Census 2000 Evaluation F.15
August 19, 2003

An Assessment of Addresses on the Master Address File "Missing" in the Census or Geocoded to the Wrong Collection Block

FINAL REPORT

This evaluation study reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Megan C. Ruhnke
Planning, Research,
and Evaluation Division

USCENSUSBUREAU
Helping You Make Informed Decisions

Intentionally Blank

CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

One of the results of the Accuracy and Coverage Evaluation included a representative sample of addresses that were coded as "missing" from the census. This was a result of the independent listing, matching and field work that was conducted as part of the Accuracy and Coverage Evaluation work. Our evaluation conducted additional research to better understand these "missing" addresses and to examine the reasons for their status of "missing" after the Accuracy and Coverage Evaluation Final Housing Unit work was completed. We matched the addresses coded as "missing" to all non-duplicate housing units on the Master Address File in a larger geographic search area than the one used by the Accuracy and Coverage Evaluation. We searched for matches in the tract which included each address, and all surrounding tracts.

Our main focus in understanding these "missing" addresses was to determine if they were actually included in the census as housing units, but were incorrectly geocoded to a collection block outside of the Accuracy and Coverage Evaluation geographic search area. We were able to do this because we expanded the Accuracy and Coverage Evaluation search area so that we could find cases of geocoding error not found during the Accuracy and Coverage Evaluation. Our work also allowed us to better understand addresses on the Master Address File, or on the Decennial Master Address File, but excluded from the census. We were able to do this because we did not limit our matching to census units only, but included other addresses on the Master Address File. As a by-product of our work, we identified some census addresses that matched to an Accuracy and Coverage Evaluation "missing" unit geocoded to the same collection block, but not included in the census address list used for the Accuracy and Coverage Evaluation address matching. We attempt to explain why this happened.

What is the total estimated percentage of census addresses geocoded to the incorrect Census 2000 collection block?

The Accuracy and Coverage Evaluation limited its matching of addresses to within the block cluster, or in some cases, to one ring of surrounding blocks. We were able to improve on the estimate of geocoding error in this evaluation by matching addresses in a larger geographic search area, and therefore finding more cases of units geocoded in error in the census. The estimated percentage of census addresses that were geocoded to the incorrect Census 2000 collection block is 4.8 percent (standard error is 0.3 percent).

Did the geocoding error estimate vary by type of enumeration area?

Yes. The estimated percentage of geocoding error in the census is significantly higher in Mailout/Mailback enumeration areas (5.5 percent) than in Update/Leave (1.7 percent) or List/Enumerate areas (1.2 percent).

To some extent, we expect less geocoding error in Update/Leave and List/Enumerate areas because our address list was created on the ground through field operations, and therefore geocoding was based on first-hand field observation. This is different from Mailout/Mailback areas where geocoding was based on a combination of procedures, including an automated geocoding process. That combination could contribute to the higher geocoding error estimate in that type of enumeration area.

However, a lower geocoding error estimate in Update/Leave and List/Enumerate areas could also come about because those areas have a higher occurrence of non-city-style addresses, which makes it harder for us to detect geocoding errors due to matching limitations. We were very limited in our ability to match the rural addresses, and were therefore unable to find as many cases of geocoding error in rural areas as we were in Mailout/Mailback areas. A greater population of large multi-unit structures in Mailout/Mailback areas could also contribute to a higher geocoding error estimate of housing units in those areas.

Did the geocoding error estimate vary by size of structure?

Yes, geocoding error is more prevalent among housing units in multi-unit structures. Housing units in both small and large multi-unit structures have a significantly higher geocoding error estimate than single units or housing units in two-unit structures. Additionally, large multi-units (housing units in structures with ten or more units) have a significantly higher geocoding error estimate than small multi-units (housing units in structures with three to nine units). The geocoding error estimate for both single housing units or two-unit structures is about three percent, for small multi-unit structures is about five percent, and for large multi-unit structures is about 11 percent.

We would expect geocoding error to be higher for units in multi-unit structures because geocoding error is a structure-based problem. Geocoding the structure to the wrong block causes every unit in that structure to be geocoded to the wrong block. The larger the structure is, the larger the number of geocoding error cases there will be if the structure is geocoded to the incorrect block.

Did the geocoding error estimate vary by census region or Regional Office?

Yes, geocoding error of census addresses is less frequent in certain regions of the country. The geocoding error estimate for the Midwest (3.8 percent) is significantly lower than the geocoding error estimate for the South (5.7 percent). There are no other significant differences.

Geocoding error estimates also differ for some of the Regional Offices. The Boston and Kansas City Regional Offices both had a significantly lower geocoding error estimate than the national estimate of 4.8 percent.

One might think that the differences between census regions and Regional Offices are driven by the differences we saw in geocoding errors between different structure sizes or Type of Enumeration Areas. We attempted to analyze this, but found no definitive results.

What else were we able to learn by matching addresses coded as “missing” during the Accuracy and Coverage Evaluation to the Master Address File?

As previously stated, one of the reasons addresses were coded as "missing" from the census during the Accuracy and Coverage Evaluation was because they were incorrectly geocoded in the census to a collection block outside of the scope of the Accuracy and Coverage Evaluation's geographic search area. The addresses that fell into that category were discussed above.

However, there were also some "missing" addresses that we found in this evaluation that were geocoded correctly. That is, they were geocoded to a block within the Accuracy and Coverage Evaluation's scope, but were not included in the census address list used for the Accuracy and Coverage Evaluation address matching.

Additionally, since we did not limit our matching to only census addresses, but included all non-duplicate housing units on the Master Address File, we were able to examine addresses that we had on the Master Address File or the Decennial Master Address File but were excluded from the census. These situations are discussed in the questions below.

Why are there addresses in the census, geocoded to an Accuracy and Coverage Evaluation sample block, but coded as "missing" by the Accuracy and Coverage Evaluation?

About 8,900 of the units coded as "missing" by the Accuracy and Coverage Evaluation were matched to units on the Master Address File during this evaluation. About 4,800 of them were matched to addresses that were included in Census 2000. Of those census matches, about 3,100 were geocoded in error in the census to a collection block that was different than the block provided by the Accuracy and Coverage Evaluation. The other 1,700 units were matched to census addresses that were geocoded to the same block as the Accuracy and Coverage Evaluation "missing" addresses. There are two primary reasons that these census units were not included in the census address list used for the Accuracy and Coverage Evaluation address matching.

The first reason is that some of these units were identified as potential duplicates during the Census 2000 Housing Unit Unduplication operations and were therefore kept out of the Accuracy and Coverage Evaluation Final Housing Unit matching operation. About 78 percent of the matches to in-census units in the same block were potential duplicates that ultimately were reinstated in the census.

The remaining 22 percent of the in-census matches to Accuracy and Coverage Evaluation "missing" units in the same block were not reinstated duplicates. A reason that these units were excluded from the address list used for the Accuracy and Coverage Evaluation address matching is that they were not geocoded to an Accuracy and Coverage Evaluation sample block at the time of the Final Housing Unit matching, but were moved into an Accuracy and Coverage Evaluation sample block in time for our evaluation work. Examples of this include units that were moved to a different block following the Geographic Misallocation operation.

How many addresses, coded as "missing" from the census during the Accuracy and Coverage Evaluation, did we have on the Master Address File, but exclude from the census?

Of the approximately 8,900 addresses coded as "missing" by the Accuracy and Coverage Evaluation that we matched to the Master Address File in this evaluation, about 4,000 were not included in Census 2000. That is, these units were listed and confirmed as good, residential addresses during the Accuracy and Coverage Evaluation, but the Census Bureau's rules for creating the Decennial Master Address File and the Hundred Percent Census Unedited File excluded them from the census.
Those units represent a weighted estimate of 1.3 million units coded as erroneously excluded from

the census as measured by the Accuracy and Coverage Evaluation, and this evaluation.

About 28 percent of those cases we coded as erroneously excluded units were never delivered to the Decennial Master Address File. There are a number of reasons units on the Master Address File would have not been sent to the Decennial Master Address File as a result of the Census Bureau's rules for developing the Census 2000 address frame. One of the reasons a unit would not be included on the Decennial Master Address File is if it was coded by the United States Postal Service as nonresidential on the Delivery Sequence File. We excluded those addresses from the original census address list because it would not be prudent to mail questionnaires to all nonresidential addresses. We relied on field listing operations to add those units if they were actually residential units by Census Bureau definitions.

