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

December 30, 2002

MASTER FILE

DSSD A.C.E. REVISION II MEMORANDUM SERIES # PP - 15

PRED CENSUS AND SURVEY MEASUREMENT STAFF MEMORANDUM SERIES:
CSM-A.C.E. Revision II -11R

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

From: Mary H. Mulry ***signed 12/30/02***
Chair, A.C.E. Revision II Quality Indicators Group
Statistical Research Division

Through: David Hubble ***signed 12/30/02***
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 Census and Administrative Records Duplication Study Plan

Attached is the A.C.E. Revision II Census and Administrative Records Duplication Study Plan. 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

Census and Administrative Records Duplication Study

Susanne L. Bean and D. Mark Bauder, Planning, Research, and Evaluation Division

1. BACKGROUND

The primary goal of the Census and Administrative Records Duplication Study (CARDS) is to use administrative records to examine the quality of the estimates of duplicate enumerations that will be used in the Accuracy and Coverage Evaluation (A.C.E.) Revision II estimates.

1.1 A.C.E. Revision II Estimates

Based on findings from the Executive Steering Committee for A.C.E. Policy (ESCAP) reports, duplicates are one of the major sources of error which the revised estimates will attempt to address. Another source of error identified in the ESCAP reports is measurement error as detected by the Measurement Error Reinterview (MER). ESCAP Report 9 (Revised): Evidence of Additional Erroneous Enumerations from the Person Duplication Study attempts to combine both sources of additional erroneous enumerations, duplicates and measurement error, to examine the impact on the Dual System Estimates (DSEs). The A.C.E. Revision II operation will extend this work to produce revised estimates that incorporate the effect of erroneous enumerations missed in the original A.C.E. estimates.

1.2 Duplication in the Census

Census 2000 Evaluation O.16: Person Duplication in the Search Area Measured by the Accuracy and Coverage Evaluation found that the estimate of duplicate census enumerations measured by A.C.E. was less than the estimate from the 1990 Post Enumeration Survey (PES). ESCAP II Report 20: Person Duplication in Census 2000 addressed this concern using the results of a computer matching operation to determine the extent of census duplication. This operation extended the search to include units which were out-of-scope for the A.C.E. but would have been in-scope for the PES. They found an additional1.2 million duplicate census enumerations in units that were out-of-scope for the A.C.E. but would have been in-scope for the PES.

The ESCAP report also found some intuitive patterns of census duplications by race/ethnicity and age/sex groups. There were higher percentages of duplicate enumerations for the Non-Hispanic Black and the Hispanic domains. These were concentrated outside the one ring of surrounding blocks of a cluster but still within the same county. Duplication for persons 50 years of age or older was seen more in a different state. The 18-29 year-old categories had higher percentages of duplicate enumerations between housing units and group quarters than the other age/sex categories. The female duplication for this age group was predominantly in college dorms while the males were duplicated in college dorms, correctional facilities, and military group quarters.

A similar methodology will be used in the Further Study of Person Duplication in Census 2000 (FSPD) to estimate and identify duplication in order to make adjustments for the A.C.E. Revision II estimates. Using a computer matching algorithm, the study will perform a national match of E-sample and P-sample records to census enumerations on the Hundred Percent Census Unedited File (HCUF). (Note: links between the P-sample and the HCUF are referred to as duplicates in this study even though they are really matches between the two different enumeration processes.)

1.3 Census and Administrative Records Duplication Study (CARDS)

CARDS will use the Census Numident File and the Statistical Administrative Records System 2000 (StARS 2000) to examine the effectiveness of the FSPD methodology. CARDS will attempt to confirm or deny duplicate links identified by the FSPD (referred to as computer duplicates here). In addition, CARDS can identify duplicates missed by the computer study.

CARDS is the first study in a series of planned research using data from the Administrative Records Duplicate Link Research project. The goals of future research using this data are to analyze the nature of the duplication to reduce census duplication in 2010 and to provide data to StARS 2000 to aid in evaluation of decisions made during the construction of the system.

