



UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

December 31, 2002

MASTER FILE

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-52

MEMORANDUM FOR

Donna Kostanich
Chair, A.C.E. Revision II Planning Group

From:

Dawn Haines *deb*
Chair, A.C.E. Revision II Estimation Subgroup

Deborah Fenstermaker *DF*
Chair, A.C.E. Revision II Duplicate Subgroup

Subject:

A.C.E. Revision II: Estimated Correct Enumeration and Residence
Probability for Duplicate Links

The attached document provides correct enumeration and residence probabilities for E-sample cases with duplicate links outside the search area and P-sample cases with links to census enumerations outside the search area, respectively.

A.C.E. Revision II

December 31, 2002

Estimated Correct Enumeration and Residence Probability for Duplicate Links

Deborah Fenstermaker and
Peter Davis

Decennial Statistical
Studies Division

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

EXECUTIVE SUMMARY

The Further Study of Person Duplication in Census 2000 identified sample cases which linked to census enumerations outside the A.C.E. search area. However, the person duplication study could not determine which member of the linked pair was in the correct location. The fact that we do not know which member of the pair is correct does not matter much for national level estimates, but does matter for estimates for geographic areas. Assuming that the linked person does exist, the probability that each member of the link is in the correct location is estimated.

Key Results

- The estimate of the overall correct enumeration probability for E-sample cases with a duplicate link outside the search area is 0.4463.
- For the P-sample nonmovers, the estimate of the overall residence probability for cases with links to census enumerations outside the search area is 0.5092.
- Accounting for E-sample cases with a duplicate link outside the search area contributes to a decrease of approximately 1 percent in the estimate of correct enumerations relative to the original estimate. This is roughly 2.5 million additional erroneous enumerations due to census duplication.
- Accounting for P-sample cases with a link to a census enumeration outside the search area contributes to an increase of approximately 0.47 percent in the estimated match rate relative to the original estimate.

These anticipated changes in the estimation components are limited to one aspect of how measurement error is being incorporated into the A.C.E. Revision II estimates. These estimates do not tell the entire story.

Assumptions

For the E sample, various assumptions could be made about what probability to assign to the two members of the duplicate pair as long as the probabilities sum to one. The simplest would have been to assign a probability of $\frac{1}{2}$ to each member of the link on the basis that each linked pair represents one person. However, thinking that we could do somewhat better, we separated the links into groups where one member of the link is thought to be correct and where it is not clear which member is correct.

For situations where it is reasonably clear one member was correct, we made these assumptions.

- For duplicate links of E-sample persons to group quarters residents, we assumed the group quarters person to be the correct enumeration and the E-sample case to be erroneous. This is the same approach used by Fay in ESCAP II Report #9.
- For duplicate links involving a person 18 years or older who was listed as a child of the householder in one source and not a child of the householder in the other source, we assumed the “not a child” was the correct enumeration and the “child” was erroneous.
- For E-sample person duplicate links who were originally coded as erroneous, we assumed this was the right code and the census enumeration outside the search area was correct.

For situations where it is not clear which member of the linked pair was correct, we assumed that over all of these duplicate links, that the resulting contribution to correct enumerations would be one half the total weighted estimate of links including those originally coded as erroneous. These situations include whole household duplication, children listed by parents living in different households and all other duplicate links.

For the P sample, there was little evidence to guide our assumptions for estimating the residence probability for P-sample cases which link to a census enumeration outside the search area. Therefore, we assumed that the P-sample residence probability for linked cases was like the E-sample correct enumeration probability for duplicate links. For the P-sample cases that matched to an E-sample case within the sample area, this seems to be a reasonable assumption. For P-sample cases that were nonmatched, the appropriateness of this assumption is less obvious. With little else to go on, we used a similar approach to the E-sample for links where it was clear which case was correct. Then, we borrowed the E-sample correct enumeration probabilities for P-sample links where it was not clear which member of the link was in the correct location.

Methods

The Further Study of Person Duplication in Census 2000 identified sample cases which linked to census enumerations outside the A.C.E. search area. However, the person duplication study could not determine which member of the linked pair was in the correct location. Assuming that the linked person does exist, the probability that each member of the link is in the correct location is estimated. This probability is used in forming the A.C.E. Revision II estimates.

The person duplication study did not identify whether the E-sample component of the link was correct or erroneous. Likewise, the duplication study did not identify whether the P-sample component of the link was at the correct Census Day residence. Thus, it was necessary to estimate two conditional probabilities:

- z_t - the probability that an E-sample case is a correct enumeration given that it is a duplicate to another census enumeration outside the search area, and
- h_t - the probability that a P-sample case is a resident on Census Day given that it links to a census enumeration outside the search area.

