PERSON RECORD DUPLICATION IN THE 2000 DECENNIAL CENSUS

John Jones and Roxanne Feldpausch
U. S. Census Bureau, Washington DC 20233[1]

Keywords: Accuracy and Coverage Evaluation; Duplicate census records

1. Introduction

This paper examines the Census 2000 duplication as measured by the 2000 Accuracy and Coverage Evaluation (A.C.E.). The A.C.E. was an operation undertaken to evaluate the coverage of Census 2000. It was comprised of the matching of an independent enumeration in a stratified sample of census block clusters against the Census 2000 enumerations in those block clusters. The 2000 A.C.E. included an initial housing unit phase, a person interview phase, a person match phase, and a final housing unit phase. For more information on the A.C.E. see Childers (2001).

The person matching phase of the A.C.E. began after census day and after the person interview phase was completed. Information on persons independently enumerated came from the person interview; these persons are also known as P-sample persons. In person matching, match and residence codes were assigned to P-sample person records and match and enumeration codes were assigned to census person records. Also, a duplicate search was performed for census person records. This search occurred within housing units in the search area and not within group quarters. We are concerned with census person records that referred to the same person as other census person records.

2. Background and Methodology

The persons enumerated in the sampled clusters by Census 2000 were divided into three groups based upon the outcome of a subsampling within large clusters and upon the selection of a subset of clusters for targeted extended search. Census persons were accordingly placed into one of the following three groups:

- **E-sample persons:** These are persons enumerated in small and medium sized block clusters and persons enumerated in large block clusters that are still in sample after within large block subsampling.
- **Non-E-sample persons:** These are persons that are out of sample after within large block subsampling.
- **Surrounding block persons:** These are persons enumerated in the surrounding blocks of clusters chosen for targeted extended search.

A census person record is said to be the duplicate of another census person record if the pair of records refer to the same person. The characteristics used to identify duplicates were name, age, gender, race, Hispanic origin, and street address. The duplicate search was restricted to duplicates of E-sample persons. When two or more records referred to the same person, all but one of them were coded as duplicates. Each E-sample person that is coded duplicate counts as one erroneous enumeration. When E-sample persons were duplicated by one or more non-E-sample persons, the person who is duplicated counts as less than one erroneous enumeration with the exact fraction depending on the number of non E-sample persons that are duplicates.

Duplicate records are linked to those records that they duplicate. These duplicated records are called primaries. Tables 5, 8, and 10 were created using a database of linked duplicate pairs. It is possible for census persons to have more than one duplicate; when this happens a separate record was created for each duplicate pair. There are three types of duplicate pair linkages. They are in descending frequency of occurrence:

- **E-sample duplicates of E-sample persons:** This is the most common type of duplicate pair. The duplicated person was either matched to a

[1]John Jones and Roxanne Feldpausch are mathematical statisticians in the Decennial Statistical Studies Division of the U. S. Census Bureau. 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.

P-sample person; not matched to the P-sample, but correctly enumerated in the Census; or not matched to the P-sample and has unresolved enumeration status. Persons who were erroneously enumerated were not duplicated.

- **Non-E-sample duplicates of E-sample persons:** This is the next most common type of duplicate pair. The duplicated E-sample person counts as a partial erroneous enumeration.

- **E-sample duplicates of surrounding block persons:** This occurs in clusters chosen for targeted extended search because of errors identified in the initial housing unit phase.

The percentage duplication featured in Tables 1–4, 6, 7, 9 and 11–13 are determined by taking the ratio of the number of E-sample person records who are duplicates in a category to the total number of E-sample person records in that category. We compute these rates by counting as duplicates E-sample persons coded as duplicates as well as E-sample persons who are linked to non-E-sample duplicates. Weighted rates reflect the probability of selection in all phases of sampling and the probability of erroneous enumeration. Standard errors of these rates were calculated using stratified Jackknife methods by the software package VPLX. VPLX uses replication methods to calculate variances of estimates derived from complex surveys as described in Fay (1990). Once these rates and their standard errors are determined, within variable comparisons are made to check for significant differences in the frequency of duplication. These comparisons are made using critical values of *t*-statistics, with overall significance level .10. These critical values are determined using a multiple comparison of means technique with a Bonferroni adjustment, per Hocking (1986, pp 108-109). The Bonferroni adjustment applies a significance level to each individual comparison that is consistent with the overall significance level.

