PROBABILISTIC MODELS FOR DETECTING CENSUS PERSON DUPLICATION

Robert E. Fay[1]
U.S. Census Bureau, 4700 Silver Hill Rd. Stop 9001, Washington, DC 20233-9001

KEY WORDS: Census duplication, Erroneous enumeration, Computer matching, Census 2000

Abstract. The net undercount of the population by the decennial census arises from the balance between: (1) omissions of persons the census should count but misses, and (2) erroneous enumerations the census incorrectly includes. Duplication is a form of erroneous enumeration; typically a duplicated person is counted correctly where they should be but also incorrectly elsewhere. Coverage measurement surveys, such as the 2000 Accuracy and Coverage Evaluation (A.C.E.), must account for the effect of census duplication in order to accurately measure the net undercount.

Because Census 2000 captured both names and dates of birth for most respondents, computer matching can identify possible duplicate enumerations. Exact matches on name and date of birth appear to be likely evidence for duplication, but such matches include persons with the same name coincidentally sharing birthdays. The paper develops probabilistic expressions for exact matches for the relative effects of duplication and coincidental sharing of birthday.

1. Introduction

In October 2001, the U.S. Census Bureau announced a decision to publish all remaining Census 2000 data products without statistically adjusting for census undercount. The Census Bureau judged the 2000 coverage survey, the Accuracy and Coverage Evaluation (A.C.E.), to be too flawed for census adjustment. In most respects, evaluation studies found the A.C.E. to be generally well executed and successful, but key studies showed the A.C.E. to be seriously deficient in its measurement of erroneous enumerations in the census.

There were two primary sources of evidence for this finding. First, a reinterview study, the Measurement Error Reinterview (MER) (Adams and Krejsa 2001), uncovered a substantial number of erroneous enumerations missed by the original A.C.E. interview. That is, the number of A.C.E. correct enumerations the MER reclassified as erroneous far exceeded the number of A.C.E. erroneous enumerations the MER reclassified as correct. Thus, the MER results implied that the A.C.E. underestimated erroneous enumerations.

Second, the Person Duplicate Study (Mule 2001) uncovered a substantial number of duplicated persons in the census. This computer matching study was feasible because Census 2000 was the first U.S. census to capture names and birthdates in computer-readable form nationally. Using the findings of the Person Duplicate Study, Feldpausch (2001) showed that the A.C.E. correctly identified some of the duplicates as erroneous enumerations, but far fewer than it should have. Thus, the Person Duplicate Study also suggested that the A.C.E. underestimated erroneous enumerations.

The Merged MER/Duplicate Study (Fay 2001a, 2002) combined the data from the MER and Person Duplicate Study to obtain a preliminary estimate of the combined effect of A.C.E. errors on the A.C.E. estimate of erroneous enumerations. In turn, Thompson, Waite, and Fay (2001) summarized the possible effect on the A.C.E. estimates of population and provided a "Revised Early Approximation" to a possible future revision of the A.C.E.

Many of the evaluation reports (including Fay 2001a, 2002) recommended specific additional work. Currently, the Census Bureau is attempting to revise its 2001 analysis, including reviewing additional MER cases clerically. The research effort will also examine and possibly correct for other issues affecting A.C.E. accuracy besides the measurement of erroneous enumerations.

One principal focus of the new work is to re-examine census duplication. The Person Duplicate Study identified duplicates through computer matching employing a complex set of strategies rather than a single algorithm. The study used two stages of computer matching: the first stage at the person level and the second at the household level. The first stage can be characterized as exact matching—persons matched at this stage had the same first and last name, month and day of birth, and generally age (Mule 2001). Intuition suggests that almost all such matches agreeing on age at a lower geographic level, such as within county, are likely to the same person enumerated twice. (The empirical research reported here essentially supports intuition in this regard.) At higher geographic levels, however, exact matching suffers from the "Linda Smith" problem. If the national number of Linda Smiths in a given birth year approximates or exceeds 365, then intuition suggests that many exact matches across state boundaries will reflect coincidental sharing of birthday. (Indeed, the number 365 is far larger than required for coincidence to be an important consideration.) Intuition is clearly at risk in judging the degree to which less frequently occurring names, such as "Robert Fay," are affected by the "Linda Smith" problem, and formal analysis is required.

