.60
Different State	0.97	0.60

Duplicate links with a Pr(NT) greater than or equal to the minimum value in Table 5.5 were assigned a statistical match duplicate probability of 1. All other links were assigned a statistical match duplicate probability of 0.

5.4.2 Exact Match Probability

Given exact matching of the census to itself, duplicate probabilities were assigned to linked records by taking into account the overall distribution of births, frequency of names and population size in a specific geographic area. Duplicate probabilities were computed separately by links within county, links within state and different county, and different states. Further, duplicate links were modeled separately by how common the last name was as well as separately by Hispanic names. Fay (2002b) gives the model and preliminary results. The following are excerpts from Fay (2002b) to give the reader a general idea of the approach.

Like the Poisson model, the new approach uses frequencies of occurrences of combinations of first and last name. The result is an estimated probability of duplication for most matches except for matches of frequently occurring names, where the probability of duplication is low and difficult to estimate with high relative precision.

This work results in a series of probability models, with parameters that can be estimated statistically from observed census data. A core model characterizes probabilities of duplication, triple enumeration (apparent enumeration of the same person three times), and other forms of multiple enumeration within a given geographic area. The other models account for duplication across domain.

The first part of the core model expresses the probability of coincidentally sharing a birthday. A second set of expressions, a model for census duplication, is built on top of the model for coincidental sharing of date of birth. The core model combines the two models to account for observed patterns of exact computer matches of census enumerations. The core model provides a basis to estimate a probability that a given computer match links the same person instead of two persons coincidentally sharing a birthday. An approximate argument allows the core model to be extended to nested geographic categories, such as (1) counties, (2) other counties within state, and (3) other states.

The result of the exact match model is a duplicate probability greater than or equal to zero, but less than one for census records that agree exactly on first name, last name, month and day of birth and two-year age intervals.

5.4.3 Combining the Two Models

The two approaches were combined to give one duplicate probability to each E- and P-sample duplicate link. Table 5.6 summarizes the results of combining of the two models. 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 as indicated by the shaded cells in Table 5.6. For non-exact matches, this was always from the statistical match modeling.

For exact matches in sample households with two or more persons, adjustments were made to account for the integration of these two methods. The exact match probabilities were determined conditionally, requiring a downward adjustment of the exact probabilities for the links which the statistical match modeling assigned a probability of zero. The amount of the downward adjustment was based on the upward adjustment made when using the statistical match probability of one instead of the exact match probability.

Table 5.6 Combining the Two Modeling Results

Type of Link	Size of Sample HU	Type of Match	Statistical Match Prob.	Exact Match Prob.
Housing Unit	1	Exact	-	[0, 1)
		NonExact	-	-
	2+	Exact	1	[0, 1)
		Exact	0	[0, 1)
		NonExact	1	-
		NonExact	0	-
Group Quarter		Exact	-	[0, 1)
		NonExact	-	-

- Modeling did not assign a value.

The results of this modeling provided, for each full E- and P-sample person who links to a census person outside the A.C.E. search area, the probability that they are in fact the same person. These probabilities, referred to as p_t in chapter 6, were used in obtaining A.C.E. Revision II estimates.

5.4.4 Reinstated and Deleted Census Records

The duplicate modeling accounted for reinstated and deleted census housing units by generating separate duplicate probabilities. One duplicate probability was computed when the sample records linked to reinstated and deleted census records. A second duplicate probability was computed without considering links to reinstated and deleted records. Under this second scenario, any links to reinstated or deleted records were assigned a duplicate probability of zero. For the exact match modeling, separate probabilities were computed based on population distributions with and without the reinstated and deleted records.

5.5 Assessment of Links

Throughout the development of this plan, the A.C.E. duplicate links found during production were the benchmark used to gauge whether the matching algorithm did a good job of finding true duplicates and minimizing the number of false links found within the block cluster.

The plan used the same method as Fay (2001, 2002) for estimating efficiency for the ESCAP II study for the E sample. Basically, this estimate measured the effectiveness of identifying A.C.E. clerical duplicates within the A.C.E. sample area and accounting for the duplicate links to reinstated and deleted records which were out-of-scope for A.C.E. Instead of getting one overall measure, measures were computed for various levels of detail including size of sample household and number of links between the units.

5.6 Forming Estimates of Duplicates

Estimates of census duplicates were formed by summing the product of the sampling weight for the E-sample person, the duplicate probability, and the multiplicity factor. Since we matched a sample of the census (E sample) to the census, a naive approach would treat each duplicate link of A to B as one duplicate. However, if we'd drawn a different sample, we could have found the B to A link. Applying a multiplicity fact of ½ in this simple case, ensured that we only treat this as one duplicate. See Mule (2002) for more details on the computation of the multiplicity factor.

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.

Feldpausch, Roxanne (2001), “ESCAP II: Census Person Duplication and the Corresponding A.C.E. Enumeration Status,” Executive Steering Committee for A.C.E. Policy II, Report 6, October 1, 2001.

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.

Probability of Not Observing a Set of Match Scores

Each E-sample household had a set of duplicate links to a particular census household. Each duplicate link had a corresponding overall match score from the second-stage matching resulting in a pattern of match scores for the sample household. How did we assess whether this observed set of match scores occurred because the links were duplicates or because the records had characteristics in common but are different people.

Objective: To estimate the probability of not observing this set of match scores or better for this E-sample household.

The hypothesis is that the higher the probability of not observing this set of match scores or better, then the more likely the links represent duplicate enumerations.

Let's say that a particular E-sample household has p duplicate links, $p \geq 2$, with observed match scores, x_1, x_2, ..., x_p.

Let Pr(NT) = the probability of not observing the set of match scores or better, ($X_1 \geq x_1$, $X_2 \geq x_2$, ..., $X_p \geq x_p$).

$$\Pr(NT) = \left[1 - \Pr\left(X_1 \geq x_1, X_2 \geq x_2, \ldots., X_p \geq x_p\right)\right]^n \qquad (1)$$

where n was the number of different census housing units that the E-sample housing unit was linked to during the second-stage match. This took into account that the more times you matched the E-sample housing unit to different housing units, the greater chance of obtaining this outcome.

We assumed independence of the individual match scores, X_1, X_2, ..., X_p, since the second-stage matching parameters gave more emphasis to first name rather than last name. Further, the parameters gave more emphasis to month and day of birth rather than age.

$$\Pr(NT) = \left[1 - \prod_{d=1}^{p} \Pr\left(X_d \geq x_d\right)\right]^n \qquad (2)$$

The probability of getting a match score X_d that was greater or equal to x_d, $\Pr(X_d \geq x_d)$, was obtained from the empirical distribution of second-stage match scores.

The probability in (2) was used for the P sample households as well.