The requirements for assigning the individual estimates to each link is documented in DSSD A.C.E. Revision II Memorandum Series #PP-25. The general methods are as follows.

The sample links were classified into five situations based on the composition of the linked members. Two situations (1 and 2) describe scenarios where one member of the link is thought to be correct and the other is incorrect. In these cases the conditional probability is set to 1 when the sample case is thought to be correct or 0 when thought to be erroneous.

Situation 1 covers links to persons in group quarters. In this situation the sample case is considered erroneous and the group quarters enumeration is considered correct. Situation 2 covers links of person records that were 18 years or older and listed as a child in one source and as “not a child” in the second source. For this type of link, the child component of the link is considered erroneous and the “not a child” component is considered correct. This was designed to handle adults who were actually living independently, but who were also listed at their parent’s residence, such as college students living off of campus.

For the other three situations (3, 4, and 5), it is not clear which enumeration of the person is in the correct location, but the duplicate study does provide a means for estimating the correct enumerations. For these three situations, we assume that overall, one-half of the E-sample cases with a duplicate link in each situation is correct. Taking into account the original coding¹, the correct enumeration probability for the E-sample links were derived separately by six control

¹For E-Sample person duplicate links who were originally coded as erroneous, we assumed this was the right code and the census enumeration outside the search area was correct.

cells: three Race/Hispanic Origin Domain groups and Tenure. Situation 3 covers whole household duplication. Situation 4 covers children who were listed more than once, such as children in joint-custody situations. Situation 5 is a catch-all group of the remaining links.

The duplicate study does not provide an estimate of correct Census Day residents in the P sample. For these cases, we borrow the resulting z_t 's for E-sample cases with duplicate links and assign that value to P-sample nonmover links which were originally considered residents. The h_t estimate is the same as the corresponding z_t estimate by control cell and situation.

The A.C.E. Revision II estimation methodology adjusts for links to census enumerations outside the search area for P-sample nonmovers. Consequently, the conditional residence probabilities were estimated only for nonmovers.

Results

Table 1 summarizes the approximate impact on the dual system estimation components for six race/Hispanic origin domain by tenure groups. This approximation is limited to one aspect of how measurement error is being incorporated into the A.C.E. Revision II estimates, correcting for links to census enumerations outside the search area. The table gives the original estimate for correct enumerations², matched nonmovers, nonmatched nonmovers, and the match rate for nonmovers from the full sample. The change estimate is the difference between the estimate adjusted for links to census enumerations outside the search area and the original estimate. The percent change estimate is the ratio of the estimated change to the original estimate.

Overall, adjusting for sample cases having a link to a census enumeration outside the search area results in decreases in the estimate of correct enumerations (1.00 percent), matched nonmovers (0.80 percent), and nonmatched nonmovers (6.85 percent). Even though the effect is dramatic between the matched and nonmatched nonmovers, the percent increase of the nonmover match rate is less than the percent decrease in the correct enumerations and correct enumeration rate. Both a decrease of the correct enumeration rate and an increase in the match rate contribute to a decrease in the dual system estimate. Again, this is an approximate impact due to correcting for measurement error due to persons having another residence.

The pattern of the impact on the six Race/Hispanic Origin and Tenure categories is similar to the total. Generally, the percent change is about the same for correct enumerations and matched nonmovers while the magnitude of the percent change for nonmatched nonmovers is larger. Yet, the match rate only increases slightly.

² Note that the estimate of correct enumerations is for the E sample and is not the same as the estimate of census correct enumerations; see Bell (2002).