2. QUESTIONS TO BE ANSWERED

This study answers the following questions:

- How much duplication was there in Census 2000?
  - How many duplicates were there overall?
  - What was the extent of duplication within the cluster and one ring of surrounding blocks? in all areas outside of the surrounding blocks? outside of the surrounding blocks but within the same county? in a different county but within the same state? beyond the same state?
  - What were the patterns of duplication for the Race/Ethnicity domains? for the Age/Sex categories? for duplicate poststrata? for one versus multiple person households? (This will be done if time permits.)

- Overall, how effective was the methodology used in the Further Study of Person Duplication in Census 2000? (This will be done if time permits.)
  - How many computer duplicates can CARDS confirm?
  - How many computer duplicates does CARDS determine to be incorrect?
  - How many computer duplicates does CARDS not have enough information to classify as confirmed or incorrect and thus are undetermined?
  - How many duplicates did CARDS find that the computer study missed?

- How effective were each of the rules selected by the duplication group? (This will be done if time permits.)

3. METHODOLOGY

3.1 Obtain Protected Identification Keys (PIKs) and StARS addresses for E- and P-Sample People

The StARS database, created by the Administrative Records Staff incorporates data from seven files:

- Internal Revenue Service Individual Master File (1040),
- IRS Information Returns File (W-2 / 1099),
- Department of Housing and Urban Development Tenant Rental Assistance Certification System File,
- Department of Housing and Urban Development's Multifamily Tenant Characteristics System File
- Center for Medicare and Medicaid Services Medicare Enrollment Database File,
- Indian Health Services Patient Registration System File,
- Selective Service System Registration File.

In addition, the "Census Numident", a lookup file was created. This file was created from the Social Security Administrations Numerical Identification File (Numident). The Numident was edited, and for confidentiality reasons a Protected Identification Key (PIK) was created for each Social Security Number (SSN). A separate file was created which contains all addresses from the IRS 1040 and 1099 files for each person. This file is called the Geokey Numident.

Via a match between the HCUF and Geokey Numident, SSNs were found for HCUF people and added to the HCUF person record. We call the resulting file the HCUF Research File. We use these files to find all StARS addresses for people in either the P-Sample or the E-Sample. The steps are as follows.

- P-Sample people are linked with Census Numident records by a matching process using address and person fields. The P-Sample people are then linked with HCUF Research File records and to StARS addresses by SSN.

- E-sample people are linked to the HCUF Research File by Census ID, then linked to StARS addresses by SSN.

- StARS addresses for each person are then obtained by linking by SSN to StARS.

For the analyses below, PIKs rather than SSNs are always used. We will use the terms "PIK" and "SSN" interchangeably.

- The above processes result in six files:
  - P-sample ID/PIK file
  - Census IDs for the P-sample PIKs
  - PIK/StARS addresses for the P-sample PIKs
  - The above three, but for E-sample people rather than P-sample

The E- and P- Sample ID/PIK files are combined with the Census ID/sample PIK to create files that are in a format that matches the output from FSPD. This format is one record per duplicate pair.

3.2 Confirming Duplicates

We attempt to confirm possible duplicates of E-Sample and P-Sample people found by FSPD. Where FSPD has proposed a duplicate (a "FSPD duplicate") of an E-Sample or P-Sample person, we determine whether the HCUF Research File has the same SSN for the sample person and The FSPD duplicate person.

- If the E- or P- sample person and The FSPD duplicate have the same SSN, we consider the FSPD duplicate link to be confirmed.

- If the FSPD duplicate has a different SSN from the E- or P- sample person, we judge the FSPD duplicate to be incorrect.

- If an E- or P- Sample person has an SSN, but the FSPD duplicate does not, each of the StARS addresses for the sample person are matched by computer against the address for the duplicate. If a StARS address for the sample person matches to the HCUF address for the duplicate, we also consider the FSPD duplicate link to be confirmed. Otherwise, the link is judged to be undetermined.

