Enumeration Status of Census 2000 Enumerations Deemed Insufficient Information for Matching and Followup

Paul Livermore Auer
U.S. Census Bureau
Washington, DC 20233-7600

ABSTRACT

The Accuracy and Coverage Evaluation (A.C.E.) was the Census Bureau's program for measuring coverage in Census 2000. The A.C.E. included an independent enumeration in a sample of block clusters referred to as the population sample, or P Sample. The A.C.E. matched the P Sample to census enumerations in order to estimate the coverage of Census 2000. Some census enumerations did not meet the A.C.E. criterion for sufficient information for matching and followup. These cases were effectively treated as erroneous enumerations in the A.C.E. processing, but did not affect estimates of net coverage error due to a corresponding balance with omissions in the dual system estimator. Since future coverage measurement programs will focus on estimating components of census coverage error, an alternative treatment of these cases is necessary (Singh 2003). This paper discusses a study designed to clerically match cases deemed insufficient information for matching and followup to the P Sample. Results from this study will be used to determine strategies for estimating components of coverage error in future censuses[1].

INTRODUCTION

A major goal and challenge for coverage measurement in 2010 is to design a survey that measures the components of coverage error, namely erroneous enumerations and omissions. Previous coverage measurement surveys, the 2000 Accuracy and Coverage Evaluation (A.C.E.) and the 1990 Post Enumeration Survey (PES), were designed primarily to estimate net census error using Dual System Estimation (DSE). To improve the accuracy of estimates of net error, our implementation of the DSE has relied on balancing some of the components of error, meaning some census omissions offset some erroneous inclusions in a manner that preserved the net error. Essentially this has entailed using a very strict definition for measuring correct enumerations. This has resulted in inflated estimates of omissions and erroneous inclusions. In order to produce more accurate estimates of erroneous enumerations and omissions, it is necessary to expand the definition of correct enumeration. A necessary condition to being considered a correct enumeration was that the enumeration had to have a complete name and at least two characteristics. Enumerations lacking a complete name and two characteristics were called insufficient information for matching and followup. The A.C.E. removed from its match processing almost 4.8 million weighted data-defined census records deemed insufficient information for matching and followup (Feldpausch 2002). They were removed from matching in order to avoid incorrect matching and/or incorrect determination of enumeration status which would have led to biases in the DSE.

Since estimation of coverage error components is a main focus of coverage measurement operations in 2010, it is necessary to estimate the number of cases deemed insufficient information for matching and followup that are correctly or erroneously enumerated. Those that remain unresolved will have to be dealt with using appropriate missing data procedures. This study attempts to match Census 2000 records deemed insufficient information for matching and followup to the P Sample using liberal matching rules. This is a first step in arriving at a more accurate estimate of erroneous enumerations. Since in this study we are matching the census to the P Sample we will not be considering the entire P Sample and will therefore not arrive at an estimate of omissions. For the 2006 Census Test, enumerations with insufficient information for matching and followup will be treated as they were in A.C.E. for net error calculations, but they will be used for estimating component errors.

BACKGROUND

The A.C.E. was comprised of two samples, a population or P Sample to measure census omissions and an enumeration or E Sample to measure census erroneous enumerations. The P Sample was obtained by independently listing housing units in a sample of

[1] This report is released to inform interested parties of ongoing research and to encourage discuss-ion of work in progress. The views expressed on statistical, methodological, technical, and operational issues are those of the author and not necessarily of the U.S. Census Bureau.

block clusters and conducting person interviews at these housing units. The P Sample included nonmovers and outmovers, but not inmovers. The E Sample consisted of selected census enumerations in those sample blocks. The P sample was matched to the census listing. P-Sample people not matching the census listing were identified as census omissions, though certain P-Sample nonmatches had an additional interview, the A.C.E. Person Followup (PFU), to confirm their Census Day residence status (Childers 2001). E-Sample enumerations that matched to P-Sample people were counted as correct enumerations. Nonmatched E-Sample people were followed up in the A.C.E. PFU interview to determine their enumeration status.

In the A.C.E. processing, cases with insufficient information for matching and followup received a match code of "KE." For the sake of brevity, I will refer to these cases as simply "KEs." Census whole person imputations have historically been referred to as "insufficient information." These imputations are to be distinguished from the E-Sample enumerations coded KE.