The Person Duplicate Study employed "Poisson weights" to address the "Linda Smith" problem. The basic approach conditioned on the frequency of occurrence of the first/last name combination within a given birth year. The author originally proposed the Poisson weights. Mule (2001) provided the most readily available account of the approach but did not detail the underlying rationale. Although bounded above at 1.0, the Poisson weights are not true probabilities, and they can in fact be negative. In practice, researchers (Mule 2001, Feldpausch 2001, Fay 2001a and 2002) dropped matches with low weights; for example, Fay (2002) excluded matches

with weights below .98, arguing that the analysis would be a lower bound on the effect of undetected duplication on the accuracy of the A.C.E. In a discussion section, Fay (2002) remarked on the importance of refining the Poisson model.

This paper presents a new approach to replace rather than refine the Poisson model. 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.

The new 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 domains.

The first part of the core model expresses the probability of coincidentally sharing a birthday. (Birthday coincidences are not as frequent as it may at first seem. For example, the results here indicate that most combinations of names and year of birth in New York State are unique.) 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.

For highly common names, the probability that a computer match is a true duplicate can be quite small, particularly between states. Treatment of these cases remains an open question, but a possible approach is to estimate an average frequency of census duplications from less common names rather than to attempt to estimate small probabilities. This aspect may be construed as a missing-data problem.

The core model is unsuited to address other questions, however. For example, studies suggest that the rate of duplication between the group quarters population (college dormitories, nursing homes, prisons, etc.) and the balance of the population in households may have been on the order of 10% of the group quarters population, approximately an order of magnitude larger than duplication within the housing unit population. A set of additional probabilistic models is described to allow extensions to different domains.

Although primarily methodological, the paper also presents a preliminary application of the core model to New York State. The application illustrates issues to be encountered in a national match.

2. Previous Use of Exact Person Matching

As previously noted, Census 2000 in the U.S. was the first to capture names nationally in computer-readable form. (A subset of names was keyed from the 1990 census for use by the 1990 Post-Enumeration Survey, but only in sample and nearby blocks. Word and Colby (1996) later studied frequency of occurrence of last names in these data.) In 2000, names were captured for three primary reasons: (1) for use in the primary selection algorithm that determined whether persons enumerated through different response modes were the same, (2) for archival purposes, (3) as in 1990, for use in the coverage measurement survey, the A.C.E. Census 2000 also asked for month, day, and year of birth, and approximately 90% of respondents provided this information.

The Housing Unit Duplication Operations (Fay 2001b) were designed and implemented in 2000 to remove duplicated housing units and the associated people from the census data prior to determining the final count for public release. The operations removed approximately 3.6 million from the preliminary population count. An important component was to identify potentially duplicated housing units through exact person matching on first and last name and date of birth. The exact matching was restricted to relatively small geographic areas, never beyond state boundaries. After the first stage of exact matching, a second stage considered how similar the other people in the potentially linked household were. The issue of coincidental sharing of date of birth was not considered. The Housing Unit Duplication Operations also temporarily set aside the enumerations of an additional 2.3 million persons, but then reinstated them into the final census count. For reasons of timing, the A.C.E. universe excluded these enumerations. Many of these enumerations duplicated enumerations in the A.C.E. universe (Mule 2001). The empirical results presented in this paper exclude duplications to the 2.3 million.

Like the Housing Unit Duplication Operations, the Person Duplicate Study included a first stage of exact matching at the person level and a second stage at the household level. The Person Duplicate Study considered exact matches at different geographic levels, including nationally. The Person Duplicate Study was the first to explicitly address the “Linda Smith” problem. At some geographic levels, such as county, all first-stage matches were accepted, but at other levels the possible effect of coincidental sharing of date of birth was represented through a weight based on a Poisson model (Mule 2001). Some of the first-stage matching rules were modified, such as accepting age agreement within one year as a match.