About 49 percent of the cases we coded as the erroneously excluded units were on the Decennial Master Address File, but were then deleted during the Census 2000 Kill Process. The goal of the Kill Process was to identify units that were most likely bad addresses and remove them from the census. An example of a unit that was deleted during the Kill Process is a case for which we received no census form and the unit was deleted in both the Nonresponse Followup and Coverage Improvement Followup operations.

About 22 percent of the cases we coded as erroneously deleted units were on the Decennial Master Address File but were determined to be potential duplicates during the Housing Unit Unduplication operations through address and person matching algorithms. We ultimately decided to exclude those units from Census 2000. The amount of erroneous deletions from the Unduplication operation as measured in this evaluation is potentially overstated. This comes from the fact that the Accuracy and Coverage Evaluation may have coded something as missing from the census, when it was actually included in the census with a different form of the address. The Unduplication operation may have recognized the duplication but removed the version of the address that the Accuracy and Coverage Evaluation listed.

Recommendations

We are currently researching the possibility of collecting Global Positioning System coordinates for addresses in the census. One reason for doing this is to help enumerators find their assignments. Another reason would be to ensure geocoding units to the correct block. If using Global Positioning System coordinates for improving geocoding is a high priority, our emphasis for the use of the Global Positioning System should not be entirely focused around rural areas. Instead, we should consider getting better geocoding for the areas with the highest geocoding error rates, which are Mailout/Mailback areas.

We also recommend research to refine procedures for identifying and deleting units we believe to be duplicates. Despite the limit stated above, the unduplication process appears to have deleted many units which should have been included in the census. However, the unduplication process was introduced very late in Census 2000 without sufficient planning. Work has already begun on building an unduplication process into the 2010 census.

1. BACKGROUND

One of the results of the Accuracy and Coverage Evaluation (A.C.E.) included a representative sample of addresses that were coded as "missing" from the census. This was a result of the independent listing, matching and field work that was conducted as part of the A.C.E. work. Our evaluation conducted additional research to better understand these "missing" addresses and to examine the reasons for their status of "missing" after the A.C.E. Final Housing Unit (FHU) work was completed. We matched the addresses coded as "missing" to all non-duplicate housing units on the Master Address File (MAF) in a larger geographic search area than the one used by the A.C.E. We searched for matches in the tract which included each address, and all surrounding tracts.

Our main focus in understanding these "missing" addresses was to determine if they were actually included in the census as housing units, but were incorrectly geocoded to a collection block outside of the A.C.E. geographic search area. We were able to do this because we expanded the A.C.E. search area so that we could find cases of geocoding error not found during the A.C.E. Our work also allowed us to better understand addresses on the MAF, or on the Decennial Master Address File (DMAF), but excluded from the census. We were able to do this because we did not limit our matching to census units only, but included other addresses on the MAF. As a by-product of our work, we identified some census addresses that matched to an A.C.E. "missing" unit geocoded to the same collection block, but not included in the census address list used for the A.C.E. address matching. We attempt to explain why this happened.

Similar operations were previously conducted to look at geocoding error and other housing unit coverage estimates. The results are discussed below.

1.1 The 1990 Housing Unit Coverage Study (HUCS)

The Housing Unit Coverage Study (Childers, 1992) of the 1990 census measured the quality of coverage of housing units (HUs) enumerated in the census. The HUCS sample consisted of two parts, the Population sample (P-sample) and the Enumeration sample (E-sample). The P-sample was an independent listing of HUs in a sample of blocks. The E-sample was the list of housing units enumerated in the 1990 census in the same sample of blocks.

The P-sample and E-sample addresses in the HUCS were computer matched, and the match results were then clerically reviewed. Nonmatched HUs, unresolved cases, and possible duplicates were sent to the field for a followup interview.

A code was assigned to each E-sample HU based on the accuracy of the geocoding. The geocoding status for the sample housing units was classified as “correct,” “incorrect - within the search area,” “incorrect - outside the search area,” “insufficient information for field interviewing,” and “unresolved.”

For Tape Address Register (TAR) areas, "incorrect - within the search area" was defined as "the correct geography was one block from the enumerated geography" and "incorrect - outside the search area" was defined as "the correct geography was more than one block from the enumerated geography." For Prelist areas and Update/Leave areas, "incorrect - within the search area" was defined as "the correct geography was one or two blocks from the enumerated geography" and "incorrect - outside the search area" was defined as "the correct geography was more than two blocks from the enumerated geography." For List/Enumerate areas, "incorrect - within the search area" was defined as "the correct geography was inside the address register area (ARA)" and "incorrect - outside the search area" was defined as "the correct geography was outside the ARA."

The HUCS found that the collection geography was correct for 93.92 percent of the housing units and incorrect for 4.54 percent of the housing units. The remaining 1.54 percent of HUs were unresolved or had insufficient information for field interviewing. "Incorrect" geography included units whose geography was incorrect both inside and outside of the search area.

The HUCS examined the percentage of incorrect geography by type of enumeration area (TEA). The estimated percentage of HUs with incorrect geography in the two Mailout/Mailback areas of the 1990 census – TAR areas and Prelist areas – was significantly different from the estimated percent of HUs with incorrect geography in List/Enumerate areas. The estimated percentage of housing units geocoded incorrectly in TAR and Prelist areas were 4.7 percent and 4.9 percent, respectively. The estimated percentage of geocoding error was 2.6 percent in List/Enumerate areas. Other than the differences just discussed, no other significant differences in geocoding error existed between the TEAs.

1.2 The 1998 MAF Quality Improvement Program (QIP)

The 1998 MAF QIP (Barrett, 1999) evaluated the effectiveness of the initial MAF in accurately reflecting housing units that existed on April 1, 1998 in Census 2000 mailout/mailback areas by producing estimates of coverage and coding errors on the MAF. The study selected a sample of counties and a subsample of block clusters within the counties. Field representatives created an independent listing of the HUs in those block clusters, and the result was matched to the addresses on the MAF at a ZIP code level.

The 1998 MAF QIP estimated that 6.23 percent of residential MAF addresses existing in the sample clusters were geocoded in error on the MAF. Another result of the QIP was an estimate of the percent of the addresses on the MAF that were coded incorrectly as non-residential. QIP estimated that only about one tenth of a percent of MAF addresses were incorrectly coded as non-residential on the MAF.

1.3 The Accuracy and Coverage Evaluation in Census 2000

The A.C.E. measured the overall and differential coverage of the U.S. population and housing in Census 2000. An independent listing of all HUs in the A.C.E. sample clusters was conducted before census day. This listing (the P-sample) was then computer-matched to the census housing units included on the January 2000 DMAF in the sample block clusters (the E-sample). A unit was considered a match if it was linked to an address within the cluster, even if the units were listed in different blocks within the cluster.

The results of the computer matching and additional information were loaded into a clerical matching software database. A clerical match was then performed on many of the sample clusters, and match codes were assigned to any A.C.E. and census units that were not matched by the computer. Any addresses that were not matched after the clerical review were sent to the field for a followup interview.

Since the initial housing unit matching occurred before the inventories of the census and A.C.E. housing units were final, another clerical matching operation was performed. Only the updates to the census and A.C.E. housing unit inventories were processed during the FHU match. These updates included the removal of units that were identified as potential duplicates in the first phase of the Housing Unit Unduplication Operations (see section 1.5).

The FHU match used only the census records that appeared in the Hundred Percent Census Unedited File (HCUF). Late census adds, or any updates that happened after the creation of the HCUF were not accounted for in the A.C.E. FHU matching operation.

There was no computer matching for the FHU match, but there was a computer processing stage to determine which addresses went to the clerical stage. After the clerical matching, a followup interview was conducted for selected cases that were not followed up during the initial phase of housing unit matching.

In addition to a search within the A.C.E. sample blocks, a targeted extended search (TES) was conducted for selected block clusters. In the clusters included in TES, P-sample nonmatches were matched to a ring of blocks surrounding the A.C.E. block cluster in order to find additional matches between the P-sample and the E-sample.