3.3 Finding Duplicate Links not Found by FSPD

We attempt to find duplicates of sample people that FSPD did not find. For each sample person, we find all HCUF Research File person records with the same SSN. Where two HCUF records have the same SSN, CARDS has identified the people as duplicates. If FSPD did not obtain a link between these two HCUF people, then we consider this to be a FSPD missed duplicate.

3.4 Evaluating the FSPD process

If time permits, this study will evaluate the rules used in the FSPD process for identifying duplicates. These include rules about how fields used in the linking process were used. These fields include name fields, date of birth fields, and age. Where CARDS finds a missed duplicate, we identify how duplicates CARDS found compare in these fields.

For some linked pairs, FSPD associated a probability the pair represents a duplicate, rather than judging it to be a duplicate or not. In those cases, we will determine whether CARDS has identified the link as a duplicate. These FSPD probabilistic links will be judged to be verified, incorrect, or undetermined, as in the above processes. We will determine how the verified, incorrect, and undetermined links vary with the probabilities assigned in the FSPD process.

3.5 Patterns of Census duplication

If time permits, this study will analyze how the rates of verified, incorrect, undetermined, and missed links vary with geographic categories, demographic categories, and housing unit or group quarters (GQ) characteristics. We will include, where possible, analyses by those groups used in (Mule, 2001).
The demographic groups include:

- **Race/Hispanic Origin:**
  - American Indian on Reservation,
  - American Indian off Reservation,
  - Hispanic,
  - Non-Hispanic Black,
  - Hawaiian and Pacific Islander, Non-Hispanic Asian, and
  - Non-Hispanic White or Some Other Race.

- **Age/Sex:**
  - age 0-17
  - age 18-29 males
  - age 18-29 females
  - age 30-49 males
  - age 30-49 females
  - age 50+ males
  - age 50+ females

These analyses will also be done for each of the above demographic categories by the following geographic groups:

- within ACE cluster,
- within surrounding blocks,
- within the same county,
- within the same state,
- in a different state.

We will also perform analyses for the demographic groups by housing unit/GQ type. These categories are:

- housing unit
- group quarters type:
  - correctional institution
  - nursing home
  - juvenile institution
  - college dorm
  - military
  - other

We may perform analyses for other groups as well.

3.6 Analyses based on the confidence of links found through StARS.

The matching process used to add SSNs to the HCUF, and the process used to link P-Sample people with StARS people, had two phases. In the first phase, address data were used to limit the search for matches for people. In the second phase, a search for matched people was done that required no agreement in address fields. In the former case, we are more confident of the link than in the latter. Thus, the analyses above will be done both with the links from just the first phase, and with all links found by StARS.

4. DATA REQUIREMENTS

- We require the following files from the PRED Administrative Records Research Staff:
  - HCUF Research File
  - Census Numident
  - Files with StARS addresses

- We require the following additional files:
  - Output of the FSPD from DSSD (Tom Mule)
  - CARDS Analysis File from PRED/DSSD

5. DIVISION RESPONSIBILITIES

- The Planning, Research, and Evaluation Division (PRED) will be responsible for managing the study, creating specifications, matching the E- and P-sample files to the administrative records, creating output files from the administrative records matching process, creating the merged analysis files, and preparing the report.

- The Decennial Statistical Studies Division (DSSD) will be responsible for creating the P-sample Input File, adding additional variables to the analysis files, and providing a file of links and estimates from the Census Computer Duplicate Study.

- PRED, DSSD, and the Statistical Research Division (SRD) will collaborate to develop analysis plans, obtain IRS approval for the matching to IRS files, and conduct the data analysis.

- See the milestone schedule for more specific information regarding responsibilities.

