Journal of Official Statistics, Vol. 22, No. 4, 2006, pp. 655–679

Evaluation of Estimates of Census Duplication Using Administrative Records Information

Mary H. Mulry[1], *Susanne L. Bean*[1], *D. Mark Bauder*[1], *Deborah Wagner*[1], *Thomas Mule*[1], *and Rita J. Petroni*[1]

The U. S. Census Bureau used administrative records to examine the quality of the estimates of duplicate enumerations in Census 2000. The estimated number of duplicates in the Census 2000 count of 281,421,906 people was 5.8 million, based on census information only. The identification of duplicates was possible because Census 2000 was the first census to use optical character recognition technology that permitted converting all the names on the census questionnaires into electronic format. Although additional field work to evaluate the duplicates was not practical by the time they were discovered, an administrative records database derived from seven government record files was available for evaluating the estimates of census duplicates. The database had census-like records and was formed using new methodology as part of the research associated with Census 2000. The evaluation using administrative records validated that there were a large number of duplicate enumerations in Census 2000. A clerical review also offered a validation of the estimates of duplicate enumerations based on only census information but raised questions about the accuracy of some types of duplicates identified only with administrative records. The results pointed to areas for refinements in evaluating census duplicates with administrative records. The validation of the estimates of duplicate enumerations through the use of administrative records demonstrated further potential for contributions from administrative records in the evaluation of census duplication.

Key words: Census 2000; undercount; overcount; Accuracy and Coverage Evaluation Survey, A.C.E. Revision II; record linkage.

1. Introduction

The U. S. Census Bureau used administrative records to examine the quality of the estimates of duplicate enumerations in Census 2000. The estimate of the number of duplicates, which used census information only, found 5,826,478 duplicates in the Census 2000 count of 281,421,906 people (U.S. Census Bureau 2003b). The evaluation using administrative records generally validated the estimate of duplicate enumerations and produced an estimate of 6,653,171 duplicates. A clerical review also offered a validation

U. S. Census Bureau, Washington, DC 20233, U.S.A. Email: mary.h.mulry@census.gov

[1] Mary H. Mulry is a Principal Researcher, Susanne L. Bean, D. Mark Bauder, and Thomas Mule are Mathematical Statisticians, Deborah Wagner is a Computer Programmer, and Rita J. Petroni is an Office Chief at the U.S. Census Bureau, Washington, DC 20233, U.S.A. This article is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

of the estimates of duplicate enumerations and provided insight about when the evaluation disagreed.

Census 2000 was the first census where a computerized search to identify duplicates was possible. The use of optical character recognition technology permitted converting all the names on the census questionnaires into electronic format. The estimation of census duplicates used a combination of computerized record linkage and estimation methodology. Though additional fieldwork was not practical, new methodology for forming an administrative records database for the U. S. provided the means for evaluating the estimates of census duplicates using computerized record linkage. The validation of the estimates of duplicate enumerations through the use of administrative records demonstrated further potential for contributions from administrative records in the evaluation of census duplication.

The causes of duplicate enumerations in the census appear to be somewhat different when a person is duplicated within the same block or within adjacent blocks than for duplication across some distance, say different parts of a county or even in different states. Duplication within the local area, consisting of the block or adjacent blocks, appears to have operational causes. A housing unit mistakenly may appear more than once in the census address file, listed twice with different addresses. Mail mis-deliveries in apartment buildings may result in a household getting two questionnaires and responding to both. Enumerators following up for nonresponse to mail questionnaires may make address errors and enumerate a household for a second time, leaving the correct household not enumerated.

In contrast, duplication across a longer distance appears to be caused more by confusion about census residency rules. In a reinterview study, limited by the amount of time since Census Day, (Smith 2004) contacted some duplicates identified electronically in different states and counties in preparation for designing the questionnaire for the 2004 Census Test (U.S. Census Bureau 2003a). The study's other limitation that applies to all the U.S. Census Bureau's fieldwork is that confidentiality restrictions forbid revealing information collected at one address to the household at a different address, even if one person may possibly be a member of both households. The study tested probes designed to help respondents report alternative addresses and was able to identify several causes of census duplication. For example, people who move around Census Day may be enumerated at both the old and new address. College students may be enumerated at their parents' address and their college address although their college address is the correct place according to census residence rules. Children in shared custody arrangements may be enumerated twice, once by each adult sharing custody. People who have more than one residence may be enumerated at both.

The estimate of duplicate census enumerations was used in the estimates of coverage error in Census 2000 produced by the U.S. Census Bureau's Accuracy and Coverage Evaluation (A.C.E.) Revision II program (U. S. Census Bureau 2003b). These estimates of duplicate enumerations demonstrated that duplicate enumerations occurred in the census much more frequently than previously observed or suspected in Census 2000 or other censuses. Since past censuses did not have the names on all census questionnaires in electronic format, coverage measurement surveys were able only to observe duplicates within their sample area (generally a block or group of contiguous blocks) or when a mover into the sample area was found to be enumerated at a reported previous address as well as at the new address. For the 1990 Census, the estimated duplication rate within the block and adjacent blocks was 1.6 percent (Hogan 1993). A separate estimate of movers enumerated in two places in the 1990 Census was not reported, although an evaluation reported some were not detected (West et al. 1991).

This article provides an overview of the methodologies for the estimation of duplicates in Census 2000 using census information only and for the evaluation that identified duplicates using information from administrative records. In addition, a comparison of the two sets of estimates of census duplicates includes the results of a clerical review. More detailed results can be found in Bean and Bauder (2002), Mulry et al. (2003), and Bauder (2004).

2. Background

For context, the U.S. Census Bureau evaluated how well Census 2000 counted the population by conducting a coverage measurement survey known as the Accuracy and Coverage Evaluation survey (A.C.E.). The U.S. Census Bureau considered adjusting the Census 2000 population total to correct for coverage error on three occasions, but each time decided not to adjust.

The original A.C.E. dual system estimates in March 2001 indicated a 1.18 percent undercount in the Census 2000 count of 281,421,906. The U.S. Census Bureau followed a prespecified decision process and decided not to adjust the census numbers for redistricting because of differences between the A.C.E. estimates and Demographic Analysis and other anomalies that could not be explained by the deadline of April 1, 2001 (U.S. Census Bureau 2001c).

Next, the U.S. Census Bureau considered adjusting Census 2000 for purposes other than redistricting when the results of evaluations planned for the summer of 2001 became available (U.S. Census Bureau 2001b). The decision in October 2001 was to use the unadjusted census for purposes other than redistricting because the evaluations of the A.C.E. found a large number of erroneous enumerations, many of them duplications, that the A.C.E. failed to detect (U.S. Census Bureau 2001a). At that time the U.S. Census Bureau issued an A.C.E. Preliminary Revised estimate of a 0.06 percent net undercount (Thompson, Waite, and Fay 2001). The U.S. Census Bureau also decided to conduct further research on census duplication and other measurement errors and to consider whether to incorporate a revised net undercount estimate into the census base used in the intercensal estimates program.