Table 1: Summary of Approximate Impact on Dual System Estimation Components

Race/Hispanic Origin Domain	Tenure	DSE Term	Original Estimate	Change	Percent Change
Domain 4: Non-Hispanic Black	Owner	CE	14,911,583	-205,844	-1.38
		M	13,312,209	-145,704	-1.09
		NM	1,379,675	-85,548	-6.20
		Match Rate	90.61%	0.44%	0.49
	Non-Owner	CE	13,996,457	-211,306	-1.51
		M	11,602,765	-148,453	-1.28
		NM	2,115,458	-115,702	-5.47
		Match Rate	84.58%	0.56%	0.66
Domain 3: Hispanic	Owner	CE	15,154,347	-141,948	-0.94
		M	13,631,851	-124,795	-0.92
		NM	1,271,353	-78,957	-6.21
		Match Rate	91.47%	0.42%	0.46
	Non-Owner	CE	15,672,304	-173,370	-1.11
		M	13,402,348	-153,218	-1.14
		NM	2,255,568	-119,637	-5.30
		Match Rate	85.59%	0.52%	0.61
Domains 7: Non-Hisp White or Other Race 5: Nat. Hawaiian or Pac. Islander 6: Non-Hispanic Asian 1: Am. Indian or AK Nat. On Res 2: Am. Indian or AK Nat. Off Res	Owner	CE	147,755,404	-1,214,804	-0.82
		M	140,142,141	-886,176	-0.63
		NM	7,674,682	-607,459	-7.92
		Match Rate	94.81%	0.36%	0.38
	Non-Owner	CE	44,606,144	-563,175	-1.26
		M	38,461,619	-381,674	-0.99
		NM	4,455,515	-304,613	-6.84
		Match Rate	89.62%	0.55%	0.62
Total		CE	252,096,238	-2,510,447	-1.00
		M	230,552,932	-1,840,020	-0.80
		NM	19,152,251	-1,311,917	-6.85
		Match Rate	92.33%	0.43%	0.47

Numbers may not add due to rounding

The P-sample components, match, nonmatch and match rate, are for nonmovers.

The estimate of correct enumerations is for the E sample and is not the same as the estimate of census correct enumerations; see Bell (2002).

Table 2 provides estimates of the overall correct enumeration probability for E-sample cases with duplicate links outside the search area as well as the residence probability for P-sample nonmovers with links to census enumerations outside the search area. The nonmovers are further disaggregated by the original match outcome. The estimated probabilities are given by the five linked situations. The first two are situations where one member of the link is chosen to be correct. In situations 3, 4, and 5, it was unclear which member of the link is correct.

The estimated overall correct enumeration probability for E-sample cases with duplicate links is 0.4463. The overall correct enumeration probability is the product of the original correct enumeration probability and the conditional correct enumeration probability. For situations 1 and 2a, the census record was chosen to be correct while for situation 2b, the E-sample record was selected to be correct. The overall correct enumeration probability for situation 2b is less than one because there are erroneous enumerations from the original coding. Situations 3, 4, and 5 were designed to have an estimated overall correct enumeration probability of 0.5 since these were the situations where it was not clear which member of the link was correct.

The estimated residence probability for P-sample nonmovers with duplicate links to a census enumeration outside the search area is 0.5092 overall, 0.5226 for matched cases, and 0.4955 for nonmatched cases. The overall residence probability is the product of the original residence probability and the conditional residence probability. Rules similar to the E sample were applied for determining the conditional residence probability to links in situations 1, 2a, and 2b. For the other three situations, the conditional correct enumeration probability was assigned to the nonmover links which were originally classified as residents. If the P-sample component of the link was originally considered to be a non-resident, then the conditional probability was zero. The resulting overall estimate of residence probability for situations 3, 4, and 5 is not 0.5 like the E sample. This happens because the proportion of originally estimated correct enumerations in each of these situations is different than the proportion of originally estimated residents.

Table 2: Overall Correct Enumeration and Residence Probability by Linked Situation

Linked Situation (E or P) ↔ (Census)		Overall CE Prob	Overall Residence Probability for Nonmovers		
			Total	Matches	NonMatches
1.	(Person in housing unit) ↔ (Person in a group quarters)	0	0	0	0
2a.	(Person 18+, child of reference person) ↔ (Person 18+, not child of reference person)	0	0	0	0
2b.	(Person 18+, not child of reference person) ↔ (Person 18+, child of reference person)	0.9166	0.9318	0.9998	0.8372
3.	(All persons in a housing unit) ↔ (All persons in another housing unit)	0.5000	0.6264	0.6357	0.6207
4.	(Child 0 - 17) ↔ (Child 0 - 17)	0.5000	0.4809	0.5423	0.3906
5.	All Remaining Linked Situations	0.5000	0.5166	0.5735	0.4437
Total		0.4463	0.5092	0.5226	0.4955

References

Bell (2002), “A.C.E. Revision II: Calculating aggregate data-defined, correct enumeration, and census inclusion rates (for groups that involve aggregation across post-strata)” DSSD Revision II Memorandum Series #PP-40, December 31, 2002.

Fay (2001), “ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study,” Executive Steering Committee for A.C.E. Policy II, Report 9, Preliminary Version, October 26, 2001.

Fay (2002), “ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study,” Executive Steering Committee for A.C.E. Policy II, Report 9, Revised Version, March 27, 2002.

Davis (2002), “A.C.E. Revision II: Integration of Duplicate Links into the Full Sample Estimation Files,” DSSD A.C.E. Revision II Memorandum Series #PP- 25, December 31, 2002.