One could produce a geocoding error estimate based on the results of the A.C.E., but it was limited geographically to the block cluster, or in some areas, one ring of adjacent blocks surrounding the cluster. The FHU match results from the A.C.E. sample were used as the starting point in this evaluation for estimating the block-level geocoding error.

The FHU match results are also used to examine the addresses that were on the MAF, listed and confirmed as good residential addresses during the A.C.E. but not included in the final Census 2000 counts.

1.4 DMAF deliverability criteria

The U.S. Census Bureau developed a set of specifications for the initial delivery of addresses from the MAF to the DMAF which occurred in July and August of 1999. These DMAF deliverability rules looked at the several operations and files from which the Census Bureau received addresses for the MAF prior to the creation of the initial DMAF. The rules looked at the following files and operations that updated and provided new addresses to the MAF:

- The 1990 Address Control File (ACF)
- The November 1997 Delivery Sequence File (DSF) from the United States Postal Service (USPS)
- The September 1998 DSF from the USPS
- The Census 2000 Block Canvassing Operation
- The 1998 Local Update of Census Addresses (LUCA)
- The Census 2000 Address Listing Operation
- LUCA 1999 Recanvass
- Census 2000 Dress Rehearsal

The DMAF deliverability criteria were intended to rectify the housing unit coverage problems observed during the Census 2000 Dress Rehearsal. The goal of the rules was to keep units that were most likely not good residential addresses out of further Census 2000 processing.

Some examples of the types of units that were *not* delivered to the DMAF include:

- Inside-the-blueline ungeocoded addresses[1]
- Units determined to be duplicates in Block Canvassing
- Units classified as LUCA 98 or LUCA 99 Recanvass deletes
- Units with a negative action from Block Canvassing, that were not added in LUCA 98, and were not determined to be residential on the first two DSF deliveries.

For the complete explanation of the DMAF deliverability rules, refer to DSSD Census 2000 Procedures and Operations Memorandum, Series #D-1.

1.5 Reasons for exclusion of units on the DMAF from Census 2000

The Census Bureau attempted enumeration for all of the units on the DMAF. When Census 2000 enumeration operations were completed, we determined which addresses should be considered valid housing units. These valid units were included in the census. The invalid housing units were excluded from the census.

[1] Inside-the-blueline TEAs include Mailout/Mailback(1), Military(6), Urban Update/Leave(7) and Urban Update/Enumerate(8). Outside-the-blueline TEAs include Update/Leave(2), List/Enumerate(3), Remote Alaska(4), Rural Update/Enumerate(5) and Mailout/Mailback converted to Update/Leave(9).

There were four operations/processes that identified housing units on the DMAF that were excluded from the census. The four operations are:

- The Kill Process – This process identified MAFIDs that most likely did not uniquely identify housing units as of census day. One example of a housing unit that was excluded from the census as a result of this process is:
  - there was no census form returned for the unit,
  - the unit was deleted in the Nonresponse Followup (NRFU) operation and,
  - the unit was confirmed as a delete in the Coverage Improvement Followup (CIFU) operation.

  A total of approximately 8.3 million HUs were killed during the Kill Process.[2]

  For more information on the types of units excluded from the census as a result of the Kill Process, refer to DSSD Census 2000 Procedures and Operations Memorandum Series #D-13*: Specification of the Kill Universe on the Decennial Master Address File for Census 2000*.

- The Assignment of Housing Unit Status – The assignment of housing unit status was performed on the housing units that were not killed during the Kill Process. However, during this assignment some units had inconsistent information. Based on the available information certain classes of housing units were ultimately excluded from the census. This process of assigning a housing unit status resulted in an additional 22,352 housing units on the DMAF being excluded from the census. An example of a unit deleted during this process would be a case that came back from NRFU as a completed case, but once the form was data-captured, we realized that there was no data for the unit.

  For more information on the types of units excluded from the census as a result of this process, refer to DSSD Census 2000 Procedures and Operations Memorandum Series #D-14*: Specifications for Assigning the Housing Units Status and Population Count on the Hundred-Percent Unedited File Prior to the Imputation for Unclassified Units*.

- Unclassified Estimation – After the process of the assignment of housing unit status, there were some units which did not have enough information to determine whether to include or exclude the unit in the census. For these cases whose status was unclassified, the housing unit status was determined through a nearest-neighbor hot-deck imputation procedure. As a result of the imputation, an additional 46,196 housing units were excluded from the census.

  For more information on the types of units excluded from the census as a result of this process, refer to DSSD Census 2000 Procedures and Operations Memorandum Series #Q-2*: Census 2000 Overview of Unclassified Estimation.*

[2] Excludes killed units in Puerto Rico and units with a surviving MAFID.

- The Housing Unit Unduplication Operation – This operation consisted of two phases. The first phase involved the identification of potential duplicates on the DMAF through address and person matching algorithms. The second phase involved the development of rules to determine which housing units would be excluded from the census. As a result of applying the rules approximately 1.4 million housing units were excluded from the census.

  For more information on the types of units excluded from the census as a result of the unduplication operation, refer to DSSD Census 2000 Procedures and Operations Memorandum Series #D-10*: Specifications for Eliminating Duplicate Records on the Hundred Percent Census Unedited File* and DSSD Census 2000 Procedures and Operations Memorandum Series #D-11*: Specification for Reinstating Addresses Flagged as Deletes on the Hundred percent Census Unedited File.*

2. METHODS

In some of the following sections, we present unweighted counts of the clerical matching and field work results. We do this only to present the magnitude and matching outcomes of the work done in this evaluation. The weighted estimates as they relate to the outcomes of our evaluation matching work are presented in the results section (Section 4).

2.1 Sample selection

The sample used in this evaluation is the same set of sample block clusters used in the A.C.E. with the exception of any clusters in Puerto Rico; *Puerto Rico is not included in this evaluation.*

For the specification of the A.C.E. block sample selection, refer to DSSD Census 2000 Procedures and Operations Memorandum, Series #R-3.

2.2 Stages of matching

Similar to the A.C.E. FHU matching operation, this evaluation had several stages. It began with a computer processing operation which created the files to be read into the clerical matching software, pulling in information from the A.C.E. independent listing (P-sample), the census file used in the A.C.E. (E-sample) and address information for units on the MAF that were not included in the A.C.E.

After computer processing was complete, a Before Followup (BFU) clerical match was performed using the Geocode Matching, Review, and Coding System (GEO MaRCS). During the BFU clerical matching, a match was performed between units that were not matched during the A.C.E. FHU match and addresses on the MAF in the search area (see section 2.3 for more details about the search area).

All A.C.E. nonmatched units were assigned match codes, and selected matched MAF addresses were designated as requiring Field Followup (FFU). After the FFU was complete, an After Followup (AFU) clerical match was performed and units were assigned final match codes to be used for estimation purposes.

2.3 The matching search area

In an attempt to optimize the occurrences of geocoding error we could find, we matched P-sample nonmatches to all geocoded, non-duplicate MAF addresses in a geographic area surrounding the address. The geographical search area for a given cluster included all census collection blocks in the census pseudotract which contained the A.C.E. block cluster, and all blocks in any pseudotract whose boundary touches the pseudotract containing the A.C.E. cluster. A surrounding pseudotract may cross county or state boundaries. A pseudotract is a 1990 census tract that was adjusted to Census 2000 collection geography. Note that we matched to all addresses on the MAF, including units that were not in the census. We matched to all addresses on the MAF so that we could determine if any of the addresses coded as "missing" from the census during the A.C.E. were actually on the MAF, but excluded from the census.

The search area used in this evaluation was larger than the one used in the A.C.E. operations. This allowed us to find additional cases of geocoding error that were not found in the A.C.E. Additionally, by matching to all addresses on the MAF, including non-census addresses, we were able to better understand the addresses coded as "missing" by the A.C.E.

2.4 Computer processing

To prepare files for use in the clerical matching software, the GEO MaRCS, the Decennial Statistical Studies Division (DSSD) did the following:

- Pulled in and merged information from the March 2001 MAF extract files, the A.C.E. FHU files, and the Geography Division (GEO) surrounding block files to create the files for this evaluation.

- Removed records from the files that should not be in the evaluation, including MAF and E-sample records known to be duplicates, group quarters or special place addresses, and addresses in the P-sample that are known to be included in error.

