# AN ANALYIS OF PERSON DUPLICATION IN CENSUS 2000

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

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

**Abstract.** Computer matching on name and date of birth can be used to identify duplicate enumerations of persons in Census 2000. Computer matching faces two important limitations. First, because name and date of birth are not always unique, computer matching links some enumerations together that do not represent duplicated persons. Second, as a consequence of inaccurate or missing names or dates of birth, computer matching fails to identify some duplicates.

Two previous papers presented probabilistic models to address the problem posed by coincidental sharing of date of birth by persons with the same name. The paper describes the application of the methods to Census 2000. Although estimates of duplication in Census 2000 are already available from a previous study, the earlier results are based on a sample whose size limits detailed analysis. Instead, this paper reports results from the full census.

In addition to applying the probabilistic models described previously, the paper investigates the degree of improvement in matching resulting from a series of edits of the reported names developed by other U.S. Census Bureau researchers. Because reduction of census duplication is a goal, the results should help in planning the 2010 Census.

## 1. Introduction

Census 2000 was the first U.S. census to incorporate the computer capture of names and dates of birth (month/day/year) as an integral part of the processing. In previous censuses, names were recorded and preserved on paper. The change to electronic capture helped to automate a number of census processes. As one example, the information was used as part of the Primary Selection Algorithm, the rules by which the number of valid enumerations was determined when more than one form was submitted for a census household. The Census 2000 coverage survey, the Accuracy and Coverage Evaluation (A.C.E.), also used the information in its matching operations.

In addition to the originally designed purposes, the capture of names and dates of birth had at least two other fortuitous consequences. First, the data were critical to the Housing Unit Duplication Operations (HUDO, Nash 2000, Fay 2001). Duplicate enumeration of some persons had been detected in previous U.S. censuses (Fay 2001 reviews evidence of duplication in the 1980 and 1990 censuses), but an examination of initial returns from Census 2000 suggested a particularly high rate of duplicate enumeration of persons in the same or nearby blocks. This geographically close duplication of persons was primarily due to duplication of housing units in the address register. During 2000, the HUDO was designed and added to the census to remove duplicated housing units and persons to the extent possible; the operation relied in large part on computer matching of names and birth dates locally to identify potential duplicates. Initially, almost 6.0 million enumerations of persons and their associated housing units were identified for potential deletion. These cases were then further screened with rules attempting to distinguish duplicated housing units from other housing units selected initially. Application of the rules deleted approximately 3.6 million enumerated persons permanently from the census (most of whom were clearly duplicate enumerations) and reinstated—that is, included in the final census files—approximately 2.3 million persons. The operation was concluded prior to the December 2000 release of the census count.

Second, computer matching of names and birth dates between the A.C.E. sample and the entire census provided data critical to the evaluation and reanalysis of the A.C.E. Mule and Fenstermaker (2003) review the final outcome of these efforts, and information on duplication is also reflected in numerous A.C.E documents.

Both of these operations actually involved two forms of computer matching, termed *exact* and *statistical* in this context. Exact matching refers to an exact agreement of name and date of birth, whereas statistical matching identifies close matches on the basis of similarity. Both HUDO and the duplicate work incorporated for the A.C.E. combined exact and statistical matching approaches.

The reanalysis of the A.C.E. required identification of duplicates across the entire country in contrast to the geographically narrow scope of the HUDO. The primary impact of the HUDO was within the same block or small neighborhood. In the HUDO, exact matches were interpreted as duplicate enumerations of persons, because the chance of two people having the same name and date of birth is virtually negligible locally. Rules for statistical matching could also be calibrated to a level appropriate for close geographic proximity and for the HUDO's primary purpose of identifying duplicated households in duplicated housing units, rather than simply duplicated persons.

The national geographic scope of the A.C.E. introduced new methodological challenges negligible in the HUDO. For the reanalysis of the A.C.E., the coincidental sharing of birthdays by two people with the same name is a critical issue in deciding the extent to which an exact match represents a duplicate enumeration, especially for enumerations occurring in different states. Similarly, thresholds for statistical matching require calibration to account for possible false matches.