The revision, A.C.E. Revision II, estimated the percent net undercount to be -0.5 percent (an overcount). Although the A.C.E. Revision II estimates appeared to be an improvement over the A.C.E. estimates, the U.S. Census Bureau decided not to adjust the census base for the intercensal estimates because of technical limitations (U.S. Census Bureau 2003b). The concern about the A.C.E. Revision II estimates focused on the limited race information in the data available for the correlation bias adjustment, the estimated extreme overcounts for some small places that could not be validated or explained, and the inconsistency with the Demographic Analysis estimate for children ages 0–9 that is believed to be highly accurate since the basis is birth records.

The evaluation of the improved estimates of duplication was part of the evaluation program for the A.C.E. Revision II estimates (Mulry and Petroni 2003). By the time a computerized search of the census provided evidence of a large number of duplicate census enumerations in October 2001 (Thompson, Waite, and Fay 2001), field tests for confirmation were not practical.

However, the Statistical Administrative Records System (StARS), created with the U.S. Census Bureau's newly developed administrative records database methodology (Leggieri, Pistiner, and Farber 2002; Judson 2000), provided the possibility of evaluating the estimates of duplicate enumerations without fieldwork. The formation of an estimate of duplicates using administrative records coincided with the preparation of the estimates of duplicate enumerations (Mule 2002) for the A.C.E. Revision II.

The A.C.E. Revision II estimation (U. S. Census Bureau 2003b) used a dual system estimator that incorporated adjustments for errors discovered after the data processing for the original A.C.E. The adjustments accounted for duplicate enumerations found by the computerized search and for measurement error (Raglin and Krejsa 2001) and matching error (Bean 2001) detected by A.C.E. evaluations conducted on a subsample of the A.C.E. sample. In addition, the A.C.E. Revision II estimation included a correction for model error, known as correlation bias, detected by Demographic Analysis (Robinson and Adlakha 2002). The A.C.E., and therefore A.C.E. Revision II, used two overlapping samples to produce the estimates, a sample of census enumerations (E-sample) for estimating erroneous enumerations and a sample of the population (P-sample) for estimating census omissions using the A.C.E. The two samples overlapped by using the same block clusters. The universe for the A.C.E. was people living in housing units and did not include those living in group quarters.

A.C.E. Revision II estimation used different methods for estimating duplicates within and outside the local area. Estimates of duplication within the local area were based on the results of clerical matching because they were better than computerized record linking. The local area of the sample block and sometimes adjacent blocks is called the A.C.E. search area since it is the area where the clerical processing was done. For estimating duplicates between an enumeration in the E-sample and another outside the local area, A.C.E. Revision II used the results of a computerized record linking. The computerized record linking for A.C.E. Revision II also estimated duplication within the A.C.E. search area, and those results are reported in this article. The processing for the original A.C.E. did not include people in housing units identified as potential duplicates by a computerized search for Census 2000 to identify housing units duplicated on the census address list (Fay 2001). Some of the housing units identified as potential duplicates were reinstated in the census address list and some were deleted. A.C.E. Revision II estimation, however, included links outside the A.C.E. search area to people in housing units reinstated in the census and those deleted. The reason for including the links to the enumerations reinstated and deleted by the computerized search is that A.C.E. Revision II estimation accounted for the possibility that the wrong record was deleted.

The evaluation of the duplication estimates used the results of another U.S. Census Bureau research project's match between the unedited census file of enumerations of persons and an administrative records file to estimate duplication in the census. Below are brief descriptions of the computerized record linking processes for A.C.E. Revision II and the evaluation with administrative records, which identified links between the E-sample and all census enumerations. Similar methodology was used to link the P-sample people with the census enumerations for the purpose of identifying possible data collection error, and the results were similar to those for the E-samples. To save space, we restrict the discussion to the E-sample record linkage. The linking used Census 2000 person records before any editing or imputation for missing characteristics.

3. Matching Methodology

Following the terminology given by Winkler (1995), all the matching algorithms used in identifying duplicates in the census were exact matching because the goal was to link records for the same person. Most, but not all, matching algorithms used probabilistic methods. The matching that did not use probabilistic methods required complete agreement of specified fields although some fields may have a tolerance, such as a range for the year of birth.

Exact matching algorithms that use probabilistic methods typically have three components: (1) criteria to identify pairs for consideration, (2) a method for measuring closeness of two records, and (3) a decision rule to classify records as matches. The criteria for identifying pairs for consideration consist of individual fields that must agree exactly, often called blocking variables. Usually several passes through the data with variations in the blocking variables avoids missing matching records because of differences in the recording of an individual's information in the two data files. The measurement of closeness of two records involves scoring the amount of agreement between matching fields, usually with a weight for each field that signifies its relative importance, and then developing a composite score. Methods for developing the composite score have been suggested by Copas and Hilton (1990), Fellegi and Sunter (1969), Newcombe (1988), and Newcombe, Kennedy, Axford, and James (1959). The decision rule involves specifying a model that uses the composite agreement score to identify a match. The goal of these models is to identify as many correct matches as possible while keeping the number of false matches very low. Methodologies classifying records as a match have been developed by Belin and Rubin (1995), Newcombe (1988), Rogot, Sorlie, and Johnson (1986), Fellegi and Sunter (1969), and Tepping (1968). Winkler (1995) discusses the advantages and disadvantages of the different methods of scoring and decision rules.

Both the duplicate estimation and the evaluation used probabilistic and complete-agreement matching in the course of their processing. The probabilistic matching algorithms used by the duplicate estimation and the evaluation to link two files are based on the Fellegi-Sunter methodology. The Statistical Research Division Matcher software developed at the U.S. Census Bureau allows each record in one file to link to one and only one record in the other file (Winkler 1999). A newer algorithm developed at the U.S. Census Bureau, known as BigMatch (Yancey 2004), handles matching very large files and has the added benefit of permitting a record in one file to link to more than one record in the other file. The evaluation used a commercial software package AutoMatch offered by MatchWare, Inc. that has had several subsequent owners and has been renamed QualityStage (IBM 2006). All three probabilistic matching software programs allow users to specify blocking variables, matching variables, weights for the matching variables, and models to use in decision rules for which linked records to designate as matches.

3.1. Linking to Estimate Census Duplicates

This section presents an overview of the two types of record linkage that produced the estimates of census duplication used for A.C.E. Revision II. Detailed descriptions may be found in U.S. Census Bureau (2004) and Fay (2002, 2003). These methods of linking used only census information collected in Census 2000. The *household-based matching* used probabilistic matching to identify duplicates when more than one person in a household was duplicated. The household-based matching was able to use the information about how likely other members of the household were duplicated in the decision on whether to classify a link as a duplicate. When only one person in a household was duplicated in another household or in a group quarters residence (called single links), the *purely person-based matching* required complete agreement on the matching variables, but had a two-year tolerance for year of birth. The purely person-based matching was used when information about other household members, if there were any, was not helpful in determining how likely a link was a duplicate. Both matching methodologies developed probabilities that the links they identified were the same person. A final step integrated the results of the two methods to produce a probability that each linked E-sample record was duplicated in the census. The A.C.E. Revision II estimation incorporated the probability of being duplicated in its estimation of whether an E-sample case was enumerated in the correct location. The details of the estimation of the probability of being enumerated in the correct location are not discussed in this article but may be found in a report published by the U.S. Census Bureau (2004).