- Re-assigned match codes to addresses in the E-sample and P-sample to be consistent with codes used in this evaluation.

2.5 Before Followup clerical matching and Field Followup

There are three types of addresses in the universe for this evaluation after the computer processing stage, which include:

- Within-cluster matches from A.C.E.
- P-sample matches to census units in a block surrounding the cluster (matches found during the A.C.E. TES)
- Unlinked P-sample units (nonmatches from A.C.E., considered to be "missing" from the census)

Only the unlinked P-sample units were attempted to be matched during BFU, and a subset of those matches were sent to followup. A subset of the A.C.E. within-cluster matches was also sent to followup based on an automatic block code comparison done by the system. The TES matches were not examined during BFU and were not sent to FFU. They were simply coded as geocoding errors automatically by the system (see Section 2.5.2 below).

2.5.1 Within-cluster matches from A.C.E.

The GEO MaRCS system automatically compared block codes for all within-cluster matches from the A.C.E. We did this because, in the A.C.E., units were considered to be matches if they were found within the block cluster, even if they were geocoded to different blocks within the cluster. All A.C.E. matches with different P-sample and E-sample block codes went to FFU to determine the correct block for the unit.

There were 261,525 within-cluster matches from the A.C.E. operation brought into our evaluation. About 98 percent of them had the same block code and did not require followup. The remaining two percent of units were matches within the cluster, but had conflicting E-sample and P-sample block codes. We sent those units to FFU to determine the correct block code. Units with the incorrect census (E-sample) block code are included as geocoding errors in our estimates.

2.5.2 The A.C.E. Targeted Extended Search matches

During the A.C.E. Targeted Extended Search, P-sample nonmatches in a subsample of A.C.E. clusters were matched to a ring of blocks surrounding the A.C.E. block cluster in order to find additional matches between the P-sample and the E-sample. Approximately 11,800 addresses were matched to census addresses during the TES. The A.C.E. conducted a field visit for TES matches to confirm that the units were located in the A.C.E. cluster, and not in the surrounding block. Therefore, we decided not to do another field followup for these units. All of those addresses are counted as geocoding errors in this evaluation.

2.5.3 P-sample nonmatches

All of the units on the P-sample that were not linked to census addresses at the end of the FHU operation are P-sample nonmatches, that the A.C.E. considered to be units that should have been included in Census 2000, but were not. As previously stated, we took those addresses and matched them to all non-duplicate housing units on the MAF in a larger geographic search area than the one used by the A.C.E. Analysts at the National Processing Center (NPC) reviewed each of those addresses and attempted to find a match during the BFU clerical matching operation.

If the analysts were unable to find a MAF address to link to the P-sample nonmatch, the unit was coded as unmatched, and was removed from our sample. If the P-sample unit was matched to a MAF address in the same collection block, the unit was coded as a match and did not require field followup. If the P-sample unit was matched to a MAF unit in a different block, or if two units were linked as a possible match, they were sent to field followup.

There were approximately 16,700 P-sample nonmatches brought into this evaluation for matching. Those addresses were considered to be good, residential addresses as of census day in the A.C.E. Housing Unit operations, but were not found on the census address list used for A.C.E. matching.

During the BFU clerical matching operation:

- Analysts matched about 30 percent of the unlinked P-sample addresses to a MAF address in the same collection block, which required no followup.

- Analysts matched about 33 percent of the unlinked P-sample addresses to a MAF address in a different collection block, which were sent for a field followup.

- Analysts found a possible match, either in the same block or a different block, for about four percent of the unlinked P-sample addresses. Those addresses were also followed up in the field.

- Analysts were unable to find a match for the remaining 33 percent of the P-sample nonmatches.

Since we were matching to such a large search area, it was possible to generate “false matches” during the BFU clerical match. This occurs because the address can legitimately occur in more than one area, typically when a search area crosses ZIP codes or counties. Our FFU procedures allowed us to catch these false matches and unlink them.

A field representative went to the block to which a unit was geocoded on the MAF. For these new matches to P-sample units in different blocks, if the unit was found in the MAF block, we considered it to be a false match, and unlinked those units. We considered these false matches because we assume that the A.C.E. address actually exists in the P-sample block. We were confident in this assumption because extensive field work was conducted during the A.C.E. on the P-sample nonmatches.

If we went to the MAF block during FFU and found the unit, we assumed that it is *not* the same unit as the P-sample nonmatched address since it cannot exist in both the P-sample and MAF blocks. If the MAF address did *not* exist in the MAF block, we considered this a true match, assumed that the unit actually existed in the P-sample block and counted it as geocoding error on the MAF.

For the A.C.E. cluster matches, FFU allowed us to determine if the address was geocoded correctly to the MAF block or incorrectly to the MAF block. For possible matches, FFU allowed us to determine if the linked P-sample and MAF addresses represented the same housing unit.

2.6 After Followup (AFU) clerical matching

Each unit that was sent to FFU was assigned a code during After Followup clerical matching.

The assigned match codes reflected characteristics such as:

- Match Status (match, nonmatch, unresolved)
- Geographic Level of match (within the block, outside of the block)
- Correct Geography (A.C.E. correct, census correct)

Of the approximately 4,500 A.C.E. cluster matches that required field followup, field staff determined that:

- The census block was incorrect for 40 percent of the matches.
- The P-sample block was incorrect for 56 percent of the matches.
- The correct block was unresolved for 4 percent of the matches.

Of the 5,500 new matches to a different block that were sent to FFU, field staff determined that:

- About 64 percent were determined to be true matches to an address in a different block, where the MAF unit is geocoded in error.

- About 36 percent of the cases were determined to be false matches. That is, the same address can legitimately occur in more than one area, so those were not cases of geocoding error.

- Less than one percent of the cases could not be resolved.

Of the 377 possible matches to an address in the same block, about 70 percent came back as confirmed matches. For about 26 percent, it was determined that the addresses did not represent the same housing unit, and for about four percent, the status was unresolved.

Of the 320 possible matches to an address in a different block, about 20 percent came back as confirmed matches. For about 76 percent, it was determined that the addresses did not represent the same housing unit, and for about three percent, the status was unresolved.

At the end of the AFU clerical matching, about 53 percent (8,869 addresses) of all the P-sample nonmatches were linked to an address on the MAF. About 46 percent remained nonmatches, and less than 0.5 percent were unresolved.

Approximately 40 percent of those 8,869 matches were geocoding errors on the MAF. The other 60 percent are matched inside the same block, representing either matches to cases left out of the census, or cases that were in the census but excluded from the A.C.E. matching.

Of the 8,869 new matches, approximately 55 percent were matched to a Census 2000 address. The remaining 45 percent matched to addresses on the MAF but not in the census.

We use the results of both the new matches from this evaluation and the A.C.E. cluster matches in our estimation.

2.7 Estimation and analysis

As discussed in previous sections, the focus of this evaluation was to examine addresses that were coded as "missing" from the census during the A.C.E. We did work to determine if those "missing" addresses were actually included in the census as housing units, but were incorrectly geocoded in the census to a collection block outside of the scope of the A.C.E. geographic search area. Our work also allowed us to examine some census addresses that matched to an A.C.E. "missing" unit geocoded to the same collection block, but were not included in the census address list used for the A.C.E. address matching. Finally, our work allowed us to determine if we had the "missing" addresses on the MAF, or the DMAF, but chose to exclude them from the census.

2.7.1 Geocoding Error Estimation

As previously stated, we conducted field work during this evaluation to determine if a census address that we matched to an A.C.E. "missing" address was geocoded in error. We also used the results from the A.C.E. matching to determine geocoding errors of units that were matched during the A.C.E. Out of that clerical matching and field followup work, we were able to produce and estimate of the percent of census addresses that are geocoded to the incorrect collection block.

Two types of geocoding errors exist:

- *geocoding errors of exclusion* – a unit that exists in a particular block but is geocoded on

the MAF outside of the sample block.

- *geocoding errors of inclusion* – a unit that exists outside of the sample block but it is geocoded on the MAF inside of the sample block.