Two previous methodological papers developed probabilistic models to address the problem. The first (Fay 2002) described methods to estimate the probability that a given exact match was a duplicate enumeration. The analysis was at the person level, ignoring information from other household members. The methods were incorporated into the final revision of the A.C.E. (Mule and Fenstermaker 2003) for some classes of duplications. The method considers both (1) the specific geographic level (for example, the specific county for duplicates occurring in the same county or the specific state for duplicates between counties in the same state) and (2) the relative frequency of the first/last name pair in the geographic area. (For example, in determining whether Jose Garcia born on the same day and enumerated in two different New Hampshire counties, the model considers the frequency of the name Jose Garcia in New Hampshire rather than nationally. For an exact match of a Jose Garcia in New Hampshire to one in California, the model considers the national frequency of the name Jose Garcia, however.)

The second paper (Fay 2003) developed methods to include household information. These results were not used, and indeed were not available in time, for the final A.C.E. revision in 2002. In the special case that two census households of 2 or more persons each are linked by a single exact match, it is possible to use statistical matching techniques to identify further potential duplicates between the remaining members. More generally, households linked by one or more exact matches can be considered for further statistical matching as long as unmatched persons remain in both households. But in the A.C.E. and the research to be reported here, exact matching was used exclusively in two situations: (1) matching household members to group quarters populations, since true matches almost always involved one person at a time, and (2) matching 1-person households to other households.

This paper describes the application of the methods from the previous two papers to Census 2000. The availability of data for the whole census will support detailed analyses of aspects of duplication in Census 2000 in far greater detail than studies based on the A.C.E. sample, whose sample size permits broad but not detailed analyses. The approach also simplifies the further study of census duplication, because it disentangles the phenomenon of census duplication from the logical complexities of the A.C.E.

Although the paper will not review the plans here, staff members at the Census Bureau are currently testing methods to reduce duplication in the 2010 Census by combining computer matching and field followup. These new methods will require testing in the experimental tests leading to 2010. At the same time, the tests will not provide effective data on all aspects of census duplication, such as duplicate enumerations between states. For some issues, then, data from Census 2000 will remain the best research source until the next census.

Previous work on duplication, including work used in the analysis of the A.C.E., has used essentially the names as captured on the census files. The names were edited under a few rules; for example, when parents filled in their full names but only their children's first names, the parents' last names were imputed to the children. Although results on duplication are available for these data, the results reported here are based on a revised file produced through the collaboration of David Word, Charles Coleman, and Robert Nunziata (2004). Their report on this work is not yet released, but they applied a series of edits to correct apparent spelling errors and occasional apparent reversals of first and last names. In aggregate, their edits increased the number of duplicates that could be identified with the methods to be described here.

## 2. Universe Descriptions and Key Variables

*2.1 The logic of two-stage matching.* Although different in several other respects, the approach used (1) in 2000 for HUDO, (2) in different cycles of work for the A.C.E., and (3) in the results to be reported here share the same logic of two-stage matching:

1. *First Stage:* search for person duplicates;
2. *Second Stage:* for 2+ households linked together by person duplicates from stage 1, attempt to identify additional duplicates.

For the work to be reported here, stage 1 is restricted to exact matches on name and birthdates, whereas the final A.C.E. work included some high-scoring statistical matches as well. In all applications, stage 1 was restricted to a universe with adequately reported data. In this study, stage 1 was restricted to enumerations with reported first and last name and full date of birth (month/day/year). Overall, approximately 91.3% of the 281.4 million person enumerations in Census 2000 were included in stage 1. Although enumerations in group quarters and persons in 1-person households in effect do not participate in stage 2, stage 2 nevertheless is open to all data-defined persons in the census, or approximately 97.9% of the count. (Data-defined persons are enumerations meeting threshold requirements for reporting characteristics.) When households or group quarters are enumerated to be a certain size, but the number of data-defined persons does not fully account for that size, then the remaining persons are imputed. These imputed persons (not data-defined) are outside of the scope of this study.