Recall that KEs were removed from matching in order to avoid incorrect matching and/or incorrect determination of enumeration status which would have led to biases in the DSE. The method in which the A.C.E. removed the KEs from its processing required that the A.C.E. estimation treat them as erroneous enumerations. Treating enumerations with insufficient information for matching and followup as erroneous does not imply that A.C.E. believed the enumerations were erroneous.

Presumably, many of the KEs corresponded to people who lived at the address on Census Day and were not otherwise counted, in other words, they could be considered correctly enumerated for estimation of census coverage error components. Logically, these correct enumerations corresponded to P-sample nonmatches. Counting correct enumerations as erroneous while generating corresponding P-Sample nonmatches does not contribute bias to a measure of net error. This is commonly referred to as "balancing" of errors.

The census required two characteristics for a person record to be data defined. The characteristics that counted toward the two were relationship, sex, race, Hispanic origin, and either age or year of birth. If a valid name was present it was also counted towards the minimum two characteristics. A name had to have at least three characters in the first and last name together for it to be considered valid by the census. Anyone who was not data defined was a whole person imputation. Only data defined census enumerations were included in the E Sample.

The A.C.E. requirement for sufficient information for matching and followup was stricter than the census's requirement for data defined. The minimum amount of data required for the data defined census people to have sufficient information for matching and followup was a complete name and two characteristics. The acceptable characteristics were the same as those that the census counted for data defined except that the name did not count towards the two characteristics.

If a data defined census person had a blank or incomplete name, that person had insufficient information for matching and followup. To meet the definition of A.C.E. sufficient information for matching and followup a complete name was defined as follows:

(1) valid first name, middle initial, and valid last name, or

(2) valid first name and valid last name, or

(3) first initial, middle initial, and valid last name.

The minimum number of characters necessary for a first name or last name to be valid was two. There were cases where a name might have met the above requirements but was nevertheless deficient and the record was coded insufficient information for matching and followup. For example, names such as Mr. Doe, Donald Duck, and any other name that was most likely false, were coded "KE" by the clerical matchers.

In the A.C.E. the census person records were reviewed both by computer and clerically to identify people with insufficient information for matching and followup. Only people with sufficient information for matching and followup were allowed to be processed in the matching and followup interviewing phase.

The A.C.E. reviewed the image of the census questionnaire for census people coded as insufficient information for matching and followup to obtain additional data that might convert them to sufficient information for matching and followup. This included looking at census rosters to get names. A.C.E. also allowed children with first names but no last names to be processed as sufficient information in a household with an adult that had a first and last name. These updates to the names were captured into the matching software.

MATCHING IN THE KE STUDY

In this study we used Census 2000 and the original A.C.E. production files. The Generic Matcher was the matching software used by the analysts.

Data Preparation

To prepare the data for the Generic Matcher we decided that the housing unit would be the preferred work unit for the analysts. In other words, the analysts would look at the entire census housing unit (a housing unit being a unique Census Identification Number (CID)) wherein a KE was located. On the census side we included every record in housing units that had at least one KE record in it. These housing units are referred to as "KE households." The 13,360 (unweighted, not including Puerto Rico) KE records were located in 8,837 housing units. On the P-Sample side we looked at all housing units in the same cluster as the KE household. The Generic Matcher displayed the KE households on the census side and the P-Sample housing units within that cluster on the A.C.E. side. This was an appropriate layout considering the match codes, matching rules, and matching software.

Matching

One of the central goals of this study was to develop and test new methodologies that deal with enumerations formerly considered insufficient information for matching and followup. A main methodological advancement of this study was the development of match codes that accurately represent the KE cases and their match status. We recommend that in future attempts to match KE cases to the P Sample, the match codes described below be compressed into a smaller subset that maintains the overall structure of the codes. We had several goals in mind when creating these match codes. We wanted them to be mutually exclusive and exhaustive. No KE record can have two distinct codes and the match codes must cover all possible outcomes. We wanted the matching criterion to be as objective as possible so that the matching would be consistent. In other words, two analysts working independently will ideally produce the same match code for any given KE record.

Matches were based on person links and housing unit links as well as on name, date of birth, household composition, and other characteristics. The match codes reflect the nature and level of confidence of the match. Match codes are broken down into four levels:

Household level (H) match codes. A household level match code means that there exists at least one person match from the original A.C.E. between persons in the E-Sample and P-Sample housing units. This level match code gives us confidence that we are searching in the correct household.

