



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-16

PRED CENSUS AND SURVEY MEASUREMENT STAFF MEMORANDUM SERIES:
CSM-A.C.E. Revision II-10R

MEMORANDUM FOR: Donna Kostanich
Chair, A.C.E. Revision II Planning and Management Group
Decennial Statistical Studies Division

From: Mary H. Mulry *signed 12/31/02* 
Chair, A.C.E. Revision II Quality Indicators Group
Statistical Research Division

Through: David Hubble *signed 12/31/02* 
Assistant Division Chief, Evaluations
Planning, Research, and Evaluation Division

Prepared By: Susanne Bean
Mathematical Statistician
Planning, Research, and Evaluation Division

Subject: A.C.E. Revision II Clerical Review of Census Duplicates Sampling
Specification

Attached is the A.C.E. Revision II Clerical Review of Census Duplicates (CRCD) Sampling Specification. Please direct any comments or questions to Susanne Bean 301-763-9590.

cc: DSSD A.C.E. Revision II Memorandum Series Distribution List
R. Killion
D. Hubble

CLERICAL REVIEW OF CENSUS DUPLICATES SAMPLING SPECIFICATION

1. BACKGROUND

The primary goal of the Clerical Review of Census Duplicates (CRCD) is for analysts at the National Processing Center (NPC) to examine the quality of the estimates of duplicate enumerations in the census. The Accuracy and Coverage Evaluation Revision II estimates will use estimates of duplicate census enumerations produced by the Further Study of Person Duplication (FSPD). The analysts will review housing units with two or more duplicate links identified by the FSPD (referred to as 2+ links) and duplicates identified by another evaluation of FSPD, the Census and Administrative Records Duplication Study (CARDS).

2. UNIVERSE

The universe of links for this review contains all E-sample eligible links from the statistical matching phase (2+ links) of the FSPD linking plus all E-sample eligible links only found by CARDS. The universe is restricted to links to target records outside of the A.C.E. surrounding blocks. For the P-sample, the universe is further restricted to links where the source record is a nonmover (resident or nonresident).

Two important notes about the CRCD universe:

- We originally believed that the target file CARDS used did not include Reinstates and Deletes. Therefore, when we attempted to restrict the CARDS only links to E-sample eligible census records we only dropped links to Group Quarters. Therefore, there are CARDS only links to Reinstates and Deletes in the realized CRCD universe that were never meant to be included.
- For the P-sample, the A.C.E. Revision II estimates will be based on duplicates of nonmover residents only. However, both nonmover residents and nonresidents are included in the CRCD universe.

3. INITIAL SELECTION OF CLERICAL REVIEW WORKLOAD

We have decided to use the Evaluation Followup (EFU) sample of clusters as the base for our selection of the clerical review workload. This simplifies the selection and weighting, while allowing us to overlap with the cases being used for other A.C.E. Revision II work.

The EFU sample sizes for this universe are:

P-sample = 126,542 links in 63,124 household (HHs)

E-sample = 120,175 links in 59,427 HHs

We estimate that given a two week operation, a reasonable workload size is about 10,000 HHs. Thus, we need to subset the universe of links to get a more manageable workload.

To reduce the clerical review workload, we propose removing from the review all links in housing units where every link in the household meets the following three conditions:

- C1. *Low Probability FSPD Only* - Link only found by FSPD and has a statistical Probability of No Trial Having Observed Outcome which is less than 2/3 of the preset FSPD statistical cutoff for determining whether to classify the link as a duplicate AND/OR
- C2. *Measurement Other Residence* - Final measurement group why code for the link indicated "Other Residence" AND/OR
- C3. *FSPD and CARDS Duplicate* - Link was found to be a duplicate by FSPD (meaning the statistical probability met the preset cutoff to be considered a duplicate) and CARDS also identified the link.

The logic behind these conditions is that we want to concentrate resources on reviewing links where we really need more information to determine whether the links are actually duplicates. The links that fail C1 are not very likely to actually be duplicates. We already have information from the Measurement Review to support the decision to consider the links that fail C2 duplicates since they admitted to having another residence. Finally, if the FSPD and CARDS are in agreement that the case is a duplicate than the record is likely a duplicate.