*2.2 Relationship between the A.C.E. and HUDO.* As previously noted, the HUDO initially identified almost 6 million enumerations for potential deletion, but in the end they permanently deleted about 3.6 million and reinstated 2.3 million. The A.C.E. naturally excluded the deleted enumerations, since they were not part of the final census, but it also treated the 2.3 million reinstated persons as out of scope. The A.C.E. universe was also restricted to the housing unit population, excluding group quarters (including both noninstitutional group quarters such as college dormitories and military barracks and institutional

group quarters such as nursing homes, hospitals, prisons, and jails). In spite of their exclusion of the HUDO deleted units from the final census, previous duplicate studies often include information about duplication between units in the final census and the deleted units. This study includes the reinstated enumerations but excludes the deletes. Table 1 summarizes the universe differences.

Table 1. Three universes encountered in 2000 duplication studies. Each universe expands on the previous one. The A.C.E. universe omits HUDO reinstates, persons in group quarters, and persons in remote Alaska. This study reports results for the final census.

| Universe | Relationship to previous universe |
|---|---|
| (1) A.C.E. Universe | |
| (2) Final Census 2000 | (2) – (1) = HUDO reinstates, GQ, remote AK |
| (3) Preliminary file before HUDO | (3) – (2) = HUDO deletes |

*2.3 Name Groups.* Word and Perkins (1996) analyzed last names based on data keyed from a sample from the 1990 census, primarily to identify predominantly Hispanic surnames. Their data set included only names of one or more Hispanic persons. Based on their data, the current analysis divides last names into four name groups, each given a mnemonic code letter.

*Group H: Common Hispanic surnames.* This group is based on the 639 most common predominantly Hispanic surnames in their data set. In their ranking, the top 10 are: Garcia, Martinez, Rodriguez, Lopez, Hernandez, Gonzalez, Perez, Sanchez, Rivera, and Ramirez. (The names occur 125 or more times in the data set analyzed by Word and Perkins.)

*Group C: Highly common non-Hispanic surnames.* This group is based on the 353 most common predominantly non-Hispanic surnames. The top 10 are: Smith, Johnson, Williams, Brown, Jones, Davis, Miller, Wilson, Anderson, and Moore. (Each name occurred 500 or more times.)

*Group F: Familiar non-Hispanic surnames.* This group is based on the next 7861 most common predominantly non-Hispanic surnames. Of these, 2,552, occurred 25 or more times in the file.

*Group N: Infrequent names, both Hispanic and non-Hispanic.* This group is defined as the residual from the other three groups. For enumerations eligible for phase 2, persons without a reported last name are also included in this group. (Persons without a reported last name are excluded from phase 1.)

*2.4 Ethnicity and Race.* For simplicity, this paper considers three mutually exclusive population groups: (1) Hispanics, reported to be of Hispanic ethnicity; (2) Non-Hispanic Blacks who reported their race as Black, possibly in combination with other races; and (3) Others, including Non-Hispanic Whites.

Table 2 shows universe sizes for the first and second phases. As expected, common Hispanic surnames are strongly correlated with Hispanic ethnicity. Perhaps surprisingly, almost 50% of non-Hispanic Blacks, but only 25% of other Non-Hispanics, are represented by the set of most common non-Hispanic surnames.

Table 2 First- and second-stage universe sizes and distributions for last names. (Note: Small differences in universe sizes between the original analysis and the files based on editing of Word et al. are not apparent when rounded to thousands.)

| | Hispanic | Non-Hispanic Black | All other |
|---|---|---|---|
| Eligible for exact (first-stage) matching, initial editing | | | |
| Universe (in 000's) | 31,031 | 30,139 | 195,827 |
| H: common Hispanic | 63.27% | 0.40% | 1.01% |
| C: highly-common non-Hispanic | 3.27% | 47.98% | 24.51% |
| F: familiar non-Hispanic | 3.26% | 18.63% | 19.58% |
| N: infrequent | 30.20% | 32.99% | 54.90% |
| Total | 100.00% | 100.00% | 100.00% |
| Eligible for exact matching, editing by Word et al. | | | |
| Universe (in 000's) | 31,031 | 30,139 | 195,827 |
| H: common Hispanic | 65.55% | 0.42% | 1.05% |
| C: highly-common non-Hispanic | 3.37% | 49.61% | 25.16% |
| F: familiar non-Hispanic | 3.22% | 19.12% | 20.04% |
| N: infrequent | 27.86% | 30.84% | 53.75% |
| Total | 100.00% | 100.00% | 100.00% |
| Eligible for second-stage matching/data defined | | | |
| Universe (in 000's) | 33,967 | 33,393 | 208,288 |

In general, the editing by Word and his colleagues reduced the size of Group N compared to the unedited data, but the effect appears larger for Hispanics and Non-Hispanic Blacks. In other words, editing differentially reduced the proportion of uncommon names for these two groups. Finally, it is of interest that Group C is larger than Group F, even though Group F includes over 20 times more surnames.