In this evaluation, we only look at geocoding errors of exclusion for units included in Census 2000. That is, units that exist in an A.C.E. sample block, but are geocoded on the MAF to a different Census 2000 collection block.

$$geocoding\ error = \frac{\text{all housing units in the census that exist in the A.C.E. sample clusters that are geocoded to an incorrect block}}{\text{all housing units in the census that exist in the A.C.E. sample clusters}}$$

We define "exist in the sample clusters" as any census address that was confirmed to exist in the sample cluster during the A.C.E. or during our work for this evaluation. This includes:

- census housing units that matched to the P-sample in the A.C.E., regardless of whether the census had that address in the correct block.

- census housing units that were not matched to P-sample housing units in the A.C.E., but were matched to P-sample housing units in this evaluation.

- census housing units that did not match to the P-sample in the A.C.E., but were confirmed to exist in the sample clusters in the A.C.E.

2.7.2 Census 2000 addresses excluded from the census address list used for A.C.E. matching

As a by-product of our work, we identified some census addresses that matched to an Accuracy and Coverage Evaluation "missing" unit geocoded to the same collection block, but not included in the census address list used for the Accuracy and Coverage Evaluation address matching. We attempt to explain why this happened.

Some addresses were excluded from the census address list used in the A.C.E. because at the time of A.C.E. matching they were part of a universe of addresses considered to be potential duplicates. Ultimately, a subset of these potential duplicates was reinstated in the census (see Section 1.5). Those results were available for us during this evaluation, so we are able to examine how often "missing" addresses actually matched to one of those reinstated duplicates.

Another reason a unit may not have been included in the census address list used in A.C.E. is because, at the time of the development of the address list, the census unit was geocoded to a collection block that was not included in the A.C.E. sample.

However, during further processing, the census unit's geocoding was changed to a block within the A.C.E. sample. Since those geocoding changes had already been incorporated on the MAF at the time of our evaluation, we were able to match the "missing" addresses to those units.

2.7.3 MAF addresses excluded from the final Census 2000 results

Although the estimate of geocoding error was limited to units included in Census 2000, we did include other addresses on the MAF that were not included in the census in the matching and field work in this evaluation. Some of those addresses were delivered to the DMAF and others were not. Our analysis will look at the frequency that addresses on the MAF that were not in the census matched to the P-sample addresses considered to be "missing" from the census.

We did this analysis by looking separately at units that were on the MAF but never delivered to the DMAF, and units that were delivered to the DMAF but were excluded from census. Units on the MAF that were never delivered to the DMAF represent situations where at the time units were delivered from the MAF to the DMAF, our best information told us that these were not valid housing units where we should attempt enumeration. These also include addresses from the 1990 ACF and the DSF that are in outside-the-blueline areas where we did not use those addresses as a source for building the address list, because presumably other census operations would have listed those addresses.

DMAF addresses that were excluded from the census included:

- Units that were deleted as a result of the Kill Process
- Units that were deleted as a result of the specifications for assigning the housing unit status on the HCUF
- Units that were deleted as a result of unclassified estimation
- Units that were confirmed as duplicates in the Housing Unit Unduplication operation.

2.7.4 Unresolved cases

At the end of processing, of all of the cases that were sent to the field for resolution and then clerically reviewed again, we had a number of cases where we could not determine the final status. That is, we were unclear whether they were geocoding errors or not, or whether an A.C.E. nonmatched case really matched to a case we found on the MAF in the same block or in a different block.

One approach to handling these cases would be to impute final status values for them. Because there was such a small number of unresolved cases, we examined what would happen to the estimates if we treated all of the unresolved cases the same way. That is, we wanted to see how critical their resolution was to the final estimates we would be reporting.

To determine the impact of unresolved cases on the geocoding error estimates, we computed the estimates three different ways, by:

- Excluding the unresolved cases
- Including the unresolved cases and assuming a worst-case scenario
- Including the unresolved cases and assuming a best-case scenario

To accomplish the worst-case and best-case scenario estimates, we considered an unresolved case to actually be resolved by assigning an appropriate final status in order to include these cases in the estimates. In one case, we coded them to a final status concluding they were all cases of geocoding error (the worst case scenario) and in the other case, we coded them to a final status concluding that they were all geocoded correctly (the best case scenario.)

For each statistic that we provide in this report, if we were to assign a final status to all unresolved cases, assuming geocoding error in all cases, the impact on the statistic was never greater than one-fifth of a percentage point. The impact was even less for the best-case scenario estimates (that is, assigning a final status to all unresolved cases, assuming no geocoding error in all situations led to changes in the statistics by less than one-fifth of a percentage point.)

For example, the geocoding error estimate in Rural Update/Enumerate areas excluding unresolved cases is 1.74 percent. The best-case estimate is 1.73 percent, which is better than the no-unresolved estimate by 0.01 percent, and the worst-case estimate is 1.91 percent , which is worse than the no-unresolved estimate by 0.17 percent.

Because of the small number of unresolved cases (see section 2.6) and the minimal impact they have on the estimates, unresolved cases are excluded from all numbers presented in this report.

2.7.5 Weighting

The sample for this evaluation consists of all the collection blocks in the A.C.E. sample clusters. Therefore, we used weights from A.C.E. in this evaluation. For all linked units in the universe for this evaluation, we used the P-sample unbiased cluster weight. For unlinked E-sample units, we use the E-sample trimmed weight, the TES weight and the TES 2 weight. Additionally, we used the correct enumeration probabilities for unresolved E-sample cases.

For details on how the A.C.E. sample weights were computed, refer to DSSD Census 2000 Procedures and Operations Memorandum Series #Q-80.

2.7.6 Variance estimation

To calculate the variances and standard errors for the estimates, we used a SAS program developed by DSSD staff. This program uses the stratified jackknife method to calculate standard errors for ratio estimates and accounts for the A.C.E. sample design.

2.7.7 Statistical significance testing

The differences presented in the results section of this report were tested for significance. To determine statistical significance, we constructed 90 percent confidence intervals using the standard errors of the estimates and the critical value 1.645. If the confidence intervals contained zero or did not overlap, the estimates were significantly different at the $\alpha = .10$ level. Otherwise, the estimates were not significantly different.

2.8 Original source of an address

Evaluations of the MAF-building operations required identification of the source of every address on the MAF. An original source variable, which did not exist on the MAF, was defined and created by the Planning, Research, and Evaluation Division and the DSSD. This variable identifies the first operation or file to add the address to the MAF, with the following three exceptions:

- If one operation added an address, but a later operation also identified the address in a different TEA, the first operation does not receive credit for adding this address.

- An address may not have sufficient operation information to indicate how the address was added to the MAF.

- In cases where one MAF-building operation overlapped with at least one other MAF-building operation and the address was added independently in each operation, we give credit to each operation. An example of this is the original source category "LUCA 1998 and Block Canvassing."

Therefore, the original source variable identifies the first operation or operations to add the address to the TEA in which it exists for the census, provided there is sufficient information to identify a TEA and an operation. For additional information on how this variable was defined, see the PRED TXE/2010 Memorandum Series: TXE/2010 MEMORANDUM SERIES: MAF-EXT-S-01, "Determining Original Source for the November 2000 Master Address File for Evaluation Purposes," March 5, 2001.

Due to the complicated design of the MAF, we had a limited ability to accurately determine the original source of every address.

2.9 Applying quality assurance procedures

We applied quality assurance procedures throughout the creation of this report. They encompassed how we determined evaluation methods, created specifications for project procedures and software, designed and reviewed computer systems, developed clerical and computer procedures, conducted field followup, analyzed data, and prepared this report.

3. LIMITS

3.1 We assumed the P-sample block was correct if we did not find the address in the MAF block

If a linked MAF address was not located in the MAF collection block during the field followup for this evaluation, we assumed that the address actually existed in the cluster in which it was listed during the A.C.E. Presumably, A.C.E. confirmed that these units existed in the sample cluster. However, it is possible that the unit existed elsewhere, that is, in neither the MAF or A.C.E. block. If that was the case, even though it is still a case of geocoding error on the MAF, it should not be part of our sample since our sample consists of only units that exist in the A.C.E. clusters. However, our field procedures did not allow us to make that distinction.

3.2 We assume the P-sample nonmatches are residential

We did not limit our matching to only residential MAF addresses. We allowed the analysts to link P-sample nonmatches to any address on the MAF that appeared to be a match. In FFU, we did not ask field representatives to make a determination about the residential status of the unit. Therefore, we assume that enough information was collected during the A.C.E. to determine that the address was a good, residential unit as of census day.

