HOUSING UNIT DUPLICATION IN THE 2000 DECENNIAL CENSUS

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

Keywords: Accuracy and Coverage Evaluation; Dual System Estimation

1. Introduction

This paper examines housing unit duplication in Census 2000 as measured by the 2000 Accuracy and Coverage Evaluation (A.C.E.). The Accuracy and Coverage Evaluation was an operation undertaken to evaluate the population and housing coverage of Census 2000. First, it performed an independent enumeration of housing units and people within a stratified sample of census block clusters. Then it matched this enumeration against the Census 2000 enumeration of housing units and people in those same block clusters. The Accuracy and Coverage Evaluation included an initial housing unit phase, where housing units in the sampled block clusters were matched against units listed in the January 2000 Decennial Master Address File (DMAF) in those same clusters; a person interview phase, where demographic information was collected from census day residents of housing units in the sampled block clusters; a person match phase, where persons listed in the independent enumeration were matched against the census record of persons in those same clusters; and a final housing unit phase, where updates to housing unit inventories after the end of the initial housing unit phase were processed. Estimates of housing unit and person coverage were produced after the completion of the A.C.E.

The 2000 Accuracy and Coverage Evaluation included the match of an independent enumeration of housing units in a sample of block clusters against the Census 2000 enumeration of housing units in those same clusters. The independent enumeration is known as the P-sample. After the initial housing unit phase, census housing units in A.C.E. clusters were subsampled. Units remaining in sample after this subsampling belong to the E-sample. The A.C.E. only recorded the enumeration status of E-sample units. Therefore, only E-sample units are of interest.

Section 2 gives a more detailed background and discusses the methods used to analyze the data. Section 3 documents the overall frequency of census duplication and compares this frequency within levels of important variables. Section 4 discusses the agreement of the address characteristics between linked duplicate housing unit pairs. Section 5 gives a summary and conclusions.

2. Background and Methodology

The Accuracy and Coverage Evaluation classified census housing units as either correct enumerations or erroneous enumerations. Correctly enumerated housing units have addresses that were confirmed to exist as housing units within the block cluster while erroneously enumerated housing units have addresses that are not confirmed to exist as housing within the block cluster. Duplicate housing units were counted as erroneous enumerations, and the initial and final housing unit phases of the A.C.E. conducted a search for housing unit duplicates and identified them as such. The objective of this study is to document the extent of census housing duplication, to give the characteristics of housing units most likely to be duplicates, and to identify the nature of duplicate housing unit addresses.

The housing unit phase began with an independent listing of the addresses of all of the housing units in the sample clusters. The sample clusters were stratified into small, medium, large block clusters, and clusters located on American Indian Reservations. After the independent listing, there was a reduction in the number of small block and medium block clusters. After this reduction in the number of sample clusters, housing units on the independent list of sample addresses were matched against the housing units listed on the January 2000 version of the Decennial Master Address File (DMAF) in the sampled clusters.

The housing unit matching began with a computer match of addresses that compared the independent listings with the DMAF (census) listings and identified matched addresses and possibly matched addresses. Addresses were said to match when an address from the independent list and an address from the census referred to the same housing unit. All addresses not matched by computer then came under before followup (BFU) clerical review where additional matches were made and duplicate searches within census address listings were performed. Addresses

[1]John Jones is a mathematical statistician 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 more limited review than official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

that were nonmatched, possibly matched, or determined to be possible duplicates of other addresses after BFU clerical review were sent to housing unit followup (HUFU). It was possible that two addresses not identified as duplicates by the clerical review could be identified as duplicates by the housing unit followup. Information obtained from housing unit followup was used to assign after followup (AFU) match and enumeration codes to housing units.

Duplicates occurred when two or more addresses referred to the same housing unit. When this happened, one of the addresses was regarded to be the primary (true) address, and the others were considered to be duplicate addresses. The primary address was often matched to an address on the independent list, or otherwise confirmed to be correctly enumerated in the block cluster. Duplicate linkages between the primary address and the duplicate address were generated. There were duplicate linkages between E-sample and non E-sample addresses, as well as linkages between E-sample addresses. Census addresses that were duplicates were either coded as such during clerical review or after being sent to housing unit followup for verification. Duplicate search also occurred in the final housing unit (FHU) phase. Some of the updates to census housing unit inventory created additional duplication and some of the units in relisted and list enumerate clusters were duplicates. Possible duplicate addresses were sent for confirmation during final housing unit followup (FHUFU). All data in the following tables are based upon housing unit files created after the FHU phase.