## 3. First-Stage Results with Exact Matching

As previously noted, the methods for the first stage were outlined in Fay (2002). Probabilities that exact matches were duplicates were separately estimated for duplication of persons in households, duplication in the group quarters population, and duplication between the household and group quarters population. Table 3 reports the results for

duplications between the household and group quarters population.

According to the estimated probabilities, 833 thousand exact matches correspond to approximately 670 thousand duplicates. To the level of rounding in Table 3, exact matches within county correspond almost entirely to duplicates. But the estimated probabilities suggest that more than 10% of the exact matches between counties in the same state, and over half the exact matches between states, are due to coincidental sharing of birth date rather than duplication.

As noted previously (Fay 2002), in applying the model no calculation was performed for very frequent names, effectively assigning them a probability of 0 rather than attempting to estimate a very small probability. Predictably, then, the model might underestimate in cells where many names occur with this high frequency. For group C, 128 thousand exact matches are reduced to an estimated 20 thousand duplicates by the estimated probabilities, including for cases where no probability is assigned, effectively treating the probability as zero. This cell is probably underestimated, and in fact only 10% of Group C's estimated duplicates occur between states, as opposed to 16% for Group F and N. If, in round numbers, Group C's estimate of 20 thousand between-state duplicates should be increased by perhaps 60% to bring it proportionately in line with Group F and N, about 12 thousand more duplicates would be added to the cell. It is similarly likely that between-state duplicates are underestimated for group H.

The A.C.E. sample was initially designed as a sample of housing units, and it provides an inadequate design to study characteristics of group quarters, including duplicates within the group quarters population. Consequently, an estimate of this form of duplication was previously unavailable. An approach similar to that shown in Table 3 yields an estimated 92 thousand duplicates within the group quarters population, with 58 thousand in the same block.

Table 4 presents results for duplications between persons in households. In Table 4, the importance of the estimated probabilities of duplication in the estimation of total duplicates is more substantial than in Table 3. For Group H, the estimated probabilities reduce the exact matches between county within state by more than half, and those between state by almost 90%. The effect of the estimated probabilities, including cases without assigned probability, is similarly substantial for Group C. Group N is the least affected, yet even its estimate of between-state duplicates is substantially reduced from the total number of exact matches. In spite of these large adjustments, the geographic distribution of duplicates, particularly among groups C, F, and N, is quite similar after weighting by the estimated probabilities. Without weighting by the probabilities, the geographic distributions of exact matches vary substantially; for example, over 2/3 of the exact matches for group C occur between state.

## 4. Second-Stage Results Including Statistical Matching

Fay (2003) outlined a method to estimate probabilities that a statistical match was a duplicate, conditional on the assigned matching score from the statistical match and on the number and estimated probabilities of the exact match or matches linking the households. The argument is complex and not easily summarized. Of note, however, is the role of an assumption that the presence of two or more exact matches ensured, with probability 1, that the matches were duplicates. The assumption was indirectly supported when the model estimated on the basis of the data that statistical matches with the highest level of matching were also estimated to be duplicates with probability 1 or virtually 1.

The second stage was actually accomplished in three steps:

1. Without changing the unconditional probabilities of duplication estimated for the exact matches, estimate the probabilities for statistical matches (Fay 2003).
2. Convert some unconditional probabilities of duplication to probabilities conditional on household information, by increasing the probabilities in households where household information provided supporting evidence in the form of other matched persons.
3. Reduce other unconditional probabilities to reflect the lack of support from household information in other cases.

As first example of step 2, two or more exact matches between households are treated as sufficient evidence to raise their probabilities to 1. As second example, the models for statistical matching can assign a statistical match a higher probability in the second stage than an exact match in the same households. This situation includes instances in which an exact match of a common name was not assigned a probability. Logically, the exact match has at least as much supporting evidence as the statistical match. Under these circumstances, the probability that an exact match is a duplicate is raised to the estimated probability for the statistical match.