3.1.1. Household-Based Matching to Identify Duplicates

The household-based matching methodology consisted of two stages (U.S. Census Bureau 2004). The first stage was a national match of persons using the matching software BigMatch. The first stage used several blocking criteria composed of combinations of names, first letter of names, month and day of birth, and 10-year age group. The matching variables used to link records in the E-sample with records in the census were first name, last name, middle initial, month of birth, day of birth, and computed age. Matching parameters measured the degree to which each matching variable agreed between the two records, ranging from full agreement to full disagreement. The composite match score for the linked records was the sum of the weighted matching variable scores. Full agreement of at least four matching variables was required to consider two records a duplicate link. Such a conservative approach was necessary for a computerized search of the entire country to minimize linking records having similar characteristics but which were for different people.

At the first stage of matching it was possible for one sample case to link to multiple census records. All of these links were retained for the second stage of matching. This capability allowed for the possibility of more than two enumerations of the same person in the census. If an E-sample case linked to a census record in a group quarters, the case did not go to the second stage but was included with the single links.

The second stage used the Statistical Research Division Matcher software with more discriminating matching parameters. The second stage of matching was limited to matching persons within households so the household was the only blocking criterion. The second-stage matching variables were the same as the first stage; however, the matching parameters differed. The second-stage matching parameters were derived using the Expectation-Maximization (EM) algorithm as described by Winkler (1995) with a subset of the first-stage links. For the derivation, the EM algorithm was run twice, once with first-stage links from the largest local census office in New York, which is high-density urban with a diverse population, and again with first-stage links from a local census office in Wyoming, which is low-density rural with a fairly homogeneous population. Since the estimates of second-stage parameters were very similar for the extremes represented by New York and Wyoming, the simplifying assumption of using a common set of the second stage parameters for the whole U.S. seemed reasonable and meant a substantial reduction in the complexity and amount of time for the processing. A key difference between the first- and second-stage parameters was that there was considerably less emphasis on needing the last name to agree in the second stage since this matching was within a household.

The decision rule for designating a match in the second stage applied when two or more members of a sample household linked with members of a census household. After the second-stage matching, each link between a person in a sample household and a person in a census household had an overall match score. So, for each sample household, a set of match scores was observed. The likeliness of observing any resulting set of match scores was estimated by first estimating the probability of not observing this set of match scores. The higher the probability of not observing this set of match scores, the more likely the linked records are to be duplicates.

The estimate of the probability of not observing this set of match scores assumed independence of the individual match scores within each household. This assumption was consistent with using the EM algorithm to determine the second-stage matching parameters. The probability of observing the individual match scores was estimated from the empirical distribution of individual match scores resulting from the second stage matching. Further, this measure accounted for the number of times that a unique sample household was matched to different census households within a given level of geography.

The decision rule for designating matches used critical values that varied by level of geography for the probability of not observing a set of match scores. The critical values were larger for larger geographic distances between the linked records. When the probability was greater than the critical value for the geographic distance, the link was designated a match and assigned a match probability equal to one. When the probability was smaller, the link was not designated a match and received a match probability equal to zero. The geographical levels for the distance between the linked records were block, tract, same county (outside tract), same state (outside county), and different state.

3.1.2. Purely Person-Based Matching to Identify Duplicates

The purely person-based matching between the E-sample and the census to search for single links used software specially designed for this purpose. The requirements for linking were that the records agree exactly on first name, last name, month and day of birth, and two-year age intervals. Each linked pair of census records was assigned a probability between zero and one that they were the same person. The methodology for estimating the probability of being the same person took into account the overall distribution of births, frequency of names and population size in a specific geographic area. This approach took into account the possibility of coincidental agreement of names and birth dates. Duplicate probabilities were computed separately by geographic distance between the linked records, such as (1) in the same county, (2) in different counties within state, and (3) in different states. Further, the probabilities of being a duplicate were modeled separately by how common the last name was as well as separately for Hispanic names. Fay (2002, 2003) gives details about the methodology for estimating the probability of coincidental agreement of names and birth dates and for using that estimate in estimating the probability that the linked records are duplicate enumerations.

3.1.3. Merging Results from the Two Matching Methods

The results of the household-based matching and the purely person-based matching were merged to give one duplicate probability to each E-sample record with a duplicate link (U.S. Census Bureau 2004). The basic principle in assigning the final duplicate probability was to take the larger of the two estimates of the probability of being a duplicate, although the probability for some cases received an adjustment to account for the merging of the results from the two matching methods. The motivation for taking the higher probability was to leverage the individual strengths of the two methods. The household-based matching incorporated information about other household members and could identify duplicates that the purely person-based matching would miss or assign a lower probability of being a duplicate. In contrast, the purely person-based matching method was able to find single links that the household-based matching would not detect.

For the matches identified by the household-based matching, the larger of the two probabilities was always one. For the links to group quarters in the first stage of the household-based matching and single links, the probability of being a duplicate was equal to the estimate from the purely person-based matching modeling. For links in multiple-person households found by purely person-based matching but not found by the household-based matching, the nonzero duplicate probabilities from the purely person-based matching were adjusted downward since the estimated probabilities were conditional on the observed number of linked pairs that were possible duplicates. The aggregate decrease in the probabilities equaled the aggregate increase in the probabilities for links that received the household-based matching probability of one instead of the purely person-based matching probability. Without the downward adjustment, the estimated total number of duplicates from the purely person-based matching would have exceeded the conditional expected number of duplicates among the observed links that accounted for coincidental agreement of name and birthday.

3.2. Linking with Administrative Records Information for Evaluation

This section presents an overview of the Statistical Administrative Records System (StARS) and how it was used to evaluate the estimates of census duplication. The methodology for StARS was created for a test of the feasibility of an administrative record census as part of the research, evaluation, and experimentation program associated with Census 2000. The National Research Council suggested research on the feasibility of an administrative records census be done in conjunction with Census 2000 (Steffey and Bradburn 1994; Edmonston and Schultze 1995). Earlier Alvey and Scheuren (1982) described methodology for taking the census using records from the Internal Revenue Service supplemented by those of several other agencies. Scheuren (1999) later updated the discussion of methodology for census taking using administrative records.