In addition, the matching rules for the Person Duplicate Study employed global geographic distinctions, with different rules for duplicates across states and within states. The new work continues to divide matching geographically, but assigns probabilities of duplication in a way that recognizes that some counties in New York are more populous than small states such as Delaware.

3. The Core Model for Exact Matching

For purposes of the A.C.E., we are interested in duplicates within the A.C.E. universe in the census, which is somewhat smaller than the complete census as enumerated. Let the population or universe, *u*, denote persons enumerated in Census 2000 one or more times with a complete name, month, day, and year of birth reported. The universe is thus a set of people rather than census enumerations.

The popularity of both first and last names and combinations of them varies widely. Generally, for any given name the distribution of month and day of birth within a year of birth is approximately uniform, and the model assumes a uniform distribution.

It is natural to consider each birth year separately, but slightly more information is available by pooling neighboring birth years into age intervals. Let *age* be the age in years on Census Day, April 1, 2000. Age thus defines one possible age interval of persons born during one of $d = 365$ possible days, ignoring the effect of leap years. More generally, it is possible to consider alternative age intervals defined in integral years, with d a multiple of 365. For $d = 730$, 0- and 1-year-olds are grouped into the first 2-year interval, 2- and 3-year-olds into the second, *etc.* Most empirical results reported in this paper are based on $d = 730$.

Model for birthday coincidence. For any given geographic area, *A*, choice of age interval, *d*, and specific age interval, *i*, let *x1, x2, x3, x4,...* denote the fixed but unknown distribution of births during the year. The model for coincidental sharing of birthday accounts for the counts, $y_{a,b}$, of shared birthdays in the population.

More formally, for a given geographic area, *A*; choice of age intervals, *d*; and specific age interval, *i*; let $S_{a,b}(A,d,i)$, or more simply $S_{a,b}$, be a set of (*name*, *age interval*) pairs in *A* such that for each (*name*, *age interval*) in $S_{a,b}$, there are (1) exactly *a* persons with name, *name* and age in *i*, and (2) *b* pairs with identical birthdays (including identical year of birth). Similarly, let $S_{a,b,c}$ be a set of (*name*, *age interval*) pairs in *A* such that for each (*name*, *age interval*) in $S_{a,b,c}$, there are (1) exactly *a* persons with name, *name* and age in *i;* (2) *b* pairs with identical birthdays; and (3) *c* triplets with identical birthdays. More complex sets, $S_{a,b,c,d}$, may be similarly defined, although the primary interest will be in the combinations up through triplets. Note that $S_{a,b} = S_{a,b,0}$. Let S_a be a set of (*name*, *age interval*) pairs in *A* such that for each (*name*, *age interval*) in $S_{a,b}$, there are exactly *a* persons with name, *name*. For example, S_3 comprises the disjoint sets $S_{3,0}$, $S_{3,1}$, and $S_{3,0,1}$.

Let $y_{a,b}$ denote the number of (*name*, *age*) combinations in set $S_{a,b}$; $y_{a,b,c}$, the number of (*name*, *age*) combinations in set $S_{a,b,c}$, etc. For example,

$$y_3 = y_{3,0} + y_{3,1} + y_{3,0,1}$$

The expected values, $x_{a,b} = \mathrm{E}(y_{a,b})$, under this model, conditioning on the number of births for the given age, are given by

$$x_{1,0} = x_1$$

$$x_{2,0} = ((d\text{-}1)/d)\, x_2$$

$$x_{2,1} = (1/d)\, x_2$$

$$x_{3,0} = ((d\text{-}1)(d\text{-}2)/d^2)\, x_3$$

$$x_{3,1} = 3\,((d\text{-}1)/d^2)\, x_3$$

$$x_{3,0,1} = (1/d^2)\, x_3$$

...

The appendix provides detailed expressions for all possible terms up to $x_{5,0,0,0,1}$, and for $x_{6,0}$, $x_{6,1}$, $x_{7,0}$, *and* $x_{7,1}$.