Step 3 is necessary because increasing unconditional probabilities of duplication to conditional ones in step 2 requires reducing conditional probabilities for other cases to maintain the unbiasedness of the overall estimate. There are arguably different ways to do so, but the following approach was taken. First, exact matches that could neither be supported nor challenged by household information—namely exact matches involving 1-person households—were treated as associated with neutral information. The conditional probability was assigned to be the same as the unconditional probability. Second, the unconditional probabilities of exact matches with probabilities greater than the estimated probability from statistical matching were also left unchanged. Here, the statistical matching case lends some support to the exact match case, but not as much as in instances in step 2 where the estimated probability for the statistical match exceeds that for the exact match.

Exact matches involving 2+-person households at both

ends can be considered unsupported if no other members match exactly or statistically. The unconditional probabilities were adjusted downward for this group.

In summary, the following groups are of interest in steps 2 and 3:

1. Neutral household information: Exact match, only possible 1-person to 1+ persons, or exact match supported by one or more statistical matches with lower probability (unchanged);
2. Statistical match in combination with exact match: conditional probability already estimated (unchanged);
3. 2+ Exact matches: conditional probability to be increased to 1 (step 2);
4. Exact match supported by statistical match, probability to be increased to probability of statistical match (step 2);
5. Single exact match, unsupported where other matches possible (step 3).

For exact matches where the estimated unconditional probability is unassigned, which is effectively zero, no adjustments are required in step 3. For example, when two or more exact matches with unassigned probabilities occur between households, the household information becomes the basis to assign probabilities of 1 to cases unconditionally estimated effectively to be zero.

Table 5 shows the result of these operations for the estimation of between state duplicates. The first section of the table shows the large number of exact matches involving names appearing so frequently that no estimate was made for the probability. A small number of these, only a few percent, are assigned a positive probability based on household information.

The application of unconditional probabilities to the exact matches reduces the 176 thousand exact matches in 2+ exact households to 138 thousand. In the final section of the table, this estimate is raised back to 176 thousand, supplemented by an additional 20 thousand estimated for exact matches with unassigned probability originally. The cell for "exact supported by statistical" is similarly raised to 42 thousand from 34 thousand. To compensate, the cell corresponding to unsupported exact matches is lowered from 173 to 127 thousand by step 3.

Table 6 provides a summary across geography. Mule and Fenstermaker (2003) estimate approximately 5,211 thousand duplicates for the household universe, an estimate approximately 15% more than the 4,532 shown in Table 6.

Table 3 shows an estimate of 670 thousand duplicates between the housing unit and group quarters population. The corresponding estimate by Mule and Fenstermaker, 616 thousand (s.e. 46 thousand) is based on a previous exact matching.

As noted earlier, this study provides an estimate of 92 thousand duplicates within group quarters, an estimate that the previous study could not provide.

Figures 1 and 2 provide duplication rates between housing units by single years of age and geography. Duplication of children claimed by multiple households, of young adults in their 20s, and of older in their 50s and 60s through multiple residences appear at interpretable geographic levels.

## 6. Discussion

The use of exact and statistical matching achieves results similar to, although a bit below, the A.C.E. estimates. Although differences could be investigated further, it is possible that the A.C.E. duplication study gains the advantage from combining statistical matching with exact matching in the first stage of the A.C.E. duplicate study. But the advantage of the current data set is to permit more detailed analysis. Of immediate interest is to further investigate differences by ethnicity and race in the patterns of duplication.

Note: (1) This paper reports the results of research and analysis undertaken by a member of the U.S. Census Bureau staff. It has undergone a 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.

## References


Fay, R.E. (2001), "The 2000 Housing Unit Duplication Operations and Their Effect on the Accuracy of the Population Count," *2001 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA.

_____(2002), "Probabilistic Models for Detecting Census Person Duplication," *2002 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA, pp. 969-974.

_____ (2003), "Probabilistic Models for Detecting Census Duplication at the Person and Household Levels," *2003 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA, pp. 1391-1398.

Mule, T. (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.