3.3 The rate of erroneously deleted units from the Unduplication operation may be overstated.

Near the end of the census, the Census Bureau conducted an Unduplication operation to remove duplicate addresses. See Section 1.5 for more details about this operation.

In this operation, if the Census Bureau believed it had a duplicate address (with a variation on how the address was presented), it had to decide which unit to retain. This decision was not based on specific knowledge of what address information was posted at the address. If the address listed in the A.C.E. looked like the one that was deleted during the Unduplication process and not like the one that was retained, the retained address may have been coded as nonexistent during the A.C.E. Furthermore, the independently-listed A.C.E. unit may have been coded as listed as missing from the census.

In this evaluation, we matched the A.C.E. "missing" units against all addresses in the MAF, including units deleted in the Unduplication operation. Therefore, it is possible that we could have matched an A.C.E. "missing" unit with the deleted unit in the duplicate pair from the Unduplication operation. From the matching in this evaluation, we would have concluded that a unit deleted in the Unduplication operation was erroneously deleted. When in fact a different version of that address was in the census and the decision to delete a potential duplicate was the correct decision. To the extent that this happened, the amount of erroneous deletions from the Unduplication operation as measured in this evaluation is overstated.

For example, if the census initially listed both "101 Main Street, Apt. A" and "101 Main Street, Basement," and one of these was determined to be a duplicate during the Unduplication operation, we may have arbitrarily deleted "101 Main Street, Basement." If the A.C.E. independently listed "101 Main Street, Basement, but did list "101 Main Street, Apt. A", the A.C.E. may have concluded that the "Basement" address was missing from the census and the "Apt. A" address was an erroneous enumeration. When this evaluation matched A.C.E. cases coded as "missing" against all units in the MAF, "101 Main Street, Basement" would have been included in our matching. In this case, if we matched to it, we would have concluded that the unit was left out of the census in error. In this specific situation, the reality is that we had the unit in the census, but neither the A.C.E. nor this evaluation recognized this.

3.4 We are limited in our ability to match non-city-style addresses

We were limited in our ability to match units with non-city-style address information, and therefore are likely to be missing instances of geocoding error and not identifying links to units on the MAF that were excluded from the census. This is especially difficult to do when matching outside of the sample block. Map spots and block codes are key identifiers of units with non-city-style addresses.

3.5 E-sample nonmatches were not checked for geocoding error

E-sample nonmatches that were coded as correct enumerations in the A.C.E. were only verified to exist in the A.C.E. cluster, not the block. In our evaluation, we checked the matches from the FHU operation for geocoding error within the A.C.E. cluster, but not E-sample correct enumerations. So, for the purpose of our estimation, we assume the E-sample correct enumerations are correctly geocoded to the MAF block, which would contribute to an underestimation of geocoding error.

3.6 The basic street address size variable was overstated

The variable showing the number of units at a basic street address (BSA) on the MAF included all addresses indicated as DMAF deliverable during the census process. Only a subset of these addresses remained in the census. Therefore, the size of BSA variable on the MAF is overstated relative to the size of BSA as of the end of the census.

Additionally, the size of BSA variable was only determined for units with city-style address information. Units with non-city-style addresses are considered single units even if they were part of a multi-unit structure.

3.7 We are unable to determine which Census 2000 operation provided the incorrect geocode

It would be interesting to look at how successful individual Census 2000 operations were at providing or confirming a correct geocode for the units included in the operation. This research could be done by examining the block code that was provided or confirmed for a given unit in an operation. Then we would look at the final geocoding error status of that unit, and be able to determine if a correct block was provided or confirmed during the operation. However, due to a problem with the block flag variable on the March 2001 MAF extracts, we are unable to determine whether the block provided by an operation was the same as the final block code for a unit.

3.8 We are unable to estimate the geocoding error associated with Group Quarters (GQs).

Our evaluation work was based on the A.C.E. sample. Since GQs were not included in the A.C.E. sample, we were not able to measure the accuracy of the geocodes associated with GQ addresses.

4. RESULTS

4.1 What is the total estimated percentage of census addresses geocoded to the incorrect Census 2000 collection block?

The Accuracy and Coverage Evaluation limited its matching of addresses to within the block cluster, or in some cases, to one ring of surrounding blocks. As previously mentioned, we were able to improve on the estimate of geocoding error in this evaluation by matching addresses in a larger geographic search area, and therefore finding more cases of units geocoded in error in the census. The estimated percentage of census addresses that were geocoded to the incorrect Census 2000 collection block is 4.8 percent (standard error is 0.3 percent).

4.2 Did the geocoding error estimate vary by TEA?

The geocoding error estimate varied among the different types of Census 2000 enumeration areas. Below, Table 1 presents the estimated percentage of housing units geocoded erroneously and their standard errors for the different TEAs.

Table 1. Census geocoding error estimates by TEA

TEA	% Geocoding error*
All TEAs	**4.80 (0.27)**
Mailout/Mailback	5.52 (0.33)
Update/Leave	1.71 (0.16)
List/Enumerate	1.15 (0.81)
Update/Enumerate (Rural & Urban)	1.70 (0.60)
Urban Update/Leave	11.59 (7.93)

* Percentages are weighted; standard errors in parentheses

As Table 1 shows, the estimated percentage of geocoding error in the census is significantly higher in Mailout/Mailback enumeration areas than in Update/Leave or List/Enumerate areas.

To some extent, we expect less geocoding error in Update/Leave and List/Enumerate areas because our address list was created on the ground through field operations, and therefore geocoding was based on first-hand field observation. This is different from Mailout/Mailback areas where geocoding was based on a combination of procedures, including an automated geocoding process. That combination could contribute to the higher geocoding error estimate in that enumeration area.

However, a lower geocoding error estimate in Update/Leave and List/Enumerate areas could also come about because those areas have a higher occurrence of non-city-style addresses. We were limited in our ability to match rural addresses, and were therefore unable to find as many cases of geocoding error in more rural areas as we are in Mailout/Mailback areas. A greater population of large multi-unit structures in Mailout/Mailback areas could also contribute to a higher geocoding error estimate in those areas (see Section 4.3).

Although the geocoding error estimate for Urban Update/Leave areas is high (11.6 percent), there is a high standard error associated with the estimate (7.9 percent). Therefore, the estimate for Urban Update/Leave areas is not statistically different from any other TEA.

4.3 Did the geocoding error estimate vary by size of structure?

When looking at geocoding error by single versus multi-unit structures, we are unable to compute structure-level statistics. No variable currently exists on the MAF to identify which housing units belong to a structure. Therefore, the estimates presented below are housing unit level estimates. To improve the MAF in the future, we recommend that information be added to the MAF to link units to structures.

The rate of geocoding error is higher for housing units in multi-unit structures than single units. Table 2 presents the estimated percentage of housing units geocoded erroneously and their standard errors for HUs in different size BSAs.

Table 2. Census geocoding error by size of BSA

BSA size	% Geocoding error*
All HUs	**4.80 (0.27)**
single unit	3.37 (0.14)
two units	3.17 (0.34)
small multi (3-9)	4.93 (0.69)
large multi (10+)	11.33 (1.37)

* Percentages are weighted; standard errors in parentheses

HUs in both small and large multi-unit structures have a significantly higher geocoding error estimate than single units or housing units in two-unit structures. Additionally, HUs in structures with 10 or more units have a significantly higher geocoding error estimate than HUs in structures with three to nine units.

We would expect geocoding error to be higher for units in multi-unit structures because geocoding error is a structure-based problem. Geocoding the structure to the wrong block causes every unit in that structure to be geocoded to the wrong block. The larger the structure is, the larger the number of geocoding error cases will be if the structure is geocoded to the incorrect block.

4.4 Did the geocoding error estimate vary by census region or Regional Office?

Geocoding error of census addresses is less frequent in certain regions of the country. Table 3 presents the estimated percentage of housing units geocoded erroneously and their standard errors for the different census regions.