Table 1 gives the overall weighted percentage of E-sample housing units classified as duplicates while **tables 2 through 7** give this percentage and the associated standard error for each level of the following variables:

- Region
- Sampling Stratum
- Metropolitan Statistical Area/Type of Enumeration Area (MSA/TEA) group
- Tenure and Occupancy Status
- Type of structure (Number of units at Basic Street Address)
- Race/Hispanic origin of householder (Occupied units only)

The percentage of duplication is the ratio of the weighted number of duplicates to the weighted number of housing units. Both units with final match code as duplicate and units with duplicate links to non E-sample housing units are counted as duplicates. Units with final match code as duplicate are counted as one erroneous enumeration while units with duplicate links to non E-sample housing units are counted as a partial erroneous enumeration, with the exact fraction depending upon the number of non E-sample duplicate links. Standard errors of the duplicate percentage were calculated using the stratified jackknife. For a given variable, each pair of levels was compared by a t-test with a critical value of t given below each table. The critical values are based on a multiple comparison of means technique with a Bonferroni adjustment. The overall significance level is 10 percent.

Tables 8 through 10 utilize a database of linked duplicate pairs. If an E-sample unit had *n* duplicates then the database had *n* separate records. Each record of the database contains address and housing unit characteristics of each member of the linked duplicate pair. The database was used to investigate the agreement on these characteristics of the linked pairs.

3. The Frequency of E-sample (Census) Duplication

Table 1 gives the aggregate weighted rate of E-sample housing unit duplication measured by the 2000 A.C.E. Here, rates are of the total weighted number of housing units in the E-sample and of the total weighted number of erroneously enumerated housing units in the E-sample. It also compares the percentage of erroneous enumerations that are duplicates in 2000 to that computed by the 1990 Post-Enumeration Survey.

Table 1: Percentage E-sample Housing Unit Duplication

Year	Percent of Erroneous Enumerations that were duplicates	Percent of units that were duplicate	Percent of E sample units that were duplicates
1990	33.4	2.8	0.93
2000	24.8	2.3	0.57

Tables 2-7 give weighted percentages of census housing unit duplication in the 2000 A.C.E. by important variables. They display variable level names, variable level numbers, the weighted percentage of E-sample housing units in level that are duplicates (percent duplicates), the stratified jackknife standard error (s.e.), and a list of level numbers with which a significant difference was found (differ). For a given variable, each pair of levels of each variable was compared by a t-

test with a critical value that reflects the Bonferroni criterion. These critical values of t are given below each table.

Table 2 gives weighted housing unit duplication rates by region. It shows that there were no significant regional differences in housing unit duplication.

Table 2: E-sample Housing Unit Duplication Percentages by Region

Region	Percent Duplicates (s.e.)	Rank	Differ
Northeast	0.68 (0.12)	2	none
Midwest	0.39 (0.06)	4	none
South	0.71 (0.19)	1	none
West	0.43 (0.08)	3	none

Critical value of t: 2.386

During the initial housing unit phase, sample clusters were stratified into small block clusters (less than three housing units per block), medium block clusters (from 3-79 housing units per block), large block clusters (80 or more housing units per block), and clusters located on American Indian Reservations (A.I.R.). **Table 3** gives weighted housing unit duplication rates by sampling stratum. It shows that clusters on American Indian Reservations have significantly higher housing unit duplication than medium sized clusters. There were no other significant differences.

Table 3: E-sample Housing Unit Duplication Percentages by Sampling Stratum

Stratum	Percent Duplicates (s.e.)	Rank	Differ
Small	0.64 (0.29)	3	none
Medium	0.48 (0.03)	4	1
Large	0.72 (0.20)	2	none
A.I.R.	1.50 (0.38)	1	4

Critical value of t: 2.386

Table 4 gives housing unit duplication rates by Metropolitan Statistical Area / Type of Enumeration Area (MSA/TEA) Group. Mailout/mailback (MOMB) areas have city style addresses and most of these addresses lie in large or medium metropolitan statistical areas (MSA). Other types of enumeration areas consist of isolated areas with small populations and non city style addresses. There was significantly more housing unit duplication in the more rural isolated areas.