Mule, V.T., Jr. (2002a), "Person Duplication in Census 2000," *2002 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA, pp. 3471-3476.

Mule, T. (2002b), "A.C.E. Revision II Results: Further Study of Person Duplication," DSSD A.C.E. Revision II Memorandum Series #PP-51, U.S. Census Bureau, Dec. 31, 2002, at http://www.census.gov/dmd/www/pdf/pp-51r.pdf.

Mule, V.T., Jr. and Fenstermaker, D. (2003), "Overview and Results of Further Study of Person Duplication for the A.C.E. Revision II," *2003 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA, pp. 2948-2953.

Nash, F.F. (2000), "Overview of the Duplicate Housing Unit Operations," unpublished report, U.S. Census Bureau, Census 2000 Informational Memorandum No. 78, Nov. 7, 2000.

Word, D.L., Coleman, C.D., and Nunziata, R. (2004), "Putting a Demographic Face on Names from Census 2000," unpublished draft report.

Word, D.L. and Perkins, R.C., 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 3. Exact matches and estimated duplicates between the household and group quarters populations. Duplicates are estimated by summing the estimated probabilities of duplication.

| | H: Common Hispanic | C: Highly Common Non-Hisp | F: Familiar Non-Hispanic | N: Infrequent | Total |
|---|---|---|---|---|---|
| Exact Matches (in 000's) | | | | | |
| Same block | 4 | 27 | 20 | 48 | 100 |
| Same tract, diff block | 2 | 12 | 8 | 18 | 41 |
| Same county, diff tract | 19 | 65 | 39 | 88 | 211 |
| Same state, diff co | 16 | 79 | 48 | 109 | 253 |
| Diff state | 25 | 128 | 24 | 51 | 228 |
| Total | 67 | 312 | 140 | 315 | 833 |
| Estimated Duplicates (in 000's) | | | | | |
| Same block | 4 | 27 | 20 | 48 | 100 |
| Same tract, diff block | 2 | 12 | 8 | 18 | 41 |
| Same county, diff tract | 19 | 65 | 39 | 88 | 211 |
| Same state, diff co | 9 | 68 | 43 | 104 | 224 |
| Diff state | 6 | 20 | 20 | 49 | 94 |
| Total | 41 | 192 | 130 | 307 | 670 |
| Percentage Distribution of Estimated Duplicates | | | | | |
| Same block | 11% | 14% | 15% | 16% | 15% |
| Same tract, diff block | 6% | 6% | 6% | 6% | 6% |
| Same county, diff tract | 46% | 34% | 30% | 29% | 31% |
| Same state, diff co | 23% | 35% | 33% | 34% | 33% |
| Diff state | 14% | 10% | 16% | 16% | 14% |
| Total | 100% | 100% | 100% | 100% | 100% |

Table 4. Exact matches and corresponding estimated duplicates within the household population, for the first phase of person matching.

| | H: Common Hispanic | C: Highly Common Non-Hisp | F: Familiar Non-Hispanic | N: Infrequent | Total |
|---|---|---|---|---|---|
| Exact Matches (in 000's) | | | | | |
| Same block | 144 | 318 | 217 | 553 | 1233 |
| Same tract, diff block | 27 | 96 | 60 | 140 | 323 |
| Same county, diff tract | 112 | 264 | 161 | 387 | 924 |
| Same state, diff co | 110 | 221 | 111 | 261 | 702 |
| Diff state | 418 | 1930 | 186 | 323 | 2857 |
| Total | 811 | 2829 | 734 | 1665 | 6039 |
| Estimated Duplicates (in 000's) | | | | | |
| Same block | 140 | 317 | 217 | 553 | 1226 |
| Same tract, diff block | 26 | 96 | 60 | 140 | 322 |
| Same county, diff tract | 100 | 262 | 160 | 386 | 909 |
| Same state, diff co | 42 | 155 | 106 | 254 | 556 |
| Diff state | 44 | 77 | 85 | 220 | 427 |
| Total | 352 | 906 | 628 | 1553 | 3439 |
| Percentage Distribution of Estimated Duplicates | | | | | |
| Same block | 40% | 35% | 35% | 36% | 36% |
| Same tract, diff block | 7% | 11% | 9% | 9% | 9% |
| Same county, diff tract | 28% | 29% | 26% | 25% | 26% |
| Same state, diff co | 12% | 17% | 17% | 16% | 16% |
| Diff state | 13% | 9% | 14% | 14% | 12% |
| Total | 100% | 100% | 100% | 100% | 100% |