Address level (A) match codes. An address level match code means that there does not exist a person match from the original A.C.E. between persons in these E-Sample and P-Sample housing units, however, the housing units were matched in the original A.C.E. during the housing unit match. This level match code gives us confidence that we are searching at the correct address.

Cluster level (P) match codes. A cluster level match code means that there does not exist a person match from the original A.C.E. between persons in the E-Sample and P-Sample housing units and the housing units were not matched during the original A.C.E. housing unit match, however, the E-Sample and P-Sample persons reside in the same cluster. This level match code gives us confidence that we are searching in the correct cluster.

Duplicate (KD) match codes. A duplicate match code was assigned for one of two reasons. Either the KE record matched another E-Sample record or the KE record matched a P-Sample record that was already coded a match in the original A.C.E.

Within each level of match code there are four types:

Type 1 match codes. A type 1 match code means that the determination of match status was made based on name.

Type 2 match codes. A type 2 match code means that the determination of match status was made based on date of birth.

*Type 3 match code*s. A type 3 match code means that the determination of match status was made based on characteristics other than name and date of birth.

Type 4 match codes. A type 4 match code means that the determination of match status was made not based on name, date of birth, or characteristics, but by non conflicting information between the E-Sample and P-Sample records. For instance, one record has a name, race and relationship that do not conflict with the other record that has only age and sex.

For KE records that do not match to a P-Sample record we assigned "Not a Match" codes, which are structured into roughly the same levels as the match codes.

Results

We will use results from this study to inform future discussions on the possibility of matching cases with insufficient information for matching and followup. For the purposes of illustration we define the following terms: *High Confidence Match, Medium Confidence Match, Low Confidence Match, Not a Match, and Other*. "*Other*" refers includes records that were updated to sufficient information for matching and followup and records located in incorrectly geocoded housing units. Table 1 shows the various match code levels and types along with the degree of confidence assigned. For example, the code MH1 (household-level, type 1) is a high-confidence match code, while the code MP4 (cluster-level, type 4) is a low-confidence match code.

Table 1. Match Confidence by Match Code

High	**MH1**	**MA1**	**MP1**	**KD1**
	MH2	**MA2**	**MP2**	**KD2**
Medium	**MH3**	**MA3**	**MP3**	**KD3**
Low	**MH4**	**MA4**	**MP4**	**KD4**

Table 2 presents the matching confidence of the KE records. The numbers in parenthesis represent standard errors that were calculated using the delete a cluster jackknife method. Table 2 tells us that approximately fifty percent of KE records can be matched with either high or medium confidence and that only about four percent match with low confidence. This suggests that matching KE records is quite feasible in the future. However, it does not tell us anything about the enumeration status of the KE records. For this we must investigate the original A.C.E. P-Sample residence and P-Sample match status of the matching P-Sample records.

Table 2. Match Confidence

	High Confidence Match	Medium Confidence Match	Low Confidence Match	Not a Match	Other	Total
Unweighted (SE)	3,383 (76)	3,365 (94)	558 (35)	5,713 (144)	341 (29)	13,360 (202)
Percent of Total	25.32	25.19	4.18	42.76	2.55	100
Weighted (SE)	1,292,286 (34,038)	1,306,027 (41,433)	209,133 (14,492)	1,856,338 (61,849)	99,597 (9,513)	4,763,381 (88,492)
Percent of Total	27.13	27.42	4.39	38.97	2.09	100

In this study KE records could match to P-Sample records with a variety of residence statuses. They could match to a P-Sample resident, or a P-Sample non-resident, in which case the KE record is considered to have a resolved enumeration status. They could also match to a P-Sample record with unresolved residence, in which case the KE record is considered to have unresolved enumeration status. KE records could also match to P-Sample records with a variety of match statuses. They could match to a P-Sample nonmatch. They could also match to a P-Sample match, or a P-Sample duplicate in which case the KE record is considered a duplicate. KE records could also match to other E-Sample records in which case the KE record is considered a duplicate as well. In order to estimate how many KE records may be considered correctly or

erroneously enumerated for component error estimation we need to look at the original A.C.E. match status and residence status of the records that were matched to the KEs. Table 3 displays this information.

Roughly twenty three (22.76) percent of records coded insufficient information for matching and followup in 2001 A.C.E., match (with a high level of confidence) to a person who lived at the address on Census Day and was not otherwise counted. In other words, these cases have a high probability of being correctly enumerated for component error estimation.