As an example, a population with common last names might be distributed within a county according to the following distribution (with rows x_5 and above not shown):

a	x_a	$x_{a,0}$	$x_{a,1}$	other
1	170,680	170,680		
2	6,841	6,832	9	
3	898	894	4	0
4	203	201	2	0
5+	…	…	…	…

Model for census duplication. Let $S_{a,b}^{*}(A)$ be a set of (*name*, *age*) pairs enumerated in universe *u* in the census in *A* such that for each (*name*, *age*) in $S_{a,b}^{*}(A)$, there are exactly *a* census enumerations with name, *name* and age, *age*, and *b* pairs with identically reported birthdays. Set $S_{a,b}^{*}$ differs from $S_{a,b}$ by counting census enumerations rather than persons. Define $S_{a,b,c}^{*}$, S_a^{*} , $y_{a,b}^{*}$, etc., similarly.

To start with a simple instance, assume that there is a fixed probability, *p*, that an individual person will be duplicated in the census in universe *u* in area *A*. Let p_3 denote the probability that a person is included three times in universe *u*. Empirically, $p >> p_3$, and multiple inclusion probabilities p_4 ... will be omitted from the development. Under this simplification, let $q = 1 - p - p_3$ represent the probability that the person is counted only once. As noted below, let p^{*} , p_3^{*} , and q^{*} represent analogous probabilities specific to y_1. Let x_a^{*} , $x_{a,b}^{*}$, etc., represent the expected values of y_a^{*} , $y_{a,b}^{*}$.

$$q^{*} = 1 - p^{*} - p_3^{*}$$

$$p_3 = (p/p^{*})\, p_3^{*}$$

The model for the expected values, x_a^{*} , $x_{a,b}^{*}$, of the observed y_a^{*} , $y_{a,b}^{*}$, is given by

$$x_1^{*} = q^{*} y_1$$

$$x_{2,0}^{*} = q^2 y_{2,0}$$

$$x_{2,1}^{*} = q^2 y_{2,1} + p^{*} y_1$$

$$x_{3,0}^{*} = q^3 y_{3,0}$$

$$x_{3,1}^{*} = q^3 y_{3,1} + 2 p q y_{2,0}$$

$$x_{3,0,1}^{*} = q^3 y_{3,0,1} + 2 p q y_{2,1} + p_3^{*} y_1$$

...

Suppose that y_1 includes some proportion, t, of misspelled names such that they would never match exactly to another enumeration. Under simplifying assumptions, p^* will be reduced relative to p according to

$$p^{*} = (1 - t) p$$

and

$$p_3^{*} = (1 - t) p_3$$

This feature predicts smaller values for p^* than p, and separate fitting of p and p^* empirically confirms this.

In summary, the underlying distribution for the number of births, *x1, x2, x3, x4,...,* to population u is treated as fixed but unknown. The two models are:

1. Model for birthday coincidence, for population u, predicts expected values, $x_{a,b} = E(y_{a,b})$, conditional on the distribution of births *x1, x2, x3, x4,….*
2. Model for census duplication, given the population u, and counts of birthday coincidences $y_{a,b},\ldots$, the model predicts the expected values $x_{a,b}^{*} = E(y_{a,b}^{*})$, of the observed patterns, $y_{a,b}^{*}$, of census-reported persons. Matches represent a mixture of coincidental birthdays and census duplications.