Table 3. Census geocoding error by census region

Census Region	% Geocoding error*
All Regions	**4.80 (0.27)**
Northeast	4.42 (0.58)
Midwest	3.79 (0.35)
South	5.66 (0.55)
West	4.71 (0.55)

* Percentages are weighted; standard errors in parentheses

The Midwestern area of the country has a significantly lower geocoding error estimate than the South. There are no other significant differences.

Geocoding error estimates also differ for some of the Regional Offices (ROs). The Boston and Kansas City ROs both had a significantly lower geocoding error estimate than the national estimate of 4.8 percent. The geocoding error estimates for all 12 ROs are presented below in Table 4.

Table 4. Census geocoding error by RO

Regional Office	% Geocoding error*
All ROs	**4.80 (0.27)**
Boston	3.14 (0.47)
New York	5.13 (1.82)
Philadelphia	5.40 (0.62)
Detroit	4.03 (0.78)
Chicago	4.57 (0.61)
Kansas City	3.12 (0.50)
Seattle	4.82 (0.62)
Charlotte	5.75 (1.08)
Atlanta	7.03 (1.43)
Dallas	4.71 (1.05)
Denver	3.34 (0.64)
Los Angeles	5.37 (1.41)

* Percentages are weighted; standard errors in parentheses

The Atlanta, Charlotte, and Philadelphia ROs had significantly higher geocoding error estimates than the Boston and Kansas City ROs. Atlanta was also significantly higher than Denver. There are no other significant differences among the Regional Offices.

One might think that the differences between census regions and Regional Offices are driven by the differences we saw in geocoding errors between different structure sizes or TEAs. We attempted to analyze this, but found no definitive results.

4.5 Why are there addresses in the census, geocoded to an A.C.E. sample block, but coded as "missing" by the A.C.E.?[3]

About 8,900 of the units coded as "missing" by the A.C.E. were matched to units on the MAF during this evaluation. About 4,800 of them were matched to addresses that were included in Census 2000. Of those census matches, about 3,100 were geocoded in error in the census to a collection block that was different than the block provided by the A.C.E. The other 1,700 units were matched to census addresses that were geocoded to the same block as the A.C.E. "missing" addresses. There are two primary reasons that these census units were not included in the census address list used for the A.C.E. address matching.

The first reason is that some of these units were identified as potential duplicates during the Census 2000 Housing Unit Unduplication operations. Because the Census Bureau was unable to resolve the duplicate status of these units prior to the A.C.E. FHU matching, they were all excluded from the address list used for A.C.E. matching. Once the Census Bureau decided which units were not truly duplicates, they were reinstated. In this evaluation, we were able to use the final duplicate status of those units in our analysis. We found that about 78 percent of the units we were able to match to A.C.E. "missing" addresses in the same collection block were these reinstated duplicates.

The remaining 22 percent of the in-census matches to A.C.E. "missing" units in the same block were not reinstated duplicates. A reason that these units were excluded from the address list used for the A.C.E. address matching is that they were not geocoded to an A.C.E. sample block at the time of the FHU matching, but were moved into an A.C.E. sample block in time for our evaluation work. Examples of this include units that were moved to a different block following the Geographic Misallocation operation.

[3] Counts and percentages presented in section 4.5 are unweighted. For weighted estimates and associated errors, see Appendix A.

4.6 How many addresses, coded as "missing" from the census during the A.C.E., did we have on the MAF but exclude from the census?

Of the approximately 8,900 "missing" addresses from the A.C.E. that we matched to the MAF in this evaluation, about 4,000 were not included in Census 2000.[4] That is, these units were listed and confirmed as good, residential addresses during the A.C.E., but the Census Bureau's rules for creating the DMAF and the HCUF excluded them from the census. These units represent a weighted estimate of 1.3 million units coded as erroneously excluded from the census as measured by the A.C.E. and this evaluation. We examine the census exclusion processes in Table 5.

Table 5. Non-census matches by exclusion process

Delete Process	Block Matches[#] Percent*	Tract Matches[#] Percent*	Total Percent*
Never delivered to DMAF (MAF only)	26.41 (1.87)	44.54 (5.32)	28.26 (1.79)
Killed	49.96 (1.89)	43.99 (4.81)	49.35 (1.78)
Deleted as a result of HU status assignment	0.09 (0.05)	0.08 (0.08)	0.08 (0.05)
Deleted as a result of unclassified estimation	0.63 (0.15)	1.41 (0.78)	0.71 (0.16)
Confirmed Delete in the unduplication process	22.93 (1.23)	9.98 (1.90)	21.61 (1.11)
Total non-census matches	**100.02[+]**	**100.00**	**100.01[+]**

* Percentages are weighted; standard errors in parentheses

[+] Percentages may not sum to 100 due to rounding.

[#] Block matches=Geocoded correctly; Tract matches=Geocoding error

As can be seen from the table, there is a significantly higher percentage of tract matches that were never delivered to the DMAF (45 percent) than block matches that were never delivered to the DMAF (26 percent). As noted, tract matches imply geocoding error. The fact that a higher percentage of tract matches were never delivered to the DMAF implies that it is important to geocode a unit correctly from the start. Having the correct geocode is not only important for counting a census housing unit in the correct block, but it also impacts the likelihood that an address would be in the census enumeration process at all. If a unit was geocoded incorrectly in the MAF from the start of any Census 2000 operations, census field operations had to go out and correct the geography on that unit. This means that if an enumerator deleted a unit from one block, the unit had to be added in another block. If there was enumerator error and it was not added in another block, then its chances of being sent to the DMAF decreased.

[4] This count is unweighted; for weighted estimates and associated errors, see Appendix A.

We also see from Table 5 that units removed from the census because they were considered to be duplicates made up a larger percentage of the non-geocoding-error universe (23 percent) than of the geocoding-error universe (10 percent). This difference can be attributed to the fact that the Census Bureau is less likely to conclude that addresses are duplicates when they are in different blocks.

Table 5 also shows that units excluded by the Kill Process make up a little less than half of the universe of the non-census matches. That is, of the units on the MAF that were erroneously excluded from the census, half of them – about 653,000 units – were excluded from the census as a result of the Kill Process. However, as seen in Table 6, of all the units on the DMAF that were deleted as a result of the Kill Process, only eight percent of them were killed erroneously.

Table 6. Error rates of the census exclusion processes

Delete Process	Number of matches*	Total # HUs deleted by process	% HUs deleted in error*
Never delivered to DMAF (MAF only)	373,757	25,452,489	1.47 (0.12)
Killed	652,779	8,312,547	7.85 (0.41)
Deleted as a result of HU status assignment	1,116	22,352	4.99 (3.20)
Deleted as a result of unclassified estimation	9,348	46,196	20.24 (4.51)
Confirmed Delete in the unduplication process	285,793	1,371,111	20.84 (1.03)
Total non-census matches	**1,322,793**	**35,204,695**	**3.76 (0.14)**

* Counts and percentages are weighted; standard errors in parentheses

Both the unduplication process and unclassified estimation had about 20 percent error in what they deleted, which is a significantly higher error rate than the other processes. This suggests that these processes need to be refined so that valid units are not deleted in error. This is particularly true for the unduplication process which involved over one million housing units. Despite the limit to this estimate (see section 3.3), the number of cases potentially deleted in error in the unduplication process is high. The unduplication process was introduced very late in Census 2000 without sufficient planning. Work has already begun on building an unduplication process into the 2010 census.

4.6.1 Units never delivered to the DMAF

As shown previously in Table 5, units that were never delivered to the DMAF account for about 28 percent of all the units we found to be erroneously excluded from the census. There are a number of reasons units on the MAF would have never made it to the DMAF as a result of the Census Bureau's rules for developing the Census 2000 address frame. Those rules include our decision to exclude DSF and 1990 ACF addresses from census processing in outside-the-blueline TEAs, as well as the DMAF deliverability criteria discussed in section 1.4.

In Tables 7 and 8, we look at the original source of addresses that were never delivered to the DMAF and attempt to explain why these units were removed from the census process. Because there are different rules for the development of the Census 2000 address frame for different TEAs, we look at the inside-the-blueline TEAs separately from the outside-the-blueline TEAs.