About twenty one (21.21) percent of records coded insufficient information for matching and followup in 2001 A.C.E., match (with a medium level of confidence) to a person who lived at the address on Census Day and was not otherwise counted. In other words, these cases have a high probability of being correctly enumerated for component error estimation. About six (0.29 + 0.36 + 2.57 + 2.56) percent of records deemed insufficient information for matching and followup in 2001 A.C.E. either match (with a high or medium degree of confidence) to a non-resident or were found to be duplicates. In other words, these cases have a high probability of being erroneously enumerated for component error estimation.

Table 3. Type of Matching Record by Match Confidence

Weighted Count (Standard Error) Weighted Percent of Insufficient Information Records (4,763,381)	High Confidence Match	Medium Confidence Match	Low Confidence Match
P-Sample Resident Nonmatch	1,083,909 (31,446) 22.76	1,010,195 (36,359) 21.21	156,950 (11,930) 3.29
P-Sample Non-Resident Nonmatch	13,621 (2,551) 0.29	17,031 (3,907) 0.36	715 (505) 0.02
P-Sample Unresolved Nonmatch	72,126 (6,736) 1.51	156,624 (12,746) 3.29	34,775 (6,026) 0.73
E-Sample Record, P-Sample Match, or P-Sample Duplicate	122,630 (9,554) 2.57	122,177 (11,158) 2.56	16,693 (3,424) 0.35

CLERICAL MATCHING RELIABILITY

As with any clerical matching operation, this one was subject to inconsistent coding despite our attempt to define matching rules that produce consistent results from different analysts. To evaluate this source of inconsistency, we selected a sample of block clusters and had the analysts recode the KE cases. The analysts began from scratch using the same procedures from the initial match without access to the results. The recoding was done independently of the initial matching, in other words the analysts did not work the same clusters both times. Of the 13,360 unweighted KE records 7,878 were recoded.

The diagonal in Table 4 shows the frequencies with which the initial matching and the recoding agreed on match confidence, for each level of match confidence. The off-diagonals show the frequency with which they disagreed. The initial match confidence is found on the column headings and the recoded match confidence is found on the row headings. When we add the totals on the diagonal in Table 4 (6,874) and divide by the total (7,878) we get the overall percent agreement between analysts, eighty seven percent. A discussion of the discrepancies observed most often follows:

Table 4. Recode vs. Initial Match Confidence

Cell Count
Row Percent
Column Percent

Initial Confidence / Recode Confidence	*High Confidence Match*	*Medium Confidence Match*	*Low Confidence Match*	*Not a Match*	*Other*	*Total*
High Confidence Match	1,503 91 93	87 5 5	7 0 2	38 2 1	11 1 7	1,646 100 21
Medium Confidence Match	61 3 4	1,429 80 82	164 9 44	131 7 3	2 0 1	1,787 100 23
Low Confidence Match	7 3 0	94 36 5	99 38 26	63 24 2	0 0 0	263 100 3
Not a Match	35 1 2	141 3 8	106 3 28	3,727 92 94	28 1 18	4,037 100 51
Other	10 7 1	0 0 0	0 0 0	19 13 1	116 80 74	145 100 2
Total	1,616 21 100	1,751 22 100	376 5 100	3,978 50 100	157 2 100	7,878 100 100

(1) One analyst assigned a high confidence code and the other assigned a medium confidence code. Discrepancies between high and medium confidence codes occurred most often for household and address level match codes, where one analyst made a match based on name and another analyst made a match based on characteristics. The source of the discrepancies between high and medium confidence codes appears to be different interpretations of a rule that describes the requirements the name has to meet to be the basis of a match.

(2) One analyst assigned a medium confidence code and the other analyst assigned a low confidence code. Discrepancies between medium and low confidence codes occurred most often when one analyst made an address level match based on characteristics and another analyst made an address level match based on non contradicting information. The source of the discrepancies between medium and low confidence codes appears to be different interpretations of a rule that describes the requirements for "distinguishing characteristics" in order to make a match based on characteristics.

(3) One analyst assigned a medium confidence code and the other analyst assigned a not a match code. Discrepancies between medium confidence and not a match codes occurred most often when one analyst made an address level match based on characteristics and another analyst did not make a match. The source of the discrepancies between medium confidence and not a match codes appears to be different interpretations of a rule that describes the requirements for "distinguishing characteristics" in order to make a match based on characteristics.