If we subset using the first condition only, the workload sizes would be:

- P-sample = 8,465 links in 5,523 HHs
- E-sample = 10,248 links in 6,412 HHs

Subsetting using all 3 conditions, the initial clerical review workload sizes are:

- P-sample = 6,183 links in 4,496 HHs
- E-sample = 8,911 links in 5,656 HHs

Therefore, the total initial clerical review workload is 10,152 HHs which is about the same as our estimate of what can be completed in the time allotted.

4. CONTINGENCY PLAN AND WORKLOAD ORDER

The initial clerical workload includes 10,152 work units (households). Subject matter experts advised that we must plan for the contingency that only 80 percent of this workload might be completed. Of the 2,259 total clusters in the Evaluation Followup (EFU) sample, only 1,737 clusters have clerical cases to constitute the sampling universe. We will select a sample of these clusters such that approximately 20 percent of the workload will be in reserve for the contingency plan.

4.1 Certainty Criteria

We will use the following certainty criteria to exclude cases from the universe of potential reserve clusters:

- Clusters with at least 25 Clerical cases
- Clusters with at least one link with weight (using the definition below) greater than 9,000
- Clusters with a total weight (using the definition below) greater than 80,000

The term “weight” refers to the P- or E-sample baseline weight (person weight calculated using the production results for just the EFU sample). The term “total weight” refers to the combined weighted total number of P- and E-sample links based on the baseline weight.

These criteria attempt to retain with certainty clusters that contain most of the clerical cases and most of the total weight.

4.2 Selection of Reserve Group

To pick the reserve group, we will sort the non-certainty clusters in the same order in which they were sorted for selection into the EFU sample, by sampling strata and cluster size (STRATUM and SORTVAR copied from the BFUSAMP file maintained by PRED). Once sorted, we will select a 1-in-4 systematic subsample to place clusters into the reserve group. The remaining clusters and the certainty clusters will be the group of clusters clerks will work first. We call this group of clusters the main group.

4.3 Workload Order and Bias Concern

If the work was simply assigned in cluster number order, learning curve and coder effects could bias the results. (Note: Cluster numbers are associated with area of the country.)

To mitigate both the learning curve and coder effect concerns, we will assign “batch” characters for each source cluster based on the last digit of the cluster number. The batch character will be in the first position of the work unit identifier, thus sorting by work unit identifier will not put the work in cluster order.

Specifically, once we have separated the clusters into the main and reserve groups, batch will be assigned as follows:

Group	Last Digit of Source Cluster									
	0	1	2	3	4	5	6	7	8	9
Main	a	b	c	d	e	f	g	h	I	j
Reserve	k	l	m	n	o	p	q	r	s	t

4.4 Application of Contingency Plan

The software contractor (Gunnison) will allocate the units in batches a - j (the main group) to each analyst's directory at the beginning of the operation. They will place all units in batches k-t (the reserve group) in a separate directory to be assigned to analysts later in the operation if time permits.

4.5 Results of Contingency Plan

Of the 1,737 clusters requiring clerical review, we selected 409 clusters for the reserve group and 1,328 clusters for the main group (of which 100 clusters were selected with certainty).

The reserve group size for the contingency plan is:

P- & E combined = 3,004 links in 2,031 HHs.

This meets the expectation of 20 percent of the workload being in the reserve group (2,031 / 10,152 \approx 20%).

Note: In the end, we were able to complete all 10,152 work units in the initial clerical review workload (all units in both the main and reserve groups).

5. ADDITIONAL CASES FOR CLERICAL REVIEW

Towards the end of the operation, we decided to review some additional cases that were originally excluded from the clerical review. Since the analysts finished the initial clerical review workload a little earlier than anticipated and the additional cases would improve the analysis, we decided to add back the links that were originally excluded due to conditions C2 and C3 in Section 3 above.