The overall objective is to analyze the observed $y_a^{*}, y_{a,b}^{*}$, etc., to estimate the number of duplicates in the census within universe u. Combining the models

$$x_1^{*} = q^{*} x_1$$

$$x_{2,0}^{*} = ((d\text{-}1)/d)\, q^2 x_2$$

$$x_{2,1}^{*} = (1/d)\, q^2 x_2 + p^{*} x_1$$

$$x_{3,0}^{*} = ((d\text{-}1)(d\text{-}2)/d^2)\, q^3 x_3$$

$$x_{3,1}^{*} = 3\,((d\text{-}1)/d^2)\, q^3 x_3 + 2\,((d\text{-}1)/d)\, p\, q\, x_2$$

$$x_{3,0,1}^{*} = (1/d^2)\, q^3 x_3 + 2\,(1/d)\, p\, q\, x_2 + p_3^{*} x_1$$

Continuing the preceding example, the observed census frequencies might appear as follows

a	y_a^{*}	$y_{1,0}^{*}$	$y_{2,1}^{*}$	other
1	169,199	169,199		
2	8,180	6,706	1,474	
3	1,013	870	127	16
4	226	194	29	3
5+	…	…	…	…

Calculation of probabilities of duplication. The proposed approach is to solve for p, q, etc. based on cells y_1^{*}, $y_{2,1}^{*}$, $y_{2,0}^{*}$, $y_{3,0}^{*}$, $y_{3,1}^{*}$, and $y_{3,0,1}^{*}$. Values of $x_4, x_5,\ldots$ can be estimated from $x_{4,0}^{*}$, $x_{5,0}^{*}$, and the estimate of q. Conditional probabilities of duplication (compared to coincidental agreement) are estimated by expressions such as

$$p_{2,1} = p^{*} x_1 / ((1/d)\, q^2 x_2 + p^{*} x_1)$$

$$p_{3,1} = 2\,((d\text{-}1)/d)\, p\, q\, x_2 / (3\,((d\text{-}1)/d^2)\, q^3 x_3 + 2\,((d\text{-}1)/d)\, p\, q\, x_2)$$

Such probabilities are average values over u. Conditioning on additional characteristics, such as the specific last name, may alter the probabilities somewhat.

Extension to nested situations. The preceding equations state the model in a form suitable for application at one geographic level, counties, for example. When the model is fitted first at a county level and then at a state level to detect duplicates between counties, a modification is necessary to prevent census duplication within county from distorting the probabilities of duplication between counties. A simple modification is to count all occurrences of a name and date of birth within a county only once. The core model can then be applied to the counts of observed pattern of birthdays across counties and days. This approach slightly distorts the assumptions of the core model, however. For example, triplets are extremely unlikely to be observed in Delaware, since a triplet requires an observation in each of the three counties of Delaware. On the other hand, the most populous of the three counties, New Castle, contributes disproportionately to the number of unique births in the interval. The models in the next section avoid this effect.

4. Probabilistic Models for Matching Across Domains

Preliminary empirical evidence suggests that true triplicate enumeration, enumeration of the same person three times, occurred much more rarely than duplication in Census 2000, at a rate roughly two orders of magnitude less. Consequently, models focused primarily on duplication can provide a useful approximation, even if they ignore triplicate enumeration as a separate possibility.

Previous work indicates a relatively high rate of duplication between the group quarters population (college dormitories, nursing homes, prisons, *etc.*) and the remaining household population living in housing units (Feldpausch 2001). Unfortunately, the core model combines these populations and treats them as a homogeneous population within a geographic area. Instead, for any given county, the census enumerations

may be divided in the following manner:

	Group Quarters	Household
Different states	GQ_n	HH_n
Diff. counties within state	GQ_s	HH_s
Within county	GQ_c	HH_c

Consider first the problem of duplication within state. The proposed approach introduces models 2, 3, and 4 and divides the problem in this manner:

Matching	Probabilistic model
Within HH_c	Core model
Within $GQ_c + GQ_s$	Core model
Between HH_c and HH_s	Model 2
Between HH_c and GQ_c	Model 3
Between HH_c and GQ_s	Model 4

The core model is applied at the lowest level of geography considered for the population group. Because the size of the group quarters population is small relative to the housing unit population, and because the rate of duplication within the group quarters population is low, for this population it is statistically advantageous to apply the core model at the state level without regard to county.