3. The Frequency of E-sample Duplication

Table 1 gives the aggregate weighted rate of duplication in the E-sample for the 1990 Post Enumeration Survey and the 2000 A.C.E. Here, rates are of the total (weighted) number of persons in the E-sample and are expressed in percentages. It shows that the relative amount of person duplication has fallen since 1990. This may be attributable to the Housing Unit Duplication Operation that deleted duplicate housing units and the people enumerated in them (see Nash 2001).

Table 1: Overall Percent Duplication

Year	Weighted Percentage Duplication
1990	1.60
2000	0.76

Tables 2–4 give weighted rates of duplication in the 2000 A.C.E. by regional office, size of metropolitan area, and type of census return, respectively. Tables that give person duplicate frequencies display variable level names, the percentage duplication ("Percent"), the stratified Jackknife standard error ("s.e."), the rank of the percentage duplication frequency in descending order ("Rank"), and the ranks of levels with which a significant difference was found ("Differ"). Each pair of levels of each variable was compared by a *t*-test with a critical value of t given below each table. The value of *t* is calculated as the difference in percentages divided by the sum of the standard errors of those percentages.

Table 2 gives weighted duplication rates by A.C.E. regional office. It shows that the New York and Boston offices have the highest rates of duplication while the Detroit and Los Angeles offices have the lowest. In fact, the New York office has significantly higher duplication rates than all other regional offices. The Boston office has significantly higher duplicate rates than five of the remaining ten regional offices (excluding New York). It appears that census duplication is more frequent in the Northeast.

Table 3 gives weighted duplication rates by size of the metropolitan statistical area (MSA). The possible metropolitan area sizes are large, medium, small, and non-MSA. The level non-MSA is a close approximation to rural, sparsely populated locations. Table 3 shows that large metropolitan areas and rural areas have higher duplication rates than small and medium metropolitan areas. The differences are significant. The results of Table 3 reinforce the results of Table 2 because the New York and Boston regional offices have cities that are in large MSA's.

Table 2: Regional A.C.E. Office Weighted Duplication Rates by Rank with Percentage

Regional Office	Percent (s.e.)	Rank	Differ
New York	2.04 (0.14)	1	all
Boston	1.07 (0.16)	2	1,10,11,12
Dallas	0.82 (0.08)	3	1,12
Seattle	0.76 (0.10)	4	1
Atlanta	0.72 (0.07)	5	1
Chicago	0.71 (0.07)	6	1
Charlotte	0.69 (0.06)	7	1
Philadelphia	0.61 (0.06)	8	1
Kansas City	0.59 (0.08)	9	1
Denver	0.51 (0.06)	10	1,2
Los Angeles	0.48 (0.05)	11	1,2
Detroit	0.44 (0.05)	12	1,2,3

Critical value of *t*: 3.164

Table 3: Metropolitan Area Size Weighted Duplication Rates by Rank with Percentage

MSA Size	Percent (s.e.)	Rank	Differ
Non-MSA	0.95 (0.06)	1	3,4
Large MSA	0.92 (0.05)	2	3,4
Small MSA	0.62 (0.06)	3	1,2
Medium MSA	0.56 (0.04)	4	1,2

Critical value of *t*: 2.386

Table 4 gives weighted duplicate rates by type of census return. Generally, a respondent could either fill out his census forms himself or have a census enumerator fill it out. Persons living in mailout/mailback areas could have enumerator-filled returns if they did not respond by mail. The several types of enumerator-filled returns have been lumped together. Table 4 shows that enumerator-filled returns have higher duplicate rates than mail returns. The difference is significant.

Table 4: Type of Census Return Weighted Duplication Rates by Rank with Percentage

Type of Return	Percent (s.e.)	Rank	Differ
Mail	1.89 (0.08)	1	2
Enumerator	0.41 (0.02)	2	1

Critical value of *t*: 1.65

Table 5 cross-classifies the return type of each duplicate pair. The rows of this cross-classification table give the return type of the primary person. The primary person is defined to be the one who was duplicated. The third row heading, "Surr," refers to pairs whose primary person was enumerated in a surrounding block. Information on the return type of these primary people is unavailable. The columns of this table give the return type of the duplicate person. The third column heading, "Subsam," refers to pairs whose duplicate person was enumerated in a housing unit that was subsampled out of the E-sample. Information on the return type of these person duplicates is unavailable. The second row and column heading, "Enum," refers to enumerator filled returns. Table 5 shows that over half (56.1%) of the duplicate pairs are mail and enumerator.