Table 5. Estimation for the second phase of matching for between state duplicates for the housing unit population. The second set of rows show the status of matches, exact and statistical, before adjustment of the exact probabilities, described as steps 2 and 3 in the paper. The third set of rows show the effect of steps 2 and 3. The columns identify the categories used in the adjustment of the exact probabilities to reflect household information. (Dash indicates logically empty.)

| | Exact, p unassigned | Exact, p estimated, hh information neutral | Stat. match | Exact in 2+ exact hh | Exact, supported by stat match | Unsupported exact | Total |
|---|---|---|---|---|---|---|---|
| Exact and Statistical Matches (000's) | | | | | | | |
| Exact, p unassigned | 2067 | - | - | 20 | 29 | - | 2116 |
| Exact, p est. | - | 135 | - | 176 | 43 | 387 | 741 |
| Statistical match | - | - | 107 | - | - | - | 107 |
| Total | 2067 | 135 | 107 | 196 | 72 | 387 | 2964 |
| Estimated Duplicates (000's) Before Adjustment of Unconditional Probabilities | | | | | | | |
| Exact, p est. | - | 82 | - | 138 | 34 | 173 | 427 |
| Statistical match | - | - | 68 | | | | 68 |
| Total | - | 82 | 68 | 138 | 34 | 173 | 495 |
| Estimated Duplicates (000's) After Adjustment of Unconditional Probabilities | | | | | | | |
| Exact, p unassigned | - | - | - | 20 | 5 | - | 25 |
| Exact, p est. | - | 82 | - | 176 | 42 | 127 | 427 |
| Statistical match | - | - | 68 | - | - | - | 68 |
| Total | - | 82 | 68 | 196 | 46 | 127 | 519 |

Table 6. Estimation for the second phase of matching, total. The rows in this table correspond to those in the preceding table, but they combine results across all levels of geography.

| | Exact, p unassigned | Exact, p estimated, hh information neutral | Stat. match | Exact in 2+ exact hh | Exact, supported by stat match | Unsupported exact | Total |
|---|---|---|---|---|---|---|---|
| Exact and Statistical Matches (000's) | | | | | | | |
| Exact, p unassigned | 2139 | - | - | 24 | 31 | - | 2194 |
| Exact, p est. | - | 669 | - | 1782 | 290 | 1103 | 3845 |
| Statistical match | - | - | 1127 | - | - | - | 1127 |
| Total | 2139 | 669 | 1127 | 1806 | 321 | 1103 | 7166 |
| Estimated Duplicates (000's) Before Adjustment of Unconditional Probabilities | | | | | | | |
| Exact, p est. | - | 601 | - | 1725 | 276 | 837 | 3439 |
| Statistical match | - | - | 1060 | | | | 1060 |
| Total | - | 601 | 1060 | 1725 | 276 | 837 | 4499 |
| Estimated Duplicates (000's) After Adjustment of Unconditional Probabilities | | | | | | | |
| Exact, p unassigned | - | - | - | 24 | 6 | - | 29 |
| Exact, p est. | - | 601 | - | 1782 | 288 | 771 | 3443 |
| Statistical match | - | - | 1060 | - | - | - | 1060 |
| Total | - | 601 | 1060 | 1806 | 294 | 771 | 4532 |