The experiment in creating a census-like database with administrative records was the first of its kind at the U.S. Census Bureau. Such research appeared worthwhile because being able to construct an administrative records database with reasonable coverage of the population could aid research and evaluation and possibly support programs even if confidentiality constraints prevented an administrative records census in the future. For example, Zanutto and Zaslavsky (1996, 1997, 2001) suggested using administrative data in developing methodology for the imputation for nonresponding households. The new StARS methodology made it possible to evaluate the estimates of census duplication using administrative records.

3.2.1. Development of the Administrative Records Database

The focus of the assessment of the feasibility of an administrative records census was on examining the quality of a list compiling both housing and population information in administrative records from several agencies (Leggieri, Pistiner, and Farber 2002). The goal was to have a census-like record for each person, which meant including the data collected on the census short form, namely first name, middle initial, last name, birth date, race, and Hispanic origin. The study created the StARS1999 database for the entire U.S.A. with administrative records from 1999, and evaluated its accuracy in two sites, one composed of two counties in Maryland and the other composed of three counties in Colorado. The data were of a somewhat earlier time frame than Census Day of April 1, 2000. This was because of the large amount of time necessary to complete the acquisition of files from the contributing agencies and subsequent processing to unduplicate the records and geographically code the addresses to blocks.

StARS1999 contained 257 million people after processing to remove deceased persons and duplicates across and within the files. The final file also had 147 million addresses, of which 73 percent could be geographically coded to a block (Bye and Judson 2004).

In the two test sites, the research explored two methods for producing counts with StARS1999 that were comparable to Census 2000. One method simulated supplementing StARS1999 with fieldwork by using some data from Census 2000 while the other method simulated not using any data from fieldwork or other sources. The method with fieldwork produced a count for the two sites that was within 1 percent of the Census 2000 count. The second method that used only StARS1999 produced a count within 8 percent of the Census 2000 count for the two sites (Bye and Judson 2004). For a critique of the research on an administrative records census, see a report by the National Research Council (Cork, Cohen, and King 2004). The biggest weakness of StARS1999 was the lack of race and Hispanic origin data for many records.[2] Since many of the files did not contain this

[2] The race and Hispanic origin data have since been improved in later versions of StARS (Farber and Miller 2003).

information, models were employed to assign race and Hispanic origin to records where these fields were blank (Bye and Judson 2004).

Since the experiment showed the method of creating the database had potential, the StARS database was created again with administrative records collected in 2000 with processing refinements. StARS2000 contained 266 million records for people (Judson 2002), and therefore was closer to the Census 2000 count than StARS1999. Linking the StARS2000 to Census 2000 had two purposes. One was to improve the processing of the administrative files so they could be processed as received instead of waiting for all to arrive. The other purpose was to develop models for estimating race and Hispanic origin since most of the administrative records do not contain these characteristics. Having the time frame for the database be as close as possible to the time frame for the census facilitated producing the model. The evaluation of the estimates of duplicate enumerations in the census used the results of the linking between StARS2000 and Census 2000.

Both StARS1999 and StARS2000 databases incorporated data from seven administrative record files:[3] (1) Internal Revenue Service Individual Master File (1040), (2) IRS Information Returns File (W-2/1099), (3) Department of Housing and Urban Development Tenant Rental Assistance Certification System File, (4) Department of Housing and Urban Development's Multifamily Tenant Characteristics System File, (5) Center for Medicare and Medicaid Services Medicare Enrollment Database File, (6) Indian Health Services Patient Registration System File, and (7) Selective Service System Registration File. Many people had a record in more than one of the record systems. Since each system contained Social Security Numbers, a simple match on these numbers identified duplicate entries for the same person. When a person had more than one record, an algorithm chose the "best" record for a person, usually the most recent entry, although all the addresses for a person were retained. A detailed description of the processing and selection of the "best" record may be found in Berning and Cook (2003). The research on the feasibility of an administrative records census used the "best" address, but the evaluation of census duplication used all the addresses that appeared in any of the sources for a person.

In addition, the U.S. Census Bureau merged address data from tax forms with the Social Security Administration's Numerical Identification File (Numident).[4] For confidentiality, the Numident was edited, and a unique identification number (ID number) was created for each Social Security Number. Then an address variable was added for each address from the IRS 1040 and 1099 files from StARS2000 for each person. Also, the ID numbers were used in all versions of the StARS database instead of Social Security Numbers to protect privacy.

There are some limitations in the use of StARS2000 in the identification of duplicates in Census 2000 (Berning and Cook 2003). The time frame for Census 2000 and that for the administrative files were not exactly the same. The reference date for Census 2000 was Census Day, April 1, 2000. The contents of the seven administrative records files provided

[3] The Census Bureau obtains administrative data for its StARS database as authorized by Title 13 U.S.C., Section 6 and supported by provisions of the Privacy Act of 1974. Under Title 13, the U.S. Census Bureau is required to protect the confidentiality of all the information it receives directly from respondents or indirectly from administrative agencies and is permitted only to use that information for statistical purposes.
[4] Ibid

for StARS2000 corresponded as much as possible to the same reference date, April 1, 2000. However, the filing deadline for IRS 1040 forms, the major source of addresses, was April 15, 2000, and these may be filed as early as January or, with extensions, as late as August. A person may submit a 1040 early in the year and move to a different address by Census Day. The new address may not be in one of the other administrative systems at the time its files were sent for the creation of StARS2000.

Another limitation is that some people have two Social Security Numbers, and more than one person can have the same Social Security Number. If a person in sample has two Social Security Numbers, then the person also will be assigned two ID numbers in the Census Numident and the evaluation may fail to find that person's duplicate in the census. If more than one person has the Social Security Number of a person in sample, then the same ID number will be assigned to two people and the evaluation may falsely call them duplicates (Bean and Bauder 2002).

3.2.2. Administrative-Records-Based Matching

The *administrative-records-based matching* methodology used information from administrative records and had two basic steps. First, probabilistic exact matching assigned ID numbers to census records, and thereby the subset of census records in the E-sample, by matching census files to administrative records in the StARS2000 database. A second matching compared the ID numbers of records. When two records had the same ID number, they were designated a duplicate link. Both steps used the commercial matching software AutoMatch (now called QualityStage).

The matching for the first step between the census file to Numident was done in two phases, each of which used several sets of blocking criteria (Wagner 2002). For the first phase, the blocking variables ranged from nearly the full address for the early passes, then broadened the geography by blocking on parts of the address along with first characters of names and partial to full date of birth. The matching variables included name, date of birth, sex and parts of the address.

In the first phase, person data and address data were used as blocking variables. Census records that were not linked to Numident records using the address variables in blocking were then involved in the second phase of matching where the blocking fields involved only person data. The second phase blocking variables included combinations of names, first characters of names, and partial to full date of birth. The matching variables included name, date of birth, and sex (Farber and Miller 2003).

Via this match, ID numbers were found for census people and added to census person records, creating a research file. Note that some person records on the census file had no ID number assigned. This could happen in two ways. If the census record was not linked with any ID number, none could be assigned. In addition, when one census record was linked with more than one ID number (which is likely to have occurred when linking people with common names and characteristics), no ID number was assigned to the census record. In many cases, the process avoided linking different people whose person characteristics were similar. However, there was no assessment of how many false links remained, or how many true links were missed.