Table 4: E-sample Housing Unit Duplication Percentages by Metropolitan Statistical Area/Type of Enumeration Area (MSA/TEA)

MSA/TEA	Percent Duplicates (s.e.)	Rank	Differ
Large MSA MOMB	0.31 (0.04)	4	1
Medium MSA MOMB	0.35 (0.07)	3	1
Small MSA & Non MSA MOMB	0.83 (0.33)	2	none
All other TEA	1.01 (0.08)	1	3,4

Critical value of t: 2.386

Table 5 gives housing unit duplication frequency by the

type of housing unit structure. Small multiunits (2-9 units at basic address) were more frequently duplicates than single units and the difference is significant.

Table 5: E-sample Housing Unit Duplication Percentages by Type of Structure

Number of units at address	Percent Duplicates (s.e.)	Rank	Differ
1	0.36(0.03)	3	1
2-9	1.40(0.44)	1	3
10+	0.96(0.77)	2	none

Critical value of t: 2.121

Table 6 gives housing unit duplication frequency by the housing unit tenure of occupied units. It shows that vacant units are most frequently duplicates. However, owner occupied housing units have significantly less housing unit duplication than units not occupied by the owner.

Table 6: E-sample Housing Unit Duplication Percentages by Housing Tenure

Tenure	Percent Duplicates (s.e.)	Rank	Differ
Owner	0.34 (0.08)	3	1,2
Non Owner	0.62 (0.03)	2	3
Vacant	2.01 (0.70)	1	3

Critical value of t: 2.121

Table 7 gives housing unit duplication by race domain of householder, for occupied units only. It shows that there were no significant racial differences in the frequency of housing unit duplication.

Table 7: E-sample Housing Unit Duplication Percentages by Racial/Ethnic Domain (Occupied Units Only)

Domain	Percent Duplicates (s.e.)	Rank	Differ
American Indian on reservation	1.31 (0.38)	1	none
American Indian off reservation	0.61 (0.21)	2	none
Hispanic	0.58 (0.09)	3	none
Non Hispanic black	0.53 (0.06)	4	none
Native Hawaiian or Pacific Islander	0.29 (0.15)	7	none
Non Hispanic Asian	0.48 (0.13)	5	none
Non Hispanic White	0.40 (0.04)	6	none

Critical value of t: 2.815

4. Comparison of Characteristics for the Linked Primary-Duplicate Pair

Housing units coded as duplicates have been linked with their corresponding primary and the address characteristics of each have been compared. The objective is to learn about the nature of duplicate addresses. It was possible for these addresses to be identical as well as to disagree on one or more characteristics. **Table 8** gives the unweighted percentage of primary duplicate pairs that agree on each of six different address characteristics. When both of the linked addresses were missing a characteristic, the pair was not included in the percentage calculation for that characteristic. Results show extensive agreement on zip

code, and census block. Disagreement on zip code and census block suggests that the duplicate housing unit may have been incorrectly geocoded to census block. There was relatively less agreement on house number, streetname and unit designation.

Table 8: Percentage Agreement on Address Characteristics of Primary-Duplicate Pairs

Address Characteristic	Agree	Disagree
Rural Route & Box Number	12.6	87.4
Unit Designation	10.9	89.1
House Number	50.2	49.8
Streetname	54.9	45.1
Zip Code	78.2	21.8
Census Block	78.5	21.5

Table 9 gives the percentage agreement on the six address characteristics by type of enumeration area. The first three levels of the MSA/TEA variable of Table 4 were collapsed to form the mailout/mailback (MOMB) level. Results show that there was relatively more disagreement in the non mailout/mailback areas, particularly in house number and streetname.