The number of additional cases are:

P-sample = 2,282 links in 1,027 HHs

E-sample = 1,337 links in 756 HHs

Therefore, the final clerical review workload sizes are:

P-sample = 8,465 links in 5,523 HHs

E-sample = 10,248 links in 6,412 HHs

6. SAMPLING WEIGHTS AND FINAL OUTPUT

Since the contingency plan was not necessary, we can use the EFU sampling weights as the clerical review sampling weights.

I have created two SAS datasets in pred_cover2:[pred.s_bean.dups] called pdupsamprev_addvars and edupsamprev_addvars, which have the sampling weights as well as some additional variables which may be useful in the CRCD analysis. These datasets have all the links in the final clerical review workload (including those additional ones described in Section 5). Thus, the P-sample file has 8,465 links and the E-sample file has 10,248 links. Here is the layout of these files:

Variable	Description	Format	File*
	/* Identifiers */		
SID	Source ID P- Cluster (6)/ MSN (4)/ WMSN (5)/ MOVER (1)/ Person # (2) E - CID (12)/ SEQ (2)/ Person # (5)	\$19.	B
TID	Target ID CID (12)/ SEQ (2)/ Person # (5)	\$19.	B
	/* Subsetting Conditions for Initial Clerical Review Workload */		
MEASOR	Measurement Other Residence Indicator 1 = meets condition C2 0 = otherwise	1.	B
EASYDUP	FSPD and CARDS Duplicate Indicator 1 = meets condition C3 0 = otherwise	1.	B
ORIGSAMP	Initial Clerical Review Indicator 1 = Link in the initial clerical review 0 = Additional link	1.	B
	/* CARDS Classification of Link */		
DUP	Original CARDS Classification of Link C = Confirmed / Found in both FSPD and CARDS F = FSPD only M = CARDS only	\$1.	B
DUP2	Final CARDS Classification of Link Confirmed/Found in Both C = Originally classified as C (DUP=C) CA = FSPD link not confirmed by CARDS until StARS address matching phase (after clerical review) CB = CARDS link not on FSPD file of E-sample Eligible 2+ Links, but is found on FSPD file which includes Bob's exact links & links to non E-sample eligible records Other Values D = FSPD link Denied by CARDS U = FSPD link that CARDS cannot confirm or deny M = CARDS only, not in any of the FSPD links	\$2.	B

Variable	Description	Format	File*
	/* Useful FSPD Variables Which Were Also Created for CARDS */		
GEOCAT	Geographic Distance Category 1 = Within Cluster 2 = Surrounding Blocks 3 = Within County 4 = Different County, Same State 5 = Different State	1.	B
GEOCAT2	Geographic Distance Category Used for Statistical Cutoff Values 1 = Block 2 = Tract 3 = Within County 4 = Different County, Same State 5 = Different State	1.	B
ETARGET	Type of Census Record (Target) 1 = E-sample Eligible 2 = Group Quarters 3 = Reinstate 4 = Delete	\$1.	B
LINK_TO_ESAMPLE	Link to E-sample Indicator 1 = Target is an E-sample record 0 = Otherwise	1.	E
	/* Weights and Components */		
UDUPPROB	Multiplicity Factor (Always 1 for P-sample)	5.3	B
RPROB	A.C.E. Residence Probability	10.8	P
BPFINWGT	P-sample Sampling Weight (Same as EFU Baseline Weight)	15.6	P
BEFINWGT	E-sample Sampling Weight (Same as EFU Baseline Weight)	15.6	E
PCLERWGT	P-sample Final Clerical Weight w/o RPROB (BPFINWGT * UDUPPROB = BPFINWGT)	15.6	P
ECLERWGT	E-sample Final Clerical Weight (BEFINWGT * UDUPPROB)	15.6	E
RPCLERWGT	P-sample Final Clerical Weight w/ RPROB factor (BPFINWGT * UDUPPROB * RPROB = BPFINWGT * RPROB)	15.6	P

* - The notations in the “File” column indicate which output file the variable is on. B indicates it is on both files, E indicates it is on the E-sample file only, and P indicates it is on the P-sample file only.