The research file composed of census records with ID numbers provided the basis to identify duplicates by matching on ID number. Links were created between E-sample records and census records with the same ID number. Each link with an ID number was considered a duplicate with probability equal to one.

4. Results

Table 1 shows the estimates of census duplication by geographic distance between the linked records and type of census record while Table 2 shows the evaluation estimates developed using administrative records information. The weights used in the estimation are the product of the A.C.E. sampling weight, the multiplicity factor that prevents double-counting (Mule 2002), and the probability of duplication. For the estimates of duplication in Table 1, the probabilities of duplication varied between zero and one. For the estimates from evaluation with administrative records in Table 2, all the links had a probability of duplication equal to one. Standard errors were calculated using a simple jackknife method. The categories for geographic distance are: (1) within the A.C.E. block cluster, (2) outside of the A.C.E. block cluster, but within surrounding blocks, (3) outside of surrounding blocks, but within same county, (4) outside of surrounding blocks and county, but within same state, and (5) outside of surrounding blocks, in a different state. The types of census record are: (1) enumerations eligible for selection in the E-Sample because they are in housing units not identified as potential duplicates during the census, (2) enumerations in group quarters, (3) enumerations in housing units reinstated in the census but had been considered potential housing unit duplicates, and (4) enumerations in housing units deleted from the census because they were housing unit duplicates. In the first three categories, both of the linked records were in the census, but one should not have been. In the last category, the link was to a record deleted from the census although the correct record may or may not have been the one deleted.

The estimate of the number of duplicates in the census based on the merged household-based matching and purely person-based matching was approximately 5.8 million (sum of E-sample eligible, group quarters, and reinstate columns), which is approximately 0.9 million fewer than the 6.7 million estimated by the evaluation with administrative records. The evaluation with administrative records estimated more duplicates between records eligible for the E-sample (4.2 million versus 3.5 million). The evaluation found fewer duplicates between records considered as possible housing unit duplicates but reinstated in the census (1.5 million versus 1.7 million) and records deleted as housing unit duplicates (2.5 million versus 2.9 million). Within block clusters, the estimate of 1.2 million duplicates between records within households eligible for the E-sample was higher than the evaluation estimate at just under a million. However, neither estimate based on computerized linking was as effective as the A.C.E. clerical person matching since the results of their matching estimated about 1.9 million duplicates in this category (Mule 2002).

Two other differences stood out between the estimates of census duplication and the evaluation estimates. The evaluation with administrative records estimated more duplicates to group quarters and to census records in different states. One possible reason for the higher estimate of duplicates to group quarters from the administrative-records-based matching may be that the purely person-based matching assigned these links their duplicate probability, which usually was less than one, while the administrative-records-based

Table 1. Estimates of Census Duplicates Developed Using Only Census Information by Geographic Distance and Census Record Type

Geographic distance	Census record type								Total	
	E-Sample eligible		Group quarters		Reinstate		Delete			
Within cluster	1,173,344	(46,173)	76,381	(15,736)	1,058,548	(48,295)	1,967,199	(94,454)	4,275,472	(129,245)
Surrounding block	259,805	(21,718)	25,373	(9,701)	24,751	(6,971)	678,355	(57,469)	988,284	(65,896)
Same county	1,011,920	(24,292)	231,774	(39,795)	482,015	(27,797)	208,246	(20,789)	1,933,956	(59,590)
Diff. county, same state	563,270	(18,873)	190,417	(9,488)	88,331	(12,567)	35,111	(7,262)	877,129	(26,615)
Different state	527,796	(23,744)	91,793	(7,093)	20,959	(17,316)*	16,184	(4,902)	656,732	(33,930)
Total	3,536,136	(68,045)	615,738	(46,003)	1,674,604	(60,317)	2,905,096	(116,541)	8,731,572	(177,071)

Notes: The sum of the columns with headings E-sample eligible, group quarters, and reinstate is approximately 5.8 million.
Standard errors are in parentheses. *The large standard error is due to clustering. (Mule 2002)

Table 2. Evaluation Estimates of Census Duplicates Developed Using Administrative Records by Geographic Distance and Census Record Type

Geographic distance	Census record type								Total	
	E-Sample eligible		Group quarters		Reinstate		Delete			
Within cluster	998,239	(35,162)	107,305	(21,452)	920,405	(42,888)	1,681,962	(82,499)	3,707,911	(113,548)
Surrounding block	202,741	(15,516)	31,355	(11,686)	22,870	(5,926)	588,300	(48,878)	845,266	(55,656)
Same county	1,145,036	(24,177)	334,983	(47,946)	420,917	(24,624)	187,804	(18,520)	2,088,740	(64,559)
Diff. county, same state	693,540	(20,531)	307,014	(13,610)	79,986	(10,708)	35,618	(6,734)	1,116,159	(29,646)
Different state	1,183,055	(30,328)	183,917	(10,500)	21,808	(3,276)	32,472	(4,350)	1,421,251	(34,133)
Total	4,222,611	(68,660)	964,574	(57,701)	1,465,986	(52,042)	2,526,156	(102,200)	9,179,326	(169,735)

Note: The sum of the columns with headings E-sample eligible, group quarters, and reinstate is approximately 6.65 million.
Standard errors are in parentheses.

matching assigned every duplicate a probability of one. Another possibility is that the administrative-records-based matching was probabilistic and therefore may have been less strict in identifying links than the purely person-based matching, which required complete agreement on matching variables with a two-year age range.

In an attempt to explain some of the differences between estimates of duplicates and the evaluation estimates within and between states, an analysis examined the links by household composition (i.e., the number of links between the households relative to the size of the household in the E-sample). Table 3 shows the distribution by whether the link was to the same state (the first four categories of geographic distance in Tables 1 and 2) or to a different state, since the latter category is where the evaluation tended to find more duplication.

Approximately 51 percent of the links to different states found by the evaluation using administrative-records-based matching were not found by the linking that used only census information, compared with 15.4 percent of the evaluation links to the other geographic distances. In multiple-person households, the percentage of links found only by the administrative-records-based matching was smaller when at least two people had links than when only one person had a link. The percentage was even lower when all household members had a link. In other words, there was more overlap when there were multiple links in a household. This general trend held both for links to different states and for links within a state. However, more multiple links were found only by the evaluation when the links were between different states (39.2 percent versus 10.3 percent and 8.7 percent versus 2.8 percent).

In the evaluation, the links in housing units with more than two people but only one link (called single links) comprised a larger proportion of the links between states than the links within a state (60.1 percent (853,828/1,41,241) vs 23.7 percent (1,837,816/7,758,075)). Furthermore, 64.7 percent of the single links to different states were found only by the evaluation while the percentage dropped to 34.1 percent of the single links within the same state. The single links were a major source of the difference between the estimates of duplication and the evaluation estimates between different states.