(4) One analyst assigned a low confidence code and the other analyst assigned a not a match code. Discrepancies between low confidence and not a match

codes occurred most often when one analyst made an address level match based on non-conflicting information and another analyst did not make a match. The source of the discrepancies between low confidence and not a match codes appears to be different interpretations of a rule that describes the requirements for "contradictory information" in order to make a match based on non-conflicting information.

These discrepancies indicate that the matching rules used in this operation are not clear for a number of cases and are often subject to differing interpretations by the analysts. Any future matching operation that attempts to match KEs should review these matching rules, and attempt to clarify them for these types of discrepancies.

CONCLUSIONS AND RECOMMENDATIONS

Of the E-Sample records with insufficient information for matching and followup, approximately half can be assigned a match status and enumeration status for component error estimation. This suggests that matching KE records is quite feasible and we recommend adopting similar methodology as was used in this study.

There is a high unresolved rate among KE records as many of these cases did not match or matched with low confidence and were not sent to followup. It may be assumed that a similar rate will be observed in the future and since most of these records do not have a discernable name they will not be followed up. Appropriate missing data procedures will have to be applied to these cases.

As stated in Section 1., the reason for strict rules in determining sufficient information for matching and followup is so that there is enough information to make an accurate determination of enumeration status, whether correct or erroneous. For this purpose the current definition of insufficient information for matching and followup is not consistent. Consider the following example: A record with a complete name and a full date of birth is considered insufficient information for matching and followup under the current definition. A record with a complete name, race, and Hispanic origin is considered sufficient information for matching and followup. Clearly the former record contains more informative matching data than the latter. Therefore there exists the need to amend the current definition of sufficient information for matching and followup to better reflect the purpose of the rule. We recommend the following definition: To be considered sufficient information for matching and followup the data-defined census enumeration must have a complete name as defined in Section 2, and two characteristics where either race, Hispanic origin, or ancestry only count as one of these characteristics (ancestry will be a new demographic characteristic collected in the 2006 Census Test and possibly beyond.) By letting either race or Hispanic origin only count as one characteristic in A.C.E. 2000, we would have increased the number of weighted KEs by 41,654, only 0.9 percent.

As explored in Section 4 the reliability of matching KE is not perfect. Particularly troublesome are the cases when we have a medium confidence match by one analyst and not a match by a different analyst. Any attempt to match KE cases in the future should address this source of variability by refining the matching rules.

LIMITATIONS

There were several limitations to this study, some of which would not be present in a production environment. The limitations are discussed below:

(1) The analysts did not have full access to the entire E Sample for matching. This limited the duplicate coding, especially in households with other duplicates or a duplicated housing unit. It also limited knowledge of person matching in the case of apartment mix-ups and mail mis-delivery. This limitation would not be present in a production environment.

(2) The analysts did not have access to images of census forms. This limited the ability to determine and update scanning error and fictitious people. This would not be a limitation in a production environment because the analysts would either have access to images of census forms or access to the census forms themselves.

(3) The analysts did not have full access to or census records for everyone listed on the A.C.E. roster. This limited the ability to match to inmovers, people born after census day, and people in group quarters. This limitation would not be present in a production environment.

(4) There was no followup. This limited the ability to obtain more information about matches with a low level of confidence and/or duplicates. This will continue to be a limitation for any record that does not have a complete name.

(5) The analysts did not have access to person followup, or Targeted Extended Search (TES) forms for person or housing unit operations. This limited the ability to code based on insight into household or housing unit

composition. This would not be a limitation in a production environment.

(6) The independence assumption of the clerical matching reliability check is at risk due to the fact that analysts routinely discuss difficult cases with each other. Although such collaboration was discouraged, we must assume that it did occasionally occur. This limitation would be irrelevant in a production environment because discussion between analysts is typically encouraged during matching operations.

(7) The P-Sample match codes were taken from the original A.C.E. files so we were not able to take advantage of any updates made to the codes during A.C.E. Revision II. The most up to date codes would be used in a production environment, therefore this would not be a limitation in such an environment.

REFERENCES

Feldpausch, R. (2002), "E-Sample Erroneous Enumerations," Executive Steering Committee For A.C.E. Policy II (ESCAP II) Report 5.

Childers, D. (2001), "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-01.

Singh, R. (2003), "Census Coverage Measurement-Goals and Objectives (Executive Brief)," DSSD 2010 Census Coverage Measurement Memorandum Series, A-1.