Table 5: Cross-Classification by Return Type

Return of Primary	Return of Duplicate			
	Mail	**Enum**	**Subsamp**	**Total**
Mail	640	2,210	404	3,254
Enum	1,648	1,250	489	3,387
Surr	75	167	0	242
Total	2,363	3,627	893	6,883

Table 6 gives weighted duplication rates by race post-strata. Person duplication occurs most frequently among African Americans and Hispanics and less frequently among whites.

Table 6: Race/Hispanic Origin Weighted Duplication Rates by Rank with Percentage

Domain	Percent (s.e.)	Rank	Differ
African American	1.19 (0.08)	1	6,7
Hispanic	1.12 (0.07)	2	7
Asian	1.02 (0.17)	3	none
Pacific Islander	0.76 (0.21)	4	none
American Indian on reservation	0.74 (0.15)	5	none
American Indian off reservation	0.65 (0.15)	6	1
White	0.61 (0.03)	7	1,2

Critical value of *t*: 2.815

Table 7 gives weighted percentages by age/sex grouping while Table 8 cross classifies duplicate pairs by age grouping. Table 7 shows that males aged 18-29 are most likely to be person duplicates while those aged 0-17 are least likely. Table 8 shows that 77.4% of all pairs agree on age grouping. Missing values occur mainly in the duplicate record.

Table 7: Age and Sex Group Weighted Duplication Rates by Rank with Percentage

Age/Sex Category	Percent (s.e.)	Rank	Differ
18-29 Male	1.01 (0.06)	1	4,5,6,7
18-29 Female	0.88 (0.06)	2	6,7
50+ Female	0.84 (0.06)	3	7
30-49 Male	0.82 (0.04)	4	1,6,7
50+ Male	0.79 (0.04)	5	1,7
30-49 Female	0.70 (0.03)	6	1,2,4,7
0-17	0.58 (0.03)	7	all

Critical value of *t*: 2.807

Table 8: Cross-Classification by Age category

Age of Primary	Age of Duplicate				
	0-17	**18-29**	**30-49**	**50+**	**Missing**
0-17	1,177	36	10	11	116
18-29	34	1,090	43	7	131
30-49	20	49	1,519	62	233
50+	10	11	67	1,543	306
Miss	63	66	122	93	64

Tables 9 and 10 give weighted duplication rates by gender and a cross-classification of duplicate pairs by gender, respectively. Table 9 shows significantly higher duplication frequencies for males. Table 10 is similar to Table 8 in that most pairs agree on gender and that most of the missing values are on the duplicate record.

Table 9: Weighted Duplication Percentages by Gender

Gender	Percent (s.e.)	Rank	Differ
Male	0.78 (0.03)	1	2
Female	0.74 (0.03)	2	1

Critical value of *t*: 1.65

Table 10: Cross-Classification by Gender

Gender of Primary	Gender of Duplicate			
	Male	**Female**	**Missing**	**Total**
Male	3,227	107	74	3,408
Female	115	3,197	65	3,377
Missing	45	40	13	98
Total	3,387	3,344	152	6,883

4. Housing Unit Characteristics of Person Duplicates

Next, we consider the housing unit characteristics of person duplicates. Table 11 gives weighted duplication rates by the housing unit enumeration status as

determined in the final housing unit match. Housing unit enumeration status can be divided into correct enumerations (which include matched units), housing units that are duplicates of other housing units, other erroneous enumerations (which include geocoding errors), and units with unresolved enumeration status. It shows that duplication frequencies are significantly higher in duplicate housing units and significantly lower in correctly enumerated units.

Table 11: Housing Unit Enumeration Status Weighted Duplication Rates by Rank with Percentage

Housing Enumeration Status	Percent (s.e.)	Rank	Differ
Duplicate Housing Unit	32.88 (3.44)	1	all
Unresolved	7.54 (2.21)	2	1,4
Other Erroneous	3.34 (0.28)	3	1,4
Correctly Enumerated	0.54 (0.02)	4	all

Critical value of *t*: 2.386

Table 12 gives weighted duplication rates by type of basic street address. It shows that small multi-unit structures with 2 to 9 units at the basic street address have higher duplication rates than single family homes and larger apartment buildings. All pairwise differences are significant.