Since many of the links to different states found for estimation of duplication were single links, many of these links were assigned final duplication probabilities in the purely person-based matching. Due to the large geographic distance, many of these links may have been assigned probabilities less than one. However, in administrative-records-based matching all links were treated as duplicates (as if they all have a final duplicate probability of one). So even if there had been a lot of overlap between the links to different states found by the two matching processes, the estimates of duplication between states based on only census information could have been substantially lower than the evaluation estimates based on administrative records information with the same links.

Recall that the matching process assigned ID numbers in two phases: one using both address and person information and a second phase using person information only. Links formed with address information may create more confidence because this linking required similar address data as well as person data. Thus, Tables 4 and 5 show the links identified using administrative records by whether the ID numbers were assigned using address information.

Table 3. Evaluation Estimates of Census Duplicates Identified Only Using Administrative Records(Ad Rec) by Household Composition and Geographic Distance

Household composition		Geographic distance between records			
HH size	HH duplication status	Same state		Different state	
		% of all ad rec links that are ad rec only	All ad rec links	% of all ad rec links that are ad rec only	All ad rec links
1	All	36.0 (1.1)	727,889 (23,908)	54.6 (2.3)	132,379 (7,296)
2 +	All	2.8 (0.3)	3,052,411 (100,883)	8.7 (1.1)	232,581 (18,014)
	Partial – 2 + links	10.3 (0.5)	2,139,959 (64,818)	39.2 (2.3)	202,463 (11,155)
	Partial – single link	34.1 (0.7)	1,837,816 (35,799)	64.7 (1.0)	853,828 (19,821)
	Total	13.3 (0.3)	7,030,186 (151,162)	50.6 (1.0)	1,288,872 (31,980)
Total		15.4 (0.3)	7,758,075 (159,182)	51.0 (1.0)	1,421,251 (34,133)

Note: Standard errors are in parentheses.

Table 4. Evaluation Estimates of Duplicates Identified Using Administrative Records (Ad Rec) by Whether Address was Used in Assigning ID Numbers. Links Within a State Only

ID numbers assigned using address	Type of evaluation link					
	Ad rec links	Ad rec only links	% of ad rec links that are ad rec only	% of total ad rec only links	Ad rec links also found for duplicate estimates	% of total ad rec links also found for duplicate estimates
Both	5,815,854 (134,973)	805,416 (23,770)	13.8 (0.4)	67.4 (0.9)	5,010,438 (125,535)	76.4 (0.6)
One	1,369,758 (35,636)	318,126 (12,317)	23.2 (0.8)	26.6 (0.9)	1,051,632 (32,338)	16.0 (0.5)
None	572,463 (25,383)	72,240 (5,502)	12.6 (0.9)	6.0 (0.4)	500,224 (23,521)	7.6 (0.3)
Total	7,758,075 (159,182)	1,195,782 (29,173)	15.4 (0.4)	100.0	6,562,294 (146,443)	100.0

Note: Standard errors are in parentheses.

Table 5. Evaluation Estimates of Duplicates Identified with Administrative Records (Ad Rec) by Whether Address Was Used in Assigning ID Numbers. Links Between Different States Only

ID numbers assigned using address	Type of evaluation link					
	Ad rec links	Ad rec only links	% of ad rec links that are ad rec only	% of total ad rec only links	Ad rec links also found for duplicate estimates	% of total ad rec links also found for duplicate estimates
Both	199,937 (10,740)	58,436 (4,499)	29.2 (1.9)	8.1 (0.6)	141,502 (9,055)	20.3 (1.0)
One	1,092,517 (27,185)	586,635 (15,415)	53.7 (1.1)	81.0 (0.8)	505,882 (19,935)	72.6 (1.1)
None	128,797 (6,693)	79,290 (4,948)	61.6 (2.5)	11.0 (0.6)	49,506 (4,275)	7.1 (0.6)
Total	1,421,251 (34,133)	724,362 (17,831)	51.0 (1.0)	100.0	696,889 (25,540)	100.0

Note: Standard errors are in parentheses.

The categories for the links by the two phases are: (1) both of the linked records had the ID number assigned using address information, (2) only one of the linked records had the ID number assigned using address information, and (3) neither of the linked records had the ID number assigned using address information.

When viewing links within and between states separately, the use of address information in addition to person information in the administrative-records-based matching tended to increase the percentage of links also found for the estimates of duplication, but more so for links between different states. When both links were in the same state and were found using both address and person information, 13.8 percent of the links were not found by the matching for duplication estimation, compared to about 20 percent ((318,126 + 72,240)/(1,369,758 + 572,463)) for the other administrative records links. However, for links to different states, 29.2 percent of links where both records were assigned ID numbers using both address and person information were not found by the matching for duplication estimation, compared to about 55 percent ((586,635 + 79,290)/(1,092,517 + 128,797)) for the other links using administrative records information.

However, the percent of total columns (fifth and seventh columns) of Tables 4 and 5 show that when the links are tabulated by the method of assigning ID numbers separately within state and between state, similar patterns appear for those found only by the evaluation and those in the overlap found by both the administrative-records-based matching and the matching using only census information. For links within states, 23.6 percent of those in the overlap did not have the ID numbers assigned using address information for both members of the pair, while 32.6 percent of those found only by the evaluation did not have address information used for both. For links between states, 79.1 percent of those in the overlap did not have address information used in assigning ID numbers to both members of the pair, while 92.0 percent of those found only by the evaluation did not have address information used for both. So, problems among links from only administrative-records-based matching where both ID numbers were not assigned using address and person information may be present in such links found by both administrative-records-based matching and matching with only census information. More likely, the links in the overlap have a strength that comes from being discovered by two separate methodologies of identifying duplicates, even though both approaches are automated.

5. Clerical Review of Links for Duplicate Estimation and Evaluation

The U.S. Census Bureau's elite matching team reviewed a subsample of the links to enumerations outside the blocks surrounding the sample blocks from the household-based matching and purely person-based matching for the duplicate estimation and from the administrative-records-based matching for the evaluation (Byrne, Beaghen, and Mulry 2002, 2003). The elite matching team is a small group of permanent employees in the processing office who all have at least 10 years of experience with clerical matching. The team used only data collected in Census 2000 for the review, and saw none of the information from the administrative records. Each link received a designation of confirmed duplicate, denied as a duplicate, or undetermined. The team did not review duplicates to enumerations in group quarters because the analysts would not have information from other household members to use in making decisions.

The elite clerical matching team reviewed links that household-based matching declared duplicates and some of the links that were not considered a duplicate because their probability of being a match was below the critical value. The matching team agreed with 95 percent of the household-based matching E-sample duplicates in housing units eligible for the E-sample outside the search area, denied 4 percent, and found 1 percent undetermined. In fact, both the administrative-records-based matching and the elite matching team agreed with 73 percent of the duplicates in housing units eligible for the E-sample outside the search area. As for the enumerations linked but not declared duplicates, both the administrative-records-based matching and the elite matching team agreed with the decision that these cases were not duplicates 82 percent of the time. The elite matching team alone agreed that 94 percent were not duplicates, but found that 5 percent were duplicates, with 1 percent undetermined (Mulry and Petroni 2003).