6. MILESTONE SCHEDULE

ACTIVITY	WHEN	WHO
Match HCUF	6/3/02 - 6/28/02	Deb Wagner
Develop P- sample Input File specifications	6/6/02 - 7/8/02	Susanne Bean & Deb Wagner
Create P- sample Input File	6/7/02 - 6/28/02	Sue Odell
Develop analysis plans	6/19/02 - 7/26/02	Mary Mulry, Susanne Bean, Mark Bauder, Tom Mule, & Rita Petroni
Develop analysis file specifications	7/1/02 - 8/09/02	Susanne Bean
Obtain approval of Administrative Records use from IRS	7/31/02	Rita Petroni & Mary Mulry
Match P- sample	8/1/02 - 8/23/02	Deb Wagner
Create analysis files	8/9/02 - 9/3/02	Deb Wagner, Sue Odell (?)
Generate list of all addresses from STARs for E- and P-sample IDs that received PIKS	8/5/02 - 8/23/02	Deb Wagner
Provide file of computer duplicate links for evaluation	8/30/02	Tom Mule
Provide computer duplicate estimates by nation & domain	8/30/02	Tom Mule
Conduct analysis	9/4/02 - 11/1/02	Susanne Bean, Mark Bauder, Mary Mulry, Tom Mule, & Rita Petroni
Draft Report	11/4/02 - 11/22/02	Susanne Bean & Mark Bauder

7. LIMITATIONS

There are several ways in which the process outlined above may fail to detect duplicates, or may link false duplicates.

- Some person records on the HCUF Research File were not linked with StARS people. Thus, they do not have SSNs/PIKs associated with them. Because of this, some duplicates may not be found in the CARDS process.
- Not all StARS addresses were linked with addresses in the HCUF Research File. If some of the remaining StARS addresses are truly HCUF addresses, them the CARDS may fail to find some duplicates. However, many of the StARS addresses that were not linked with HCUF Research File addresses contain address data that will be available for additional matching to HCUF Research File addresses.
- Because StARS is created from administrative records, a person can be duplicated at different addresses, yet StARS fail to have records from both addresses.
- Some people have two SSNs, and more than one person can have the same SSN. If one sample person has two SSNs, then CARDS may fail to find that person's duplicate. If more than one person has the SSN of a sample person, then CARDS may falsely call them duplicates.
- In a small number of cases, an SSN was attached to several (as many as 24) HCUF Research File person records. If an E- or P- Sample person's SSN is attached to several HCUF Research File people, we may be hesitant to consider all of those to be duplicates.

8. RELATED STUDIES

- Data created for CARDS may be used in subsequent project during the planning for 2010.

- CARDS will use administrative records to examine the effectiveness of methodology used in the Further Study of Person Duplication in Census 2000 (FSPD).

- CARDS will examine one component error, the estimation of duplicate enumerations, in the A.C.E. Revision II estimates. Additional A.C.E. Revision II evaluation studies will assess other component and relative errors. One of the studies will clerically review duplicates identified by the FSPD and CARDS. (See Chapter 7 of the A.C.E. Revision II Plans.)

9. REFERENCES

Fay, Robert E. (2002). "Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee on A.C.E. Policy II Report 9 (Revised). U.S. Census Bureau, Washington, D.C.

Jones, John and Roxanne Feldpausch (2001). "Person Duplication in the Search Area Measured by the Accuracy and Coverage Evaluation," Census 2000 Evaluation O.16. Initial Draft dated July 23, 2001.

Davis, M.C. and Biemer, P. (1991b). Measurement of the Census Erroneous Enumerations - Clerical Error Made in the Assignment of Enumeration Status. 1990 Coverage Studies and Evaluation Memorandum Series, #L-2 dated July 11, 1991.

Mule, Thomas (2001). "Person Duplication in Census 2000," Executive Steering Committee on A.C.E. Policy II Report 20. U.S. Census Bureau, Washington, D.C.

Mulry, Mary (2002). "Chapter 7: Assessing the Estimates," A.C.E. Revision II Plans. Bureau of the Census internal memorandum, June 24, 2002 DRAFT.