Table 12: Number of Units at Address Weighted Duplication Rates by Rank with Percentage

Number of units	Percent (s.e.)	Rank	Differ
2-9	3.33 (0.15)	1	all
10+	1.04 (0.11)	2	all
1	0.39 (0.02)	3	all

Critical value of *t*: 2.121

Table 13 gives weighted person duplication percentages by source of census address. Addresses are added to the census by a variety of operations. Some of these operations occurred before the census while others occurred after census day. The sources of housing address of persons in the E-sample are as follows:

- **1990 Census Address Control File (1990 ACF):** Addresses that were on file at the Census Bureau in 1990.
- **Pre-Census Address Listing Operations (Address List):** This was a field operation occurring in non-mailout/mailback enumeration areas.
- **Postal Delivery Sequence Files (DSF):** This was a monthly update of addresses from the Postal Service.
- **Pre-Census Block Canvassing Operations (Block Canvass):** This was a field verification of addresses on the Master Address File as of January 1999.
- **Local Update of Census Addresses (LUCA):** An update attributable to a cooperative effort with local governments.
- **Questionnaire Delivery (QD):** Field operations where addresses were updated while census forms are hand delivered to housing units.
- **Non-Response Follow-up (NRFU):** These were address updates from enumerators visiting households that have not completed mail returns.
- **Coverage Improvement Follow-up, Telephone Questionnaire Assistance, Be Counted (CIFU, TQA, Be Counted):** These were additional census operations that began after NRFU that furnish addresses.
- **New Construction (NC):** These were housing units under construction around Census Day and recently completed.
- **Special Place or Group Quarters (SPGQ):** Addresses were furnished by the census enumeration of special places and group quarters.

Table 13 shows that person duplication is most frequent in housing units added to the census by Non-Response Follow-up, Coverage Improvement Follow-up, Special

Place enumeration and Group Quarters enumeration. All of these operations occurred after census day.

Table 13: Source of Census Address Weighted Duplication Rates by Rank with Percentage

Source of Address	Percent (s.e.)	Rank	Differ
SPGQ	9.87 (6.30)	1	none
CIFU, TQA, Be Counted	9.07 (1.70)	2	6,7,8,9,10
NRFU	8.74 (1.62)	3	6,7,8,9,10
NC	5.17 (3.10)	4	none
QD	3.73 (0.47)	5	8,9,10
Block Canvass	2.99 (0.38)	6	2,3,8,9,10
LUCA	2.32 (0.36)	7	2,3,8,9,10
Address List	0.72 (0.05)	8	2,3,5,6,7
DSF	0.61 (0.05)	9	2,3,5,6,7
1990 ACF	0.58 (0.03)	10	2,3,5,6,7

Critical value of *t:* 3.051

5. Summary and Conclusions

There are important regional differences in the rate of person duplication. Duplication is more prevalent in the Northeast. It is also more prevalent in large urban areas and in rural areas. Person duplicates are more frequently male and between 18 and 29 years old. Racial and ethnic groups that are traditionally undercounted such as African Americans and Hispanics are also more likely to be person duplicates.

Person duplicates occur most frequently in housing unit duplicates. They are most likely to be found in small multi unit housing structures. They are also the most prevalent in addresses furnished to the census by Census 2000 operations.

In most cases, person duplicates have a different return type than those that they duplicate. Otherwise, person duplicates generally share person characteristics with those they duplicate but have more missing characteristics than those they duplicate.

6. References

Childers, D. (2001). "The Design of the Census 2000 Accuracy and Coverage Evaluation (A.C.E.)" internal memorandum, U.S. Bureau of the Census

Fay, R. (1990). "VPLX: Variance Estimates for Complex Samples," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Hocking, RR (1986). *Methods and Applications of Linear Models: Regression and the Analysis of Variance* (New York: John Wiley and sons).

Nash, Fay (2001). "Overview of the Duplicate Housing Unit Operations" internal memorandum, U.S. Bureau of the Census