For the duplicates found by the purely person-based matching that also were found by the administrative-records-based matching for the evaluation, the results for the elite matching team assume that the probability of being a duplicate equals one. Under this assumption, the elite matching team agreed with 80 percent of the purely person-based matching E-sample duplicates in housing units eligible for the E-sample found outside the search area, denied 8 percent, and found 12 percent undetermined. However, all the duplicates identified by the purely person-based matching received a probability of being a duplicate that is less than one in the A.C.E. Revision II estimator.

Links for an estimated 1.2 million of the duplicates found with administrative-records-based matching were not identified by the household-based or purely person-based matching for the duplication estimation. The matching team's review raises questions about the duplicates found only by administrative-records-based matching in the evaluation. The matching team agreed on 37 percent, disagreed on 47 percent, and was undecided on 15 percent. One could question whether the matching team is correct for these cases because the reason given for 65 percent of the disagreements was "household composition". These are among the most challenging cases to resolve without field work because detecting a person who is truly a member of two different households is difficult (Mulry and Petroni 2003). However, when the links were between enumerations in the same state, the matching team agreed with 62 percent, disagreed with 23 percent, and found 15 percent undetermined. In contrast, when the links were between enumerations in different states, the matching team agreed with only 10 percent, disagreed with 75 percent, and found 15 percent undetermined (Bauder 2004).

The matching team did provide some insight about the links between states from the results of their review of the subsample of all administrative-records-based matching links, including those also found by using only census information. For the administrative-records-based matching links between enumerations in the same state, the matching team confirmed 84 percent, denied 8 percent and found 8 percent undetermined. When the evaluation links were between enumerations in different states, the matching team confirmed 37 percent, denied 49 percent, and found 14 percent undetermined. However, for the subset where the address was used along with person information to assign the ID Number to both enumerations, the matching team confirmed 75 percent of the duplicates identified by the evaluation using administrative records (Bauder 2004). The use of the address in assigning the ID number appears to be an important indicator of whether the matching team confirms a link found by administrative-records-based matching but not found by the household-based and purely person-based matching methods.

The elite matching team did not review some categories of duplicates that were incorporated in the A.C.E. Revision II estimator. Links found by the purely person-based matching that accounted for an estimated 86,883 duplicates were not confirmed (i.e., they were denied or undetermined) by the administrative-records-based matching not submitted for review by the matching team. The elite matching team did not review the links for an estimated 297,164 duplicates to enumerations in housing units identified as potential duplicates, some of which were in housing units reinstated in the census and the rest were deleted. Also, an estimated 513,984 duplicates to enumerations in group quarters were not eligible for review by the elite matching team (Mulry and Petroni 2003).

6. Summary

The evaluation based on administrative-records-based matching provided insight about the identification of duplicates in Census 2000 as well as insight about administrative-records-based matching. The evaluation validated the estimate of a large number of duplicates in Census 2000 detected by a combination of two matching methods using only census information. The administrative-records-based matching proved to be effective in the confirmation and denial of duplicate enumerations and in need of refinement for finding additional duplicates. When administrative-records-based matching also found a duplicate designated by the matching using only census information, the clerical team was more likely also to agree. When the administrative-records-based matching did not find a duplicate designated by the matching using only census information, the clerical team was more likely not to find a duplicate.

Administrative-records-based matching has the information required to identify duplicates that both the methods using only census data have difficulty detecting – for example, people enumerated with different names, and people whose enumerations have reporting errors.

Although an analysis of the evaluation results and a clerical review of a sample of the duplicates outside the surrounding blocks (Byrne, Beaghen, and Mulry 2003, 2002) presented reasons to question some of the links found only by administrative-records-based matching, there was evidence that the evaluation identified some duplicates missed by matching with only census data.

Analysis of all the administrative-records-based matching links identified a subset where the confidence was high that they were duplicates. The links where the ID numbers for both members of the pair were assigned during the first phase that used address information were confirmed at a much higher rate −75 percent for links between states and 84 percent for links within states.

Further research may very well improve the effectiveness of the second phase of assigning ID numbers that did not use address information. The evaluation used the results of an assignment of ID numbers to census records designed to associate Census 2000 race and Hispanic origin data with ID numbers. The matching strategy and decision rules were developed with the goal of linking as many census records as completely as possible with ID numbers, while maintaining a reasonably low false match rate over all the census records. In contrast, a procedure designed with the goal of identifying census duplicates might need to use stricter criteria in assigning an ID number to a census record when not using any address information.

Overall, the results indicate that a record linkage methodology using administrative records information developed with the sole purpose of detecting census duplicates and built on the lessons learned from this study, the clerical review, and further research, has the potential for producing more complete and accurate results.

7. References

Alvey, W. and Scheuren, F. (1982). Background for an Administrative Record Census. Proceedings of the American Statistical Association, Social Statistics Section, 137–146.

Bauder, M. (2004). Administrative Records and Person Duplication in Census. Proceedings of the American Statistical Association, Section on Government Statistics. [CD-ROM], 1430–1437.

Bean, S.L. (2001). ESCAP II: Accuracy and Coverage Evaluation Matching Error. Executive Steering Committee for A.C.E. Policy II, Report No. 7., October 12. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/Report7.pdf

Bean, S.L. and Bauder, D.M. (2002). Census and Administrative Records Duplication Study. DSSD A.C.E. Revision II Memorandum Series #PP-44. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/pp-44r.pdf

Belin, T. and Rubin, D. (1995). A Method for Calibrating False-Match Rates in Record Linkage. Journal of the American Statistical Association, 694–699.

Berning, M. and Cook, R. (2003). Administrative Records Experiment (AREX 2000) Process Evaluation. Census 2000 Evaluation. U.S. Census Bureau, Washington, DC. http://www.census.gov/pred/www/rpts/AREX2000_Process.pdf

Bye, B. and Judson D. (2004). Results from the Administrative Records Experiment in 2000. Census 2000 Testing, Experimentation, and Evaluation Program. Topic Report No. 16, TR-16. U.S. Census Bureau, Washington, DC. http://www.census.gov/pred/www/rpts/TR16.pdf

Byrne, R., Beaghen, M., and Mulry, M.H. (2002). Clerical Review of Census Duplicates. DSSD A.C.E. Revision II Memorandum Series #PP-43. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/pp-43r.pdf

Byrne, R., Beaghen, M., and Mulry, M.H. (2003). Clerical Review of Census Duplicates. Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], 768–773.

Copas, J. and Hilton, F. (1990). Record Linkage: Statistical Models for Matching Computer Records. Journal of the Royal Statistical Society, Series A, 153, 287–320.