The new models each consider duplication between groups. Each is based on a straightforward counting argument. Model 3 is the most easily explained. If the number of unique dates of birth in a county's group quarters population is u_{gq} and the number of unique dates in the household population is u_{hh}, then the number of possible pairings between the two populations is $u_{gq}\, u_{hh}$, and the expected number of coincidental agreements is $(u_{gq}\, u_{hh})/d$. By summing the actual pairings and the expected number coincidental agreements over a suitable set of cases, it is possible to estimate from the observed data the average probability that a match is a duplicate. To obtain better conditional probabilities, the calculation can be conditioned on a measure of size, such as u_{hh}. (An initial empirical study on a subpopulation of New York favors the dichotomy $u_{hh} = 1$ vs. $u_{hh} > 1$.)

Model 2 requires a similar counting argument. If counties are indexed by c, let the total number of unique births be u_{hhc} and their sum $u_{hh.} = \sum_c u_{hhc}$. The number of possible pairs, n_p, between counties is

$$n_p = \binom{u_{hh.}}{2} - \sum_{u_{hhc}>1} \binom{u_{hhc}}{2}.$$

For example, if 3 unique birthdays for a name and age interval are observed in New Castle County, Delaware, and 1 unique birth is observed in another county, $n_p = 3$ for model 2. Extension of the core model to between county births instead effectively assumes $n_p = 6$, thus overestimating coincidental births.

Finally, the required counting argument for model 4 is

$$n_p = u_{hh.}u_{gq.} - \sum_c u_{hhc}u_{gqc}.$$

Name reversals, recording first name as last name and last as first on the census form, occur frequently enough to be of some interest. The core model was unsuitable to estimate probabilities of duplication for such reversals, but models 3 and 4 are readily applicable.

5. Illustrative Analysis for New York State

A preliminary analysis of New York State with the core model illustrates the potential use of the model in other states and for national matching.

The model may be applied to a partition of the census enumerations. For example, because matching is exact, last names were divided into three groups: (1) the 639 most frequently occurring heavily Hispanic surnames (Word and Colby 1996), (2) the 353 most frequently occurring rarely Hispanic surnames (for example, Smith, Johnson, Jones, Brown, *etc.*) (Word and Colby 1996), and (3) all others. The Census 2000 population count for the April 1, 2000 resident population of New York is 18,976,457. The study population, with sufficient name and date of birth information is 16,650,346, divided into 1,538,024, 2,834,034, and 12,278,288, for the three groups, respectively.

Table 1 summarizes the estimates from the model. Although estimated probabilities are less for frequently occurring names, the model indicates that almost all matches within county identify true census duplicates. The model shows that duplication between counties is more likely to be mixed with coincidental sharing of birthdays.

The definition of the universe, *u*, did not require reporting of middle initial (MI), and this characteristic appears irregularly. At some geographic levels, Mule (2001) excluded census matches with disagreeing MI when both forms reported one, but retained matches when one or both enumerations lacked MI. Table 1 shows results for the subset of cases where MI was reported for both. Middle initial agrees less frequently than suggested by the estimated probabilities. But MI was inconsistent a few percent of the time even for duplicates within A.C.E. clusters confirmed clerically. Hence, the somewhat reduced consistency for MI is not of immediate concern. The results for MI vary quite smoothly with variation in estimated duplicate probabilities across the table.

6. Discussion

These preliminary findings suggest additional areas for investigation. The model is built on an assumption of uniform births across the age interval. The effect of naturally occurring variation in births and parents' selection of first names merits further investigation.

Although matching is described as exact, other researchers are currently investigating the possible benefits from editing names in specific ways and using record linkage based on the Fellegi-Sunter (1969) algorithm. It is not yet clear whether estimating

the Fellegi-Sunter model produces estimated probabilities that are sufficiently well-calibrated to use in estimation, and the current project may provide an opportunity to compare the approaches.

The second phase of matching, which examines similarities of entries for households linked by one or more exact matches, merits further investigation, and information from the second phase could inform results from the first. But the development of estimated probabilities from the first stage could guide procedures in the second phase.

Note: 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 author wishes to thank William Bell, Aref Dajani, Tom Mule, and William Winkler for their comments. The title is revised from "The Effect of Person Duplication in Census 2000 on the Population Undercount."