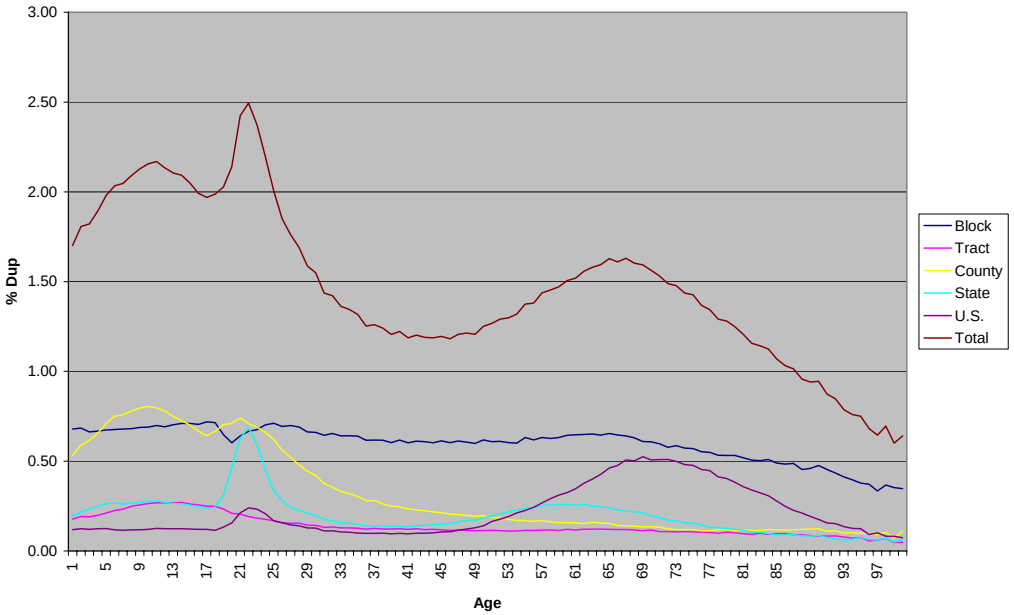


Figure 1. Person duplication rates in Census 2000 between housing units as a percent of the total population by age. Geographic categories are mutually exclusive. *Block* means the same block, *tract* the same tract but not the same block, *etc.* The combined total is shown.

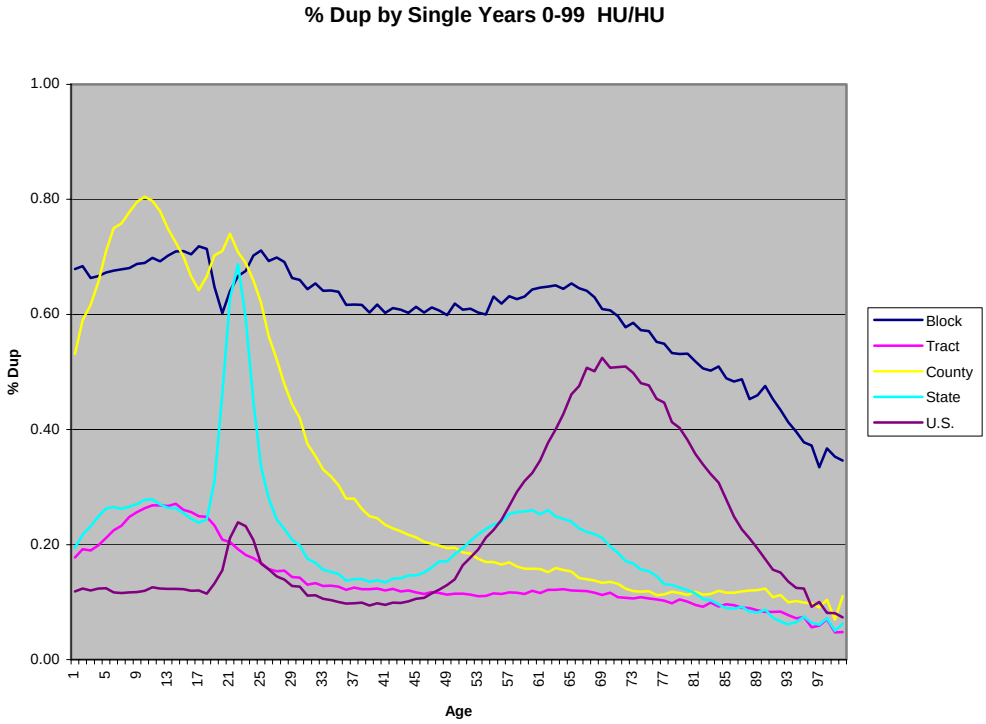


Figure 2. Percent duplication rates in Census 2000 by single years of age by geography, without an overall total. The categories are identical to those in Table 1, but the omission of a line for total allows a scale providing more detail for separate geographic categories.