Cork, D.L., Cohen, M.L., and King, B.F. (2004). Reengineering the 2010 Census: Risks and Challenges. National Academy Press. Washington, DC.

Edmonston, B. and Schultze, C. (1995). Modernizing the U.S. Census. National Academy Press. Washington, DC.

Farber, J. and Miller, E. (2003). Matching Census 2000 to Administrative Records. Proceedings of the American Statistical Association, Section on Government Statistics [CD-ROM], 1387–1390.

Fay, R.E. (2001). The 2000 Housing Unit Duplication Operations and Their Effect on the Accuracy of the Population Count. Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM].

Fay, R.E. (2002). Probabilistic Models for Detecting Census Person Duplication. Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], 969–974.

Fay, R.E. (2003). Probabilistic Models for Detecting Census Duplication at the Person and Household Levels. Proceedings of the American Statistical Association, Section on Government Statistics [CD-ROM], 1391–1398.

Fellegi, I. and Sunter, A. (1969). A Theory of Record Linkage. Journal of the American Statistical Association, 64, 1183–1210.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. Journal of the American Statistical Association, 88, 1047–1060.

IBM (2006). IBM Websphere QualityStage. Make Critical Business Decisions with Confidence Using Industry-Leading Data Quality Solution. ibm.ascential.com/-products/qualitystage.html#

Judson, D.H. (2000). The Statistical Administrative Records System: System Design and Challenges. Paper presented at the NISS/Telcordia Data Quality Conference, November.

Judson, D.H. (2002). The Statistical Administrative Records System and Administrative Records Experiment 2000: System Design, Successes, and Challenges. Paper presented at the Meeting of National Academy of Science Panel on Research on Future Census Methods for 2010.

Leggieri, C., Pistiner, A., and Farber, J. (2002). Methods for Conducting an Administrative Records Experiment in Census 2000. Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], 2709–2713.

Mule, T. (2002). Further Study of Person Duplication in Census 2000. DSSD A.C.E. Revision II Memorandum Series #PP-51. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/pp-51r.pdf

Mulry, M.H. and Petroni, R.J. (2003). Evaluation of the A.C.E. Revision II Estimates of Census 2000 Coverage Error. Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], 2960–2965.

Mulry, M.H., Bean, S.L., Bauder, D.M., Wagner, D., Mule, T., and Petroni, R.J. (2003). Census and Administrative Records Study. Proceedings of the Federal Committee on Statistical Methods Research Conference. U.S. Office of Management and Budget. Washington, DC, 36–43. http://www.fcsm.gov/03papers/MulryBean.pdf

Newcombe, H.B. (1988). Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business. Oxford University Press. Oxford, UK.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959). Automatic Linkage of Vital Records. Science, 130, 954–959.

Raglin, D.A. and Krejsa, E.A. (2001). ESCAP II: Evaluation Results for Changes in A.C.E. Enumeration Status. Executive Steering Committee for A.C.E. Policy II, Report

No. 16. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/Report3.pdf

Robinson, J.G. and Adlakha, A. (2002). Comparison of A.C.E. Revision II Results with Demographic Analysis. DSSD A.C.E. Revision II Memorandum Series #PP-41. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/pp-41r.pdf

Rogot, E., Sorlie, P.D., and Johnson, N.J. (1986). Probabilistic Methods for Matching Census Samples to the National Death Index. Journal of Chronic Diseases, 39, 719–724.

Scheuren, F. (1999). Administrative Records and Census Taking. Survey Methodology, 25, 151–160.

Smith, D.R. (2004). Long Distance Duplicate Telephone Followup Operation. DSSD 2003 Memorandum Series Chapter #A-03. July 8, 2004. U.S. Census Bureau, Washington, DC.

Steffey, D.L. and Bradburn, N.M. (1994). Counting People in the Information Age. National Academy Press: Washington, DC.

Tepping, B.J. (1968). A Model for Optimal Linkage of Records. Journal of the American Statistical Association, 63, 1321–1332.

Thompson, J., Waite, P., and Fay, R., (2001). Basis of 'Revised Early Approximations' of Undercounts Released Oct. 17, 2001. Executive Steering Committee for A.C.E. Policy II, Report 9a. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/report9a.pdf

U.S. Census Bureau (2001a). Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses. October 17. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/Recommend2.pdf

U.S. Census Bureau (2001b). Analysis Plan for Further ESCAP Deliberations Regarding the Adjustment of Census 2000 Data for Future Uses. July 26. U.S. Census Bureau. Washington, DC.

U.S. Census Bureau (2001c). Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy. March 1. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/ESCAP2.pdf

U.S. Census Bureau (2003a). 2004 Census Test Operational Plan. Dated September 29, 2003. Decennial Management Division. U.S. Census Bureau, Washington, DC. http://www.census.gov/procur/www/fdca/library/2004-test/2004-Operational-Plan-v2-9-29-03.pdf

U.S. Census Bureau (2003b). Decision on Intercensal Population Estimates. March 12. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/dipe.html

U.S. Census Bureau (2003c). Technical Assessment of A.C.E. Revision II, March 12. U.S. Census Bureau, Washington, DC. http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf

U.S. Census Bureau (2004). Accuracy and Coverage Evaluation of Census 2000: Design and Methodology. DSSD/03-DM. Issued September 2004. U.S. Census Bureau, Washington, DC. http://www.census.gov/prod/2004pubs/dssd03-dm.pdf

Wagner, D. (2002). Race Enhanced Numident Project – Match HCUF to Census Flow Charts. Dated May. U.S. Census Bureau. Washington, DC.

West, K., Mulry, M., Parmer, R., and Petrik, J. (1991). Address Reporting Error in the 1990 Post-Enumeration Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, 236–241.

Winkler, W.E. (1995). Matching and Record Linkage. In Cox et al. (eds) Business Survey Methods, John Wiley and Sons, NY: 355–384.

Winkler, W.E. (1999). Documentation for Record Linkage Software. Statistical Research Division. U.S. Census Bureau. Washington, DC.

Yancey, W.E. (2004). Bigmatch: A Program for Large-Scale Record Linkage. Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM].

Zanutto, E. and Zaslavsky, A.M. (1996). Modeling Census Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup to Impute Census Nonrespondents. Proceedings of the American Statistical Association, Section on Survey Research Methods, 538–543.

Zanutto, E. and Zaslavsky, A.M. (1997). Estimating a Population Roster from an Incomplete Census Using Mailback Questionnaires, Administrative Records, and Sampled Nonresponse Followup. Proceedings of the American Statistical Association, Section on Survey Research Methods, 754–759.

Zanutto, E. and Zaslavsky, A.M. (2001). Using Administrative Records to Impute for Nonresponse. In Groves, R., Dillman, D., Eltinge, J., and Little, R.J.A. (eds). Survey Nonresponse, NY: John Wiley and Sons, 403–416.

Received January 2004
Revised August 2006