References

Adams, Tamara and Krejsa, Elizabeth A. (2001), "ESCAP II: Results of the Person Followup and Evaluation Followup Forms Review," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 24, DSSD Census 2000 Procedures and Operations Memorandum Series #T-17, Oct. 16, 2001, U.S. Census Bureau, at http://www.census.gov/dmd/www/ReportRec2.htm.

Fay, Robert E. (2001a), "Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 9, preliminary version Oct. 26, 2001, at http://www.census.gov/dmd/www/ReportRec2.htm.

_____(2001b), "The 2000 Housing Unit Duplication Operations and Their Effect on the Accuracy of the Population Count," presented at the Joint Statistical Meetings, Atlanta, GA, Aug. 5-9, 2001, *2001 Proceedings of the Joint Statistical Meetings* on CD-ROM, American Statistical Association, Alexandria, VA.

_____(2002), "Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 9, revised version Mar. 27, 2002, at http://www.census.gov/dmd/www/ReportRec2.htm.

Feldpausch, Roxanne (2001), "Census Person Duplication and the Corresponding A.C.E. Enumeration Status," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 6, DSSD Census 2000 Procedures and Operations Memorandum Series #T-16, Oct. 12, 2001, U.S. Census Bureau, at http://www.census.gov/dmd/www/ReportRec2.htm.

Fellegi, Ivan P. and Sunter, Alan B. (1969), "A Theory of Record Linkage," *Journal of the American Statistical Association,* **64**, 1183-1210.

Mule, Thomas (2001), "Person Duplication in Census 2000," Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 20, DSSD Census 2000 Procedures and Operations Memorandum Series Q-71, Oct. 11, 2001, U.S. Census Bureau, available at http://www.census.gov/dmd/www/ReportRec2.htm.

Thompson, John H., Waite, Preston J., and Fay, Robert E. (2001), "Basis of 'Revised Early Approximation' of Undercounts Released Oct. 17, 2001," supplemental statement included in materials of the Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, available at http://www.census.gov/dmd/www/ReportRec2.htm.

Word, David L. and Perkins, R. Colby Jr. (1996), "Building a Spanish Surname List for the 1990's—A New Approach to an Old Problem," Technical Working Paper No. 13, Population Division, U.S. Census Bureau, March 1966.

Table 1. Preliminary analysis of census duplicates within New York State, based on exact matching only. The analysis uses 2-year intervals to compute the number of births in the interval. Results for all links are shown first, followed by those for links where middle initial (MI) is reported for both names. Counts are of combinations, not of persons.

# Census Births in Interval	Links	∑ prob	Aver. Prob.	Links w/MI	Rate MIs Agree	Aver. Prob w/MI
Links Within County Only						
2	106,612	106,318	.997	38,157	.928	.998
3	6,766	6,645	.982	1,953	.912	.983
4	1,272	1,133	.891	349	.851	.880
5	466	403	.865	128	.867	.894
6	222	191	.858	67	.761	.860
7	118	95	.803	29	.759	.802
Total 2-7	115,456	114,785	.994			
8+	291	(n/a)	(n/a)	90	.622	(n/a)
Links Between Counties Only						
2	43,736	42,881	.980	17,644	.931	.983
3	5,720	5,070	.886	2,327	.831	.889
4	2,427	1,888	.778	1,002	.742	.789
5	1,447	1,009	.697	606	.640	.697
6	1,038	645	.622	395	.630	.622
7	782	439	.562	340	.544	.556
Total 2-7	55,150	51,932	.942			
8+	5,499	(n/a)	(n/a)	2,451	.341	
Links Within County That Also Link Outside County						
2	776	774	.997	245	.931	.998
3	78	77	.984	20	.9	.980
4	24	22	.934	6	.8	.941
5	11	10	.896	1	1.0	1.000
6	9	8	.843	4	.75	.838
7	4	3	.815	1	0	.856
Total 2-7	902	894	.991			
8+	16			6		