Table 9: Percentage Agreement of E-sample Linked Pairs on Address Characteristics by TEA

Characteristic	MOMB		Non MOMB	
	Agree	Disagree	Agree	Disagree
Rural Route & Box Number	0.0	100.0*	12.9	87.1
Unit Designation	10.8	89.2	15.6	84.4
House number	64.9	35.1	18.6	81.4
Streetname	68.9	31.1	26.4	73.6
Zip Code	87.7	12.3	61.2	38.8
Census Block	73.3	26.7	86.1	13.9

*There were only 5 linked pairs falling in this category

Table 10 gives the percentage agreement on all six address characteristics by type of structure. The last two levels of the type of structure of Table 5 were collapsed to form the multi-unit level. Results show that there was relatively less disagreement on the house number, streetname, unit designation, and rural route characteristics in multi-units compared to single units.

Table 10: Percentage Agreement of E-sample Linked Pairs on Address Characteristics by Type of Structure

Characteristic	Single Unit		Multi Unit	
	Agree	Disagree	Agree	Disagree
Rural Route & Box Number	12.4	87.6	13.5	86.5
Unit Designation	4.3	95.7	12.3	87.7
House Number	30.2	69.8	66.8	33.2
Streetname	32.6	67.4	74.0	26.0
Zip Code	69.1	30.9	85.4	14.6
Census Block	78.1	21.9	78.9	21.1

5. Summary and Conclusions

The objective of this study was to document the extent of census housing duplication**,** and the distribution of duplicates by various characteristics of interest. Results of this study can be used to identify characteristics and geographic areas that may be most beneficial to study or target when searching for duplicates. The results can be used to guide unduplication efforts and to help correct erroneous addresses.

The major conclusions are as follows**:**

There was substantial duplication of Census 2000 housing units. Table 1 shows that nearly 25 percent of all erroneously enumerated housing units were duplicates in 2000. Although this figure is lower than the corresponding figure of 33.4 percent duplication in 1990, but it is still significant. Consequently, it is beneficial to conduct duplicate housing unit searches during census operations since successful efforts to unduplicate housing units can result in better housing unit coverage estimates.

Housing unit duplication was not uniform. It varied, sometimes rather widely, by size of urban area, by whether units were single units or multiunits, by the occupancy status of the housing unit, and by the housing tenure of the occupant.

The next major conclusions relate to the location and kind of housing units that were most likely to have duplicates**:**

There was more housing unit duplication in small cities and in rural areas. Table 4 shows that the percentage of housing unit duplication increases as you proceed from large mailout/mailback areas to small mailout/mailback areas and non mailout/mailback areas. The percentage in non mailout/mailback areas is the largest (1.01 percent), and these are sparsely populated and geographically isolated areas where census information is collected by enumerators. The results imply that duplicate search and unduplication efforts should be targeted to small cities and rural areas**.**

There was more housing unit duplication among units in multi-unit structures than among single unit structures. In particular, it was highest in the small multi-unit structures that had between 2 and 9 units at a basic street address. The results suggest that duplicate search and unduplication efforts be targeted to all multi-unit structures in small cities and non mailout/mailback areas, and primarily to small multi-unit structures in the large and medium sized cities.

There was more housing unit duplication among vacant units than among occupied units. Table 6 shows that the vacants had the highest percentage of duplication (2.01 percent). Duplication of vacant units distorts the housing unit count but not necessarily the person count. This result implies that it is beneficial to perform duplicate search on and to unduplicate vacant units.

The final conclusion concerns the nature of duplicate housing unit addresses**:**

Duplicate addresses that referred to the same housing unit seldom were identical. Perhaps this is why they were not detected as possible duplicates by the census duplicate housing unit operation (see section on Limits). Non-city-style (rural route and box number) addresses and unit designations seldom agreed. The agreement percentages were 12.6 percent and 10.9 percent, respectively. House number and street name agreed about 50 percent of the time, while zip codes and census blocks agreed most often. Both were around 78 percent. In general, mailout/mailback areas had less disagreement than non mailout/mailback areas, which suggests that there is potential for address improvement in these non mailout/mailback areas.

REFERENCES

Barrett, D., Beaghen, M., Smith, D., Burcham, J. (2001) "ESCAP II: Census 2000 Housing Unit Coverage Study" internal memorandum, U.S. Bureau of the Census

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 LinearModels: Regression and the Analysis of Variance* (New York: John Wiley and sons), pp 108-9