Table 7. Inside-the-blueline MAF-only matches by original source

Original Source	Count*	Percent*
1990 ACF	63,139	33.92 (4.75)
DSFs	118,182	63.50 (4.81)
Dress Rehearsal	394	0.21 (0.21)
Be Counted/ Telephone Questionnaire Assistance	584	0.31 (0.23)
Non-ID adds/Unknown	3,822	2.05 (0.87)
Total	**186,122**	**99.99**[+]

* Counts and percentages are weighted; standard errors in parentheses

[+] Percentages may not sum to 100 due to rounding

Table 7 shows that about 64 percent of the inside-the-blueline MAF-only matches had a DSF as the original source of the address, and about 34 percent had the 1990 ACF as the original source.

Of the 63,139 units that came from the 1990 ACF, about 17 percent were identified as duplicates during the Block Canvassing operation, and were therefore excluded from the DMAF. About 60 percent had other negative actions from the Block Canvassing operation, including deletes, nonresidential units and uninhabitable units. Those units were not deliverable to the DMAF when they also were not indicated as residential on the September 1998 DSF. About 24 percent of the 1990 ACF units were not included in the Block Canvassing universe, which suggests they were not geocoded to a census collection block at the time of the operation. Those addresses were also kept off the DMAF.[5]

Of the 118,182 units with a DSF as the original source, about 63 percent were not indicated as residential on the November 1997 DSF or September 1998 DSF, and about 27 percent were indicated as residential on the November 1997 DSF, but as non-residential on the September 1998 file. The remaining units were either coded as duplicates during the Block Canvassing operation, or were not geocoded to a block in time to be included in the Block Canvassing universe.[4]

[5]The percentages presented in this paragraph are weighted. For the weighted counts and standard errors associated with the percentages discussed in this section, see Appendix B.

Table 8. Outside-the-blueline MAF-only matches by original source

Original Source	Count*	Percent*
1990 ACF	52,445	27.95 (4.79)
DSFs	125,361	66.81 (4.73)
Questionnaire Delivery	652	0.35 (0.35)
Address Listing	7,553	4.03 (1.08)
LUCA 99 Recanvass	232	0.12 (0.12)
Non-ID adds/Unknown	1,392	0.74 (0.40)
Total	**187,635**	**100.00**

* Counts and percentages are weighted; standard errors in parentheses

Table 8 shows that almost 95 percent of MAF-only records in outside-the-blueline TEAs came from the 1990 ACF and the DSFs. By design, the Census Bureau did not use addresses from those two sources in those TEAs. These units may be represented in the census by other forms of their address, however. During the creation of the address list in these areas, a different form of the address may have been provided, but the Census Bureau was unable to match it to the form of the address that came from the 1990 ACF or the DSF during census processing. However, we were able to match it to the A.C.E. address during this evaluation work. These addresses either represent units not captured in our outside-the-blueline operations or they reflect an inability of the A.C.E. to match other forms of addresses that were included in the census.

The remaining five percent of addresses were addresses intended for use in outside-the-blueline TEAs. Those addresses were deleted during the LUCA 99 Recanvass, did not have a map spot in the Topologically Integrated Geographic Encoding and Referencing System (TIGER) or did not have sufficient address information or location descriptions to be delivered to the DMAF.

4.6.2 Units that were removed by the Kill Process[6]

Of the non-census matches that were delivered to the DMAF, most of them (about 69 percent) were removed from the census by the Kill Process (see Appendix D). As discussed in Section 1.5, the Kill Process identified MAFIDs that most likely did not uniquely identify housing units as of census day. A total of about 8.3 million HUs were deleted by the process. In this evaluation, we estimate that 652,779 of those units were deleted in error. Table 9 below presents the characteristics of those units.

[6] The counts and percentages presented in section 4.6.2 are weighted. For the weighted counts and standard errors associated with the percentages discussed in this section, see Appendix C.

Table 9. Erroneously killed addresses by reason they were killed

Kill reason	Matches to kills*	Percent*
Double Delete, no mail return	43,728	6.70 (1.17)
Old DSF Address, no mail return	66,179	10.14 (1.05)
NRFU delete, not in CIFU, not a CIFU add, no mail return	320,773	49.14 (2.42)
Update/Enumerate delete, no mail return	2,072	0.32 (0.10)
NRFU delete, CIFU delete, no mail return	127,220	19.49 (1.65)
Not in NRFU, CIFU delete, no mail return	49,187	7.54 (1.39)
Fld. Ver. delete or duplicate, not a NRFU add, not a CIFU add	39,365	6.03 (0.74)
Usual Home Elsewhere (UHE) addresses[7]	4,254	0.65 (0.21)
Totals	**652,779**	**100.02**[+]

* Counts and percentages are weighted; standard errors in parentheses

[+] Percentages may not sum to 100 due to rounding.

As seen above in Table 9, about 49 percent (320,773) of the erroneously killed units met the following criteria:

- Respondent at the address did not mail in a census form
- Unit was deleted during NRFU
- Unit was not included in the CIFU universe of addresses
- Unit was not added in CIFU

In Table 10 we look at the breakdown and error rates associated with the kill reasons for the 320,773 units discussed above.

Table 10. Error rates for a specific kill reason by other census actions

Census Actions	# matches	% matches	ALL KILLS	Percent killed in error*
Undeliverable as Addressed (UAA)	217,762	67.89 (3.38)	2,453,235	8.88 (0.70)
Update/Leave (U/L) delete	25,903	8.08 (1.24)	498,132	5.20 (0.75)
Urban Update/Leave delete	617	0.19 (0.19)	10,683	5.78 (5.78)
Not a UAA or U/L or Urban U/L delete	76,491	23.85 (3.46)	281,757	27.15 (4.91)
ALL types of "NRFU delete, not in CIFU, not a CIFU add, no mail return"	**320,773**	**100.01**[+]	**3,243,807**	**9.89 (0.73)**

* Counts and percentages are weighted; standard errors in parentheses

[+] Percentages may not sum to 100 due to rounding.

[7] Adds from the July 7, 2000 update of the DMAF which were UHE addresses that were generated from Special Place/Group Quarters which were not allowed to provide a UHE address.

As can be seen from the table, when we did not receive any information about the unit from the post office, there was a significantly higher deletion error rate (27 percent) than when the post office identified the addresses as UAA (nine percent). The higher deletion error rate in that category is probably caused by the fact that only one operation deleted the unit. These cases represent a universe of NRFU deletes that were initially coded as completed cases from NRFU. However, when we completed data capture, we realized that these units should have been deletes. This is different than the other situations represented in Table 10 because the three other situations all provided a second confirmation that the unit should be deleted. Because of the planned introduction of mobile computing devices in the 2010 census, we should be able to avoid the situation where a NRFU questionnaire is allowed to be checked-in as complete when it has no data. Therefore, this class of errors from Census 2000 is not likely to be repeated in 2010.

In general, error rates for the kill process were low. As shown previously in Table 7, we estimate that approximately eight percent of all the killed units were deleted in error. The kill reason discussed above – NRFU delete, not in the CIFU universe, not a CIFU add and no mail return – had a significantly higher error rate than the overall rate (about 10 percent). Other than that difference, there were no individual kill reasons with an error rate significantly higher than the overall estimate of eight percent (see Table C-1 in Appendix C).

5. RECOMMENDATIONS

We are currently researching the possibility of collecting Global Positioning System (GPS) coordinates for addresses in the census. One reason for doing this is to help enumerators find their assignments. Another reason would be to ensure geocoding units to the correct block. If using GPS coordinates for improving geocoding is a high priority, our emphasis for the use of the GPS should not be entirely focused around rural areas. Instead, we should consider getting better geocoding for the areas with the highest geocoding error rates, which are inside-the-blueline areas.

We also recommend research to refine procedures for identifying and deleting units we believe to be duplicates. Despite the limit discussed in section 3.3, this evaluation found that the unduplication process may have deleted many units which should have been included in the census. However, the unduplication process was introduced very late in Census 2000 without sufficient planning. Work has already begun on building an unduplication process into the 2010 census.

This evaluation estimated that the unclassified estimation process may have had a high error rate and should be reexamined in the future as well. Research has already begun to minimize the number of units that require unclassified estimation and to improve the imputation methodology for when it is needed in the 2010 census.

Research could also be done to look at how successful individual operations were at providing or confirming a correct geocode for the units included in the operation. Most Census 2000 operations either provided a block code for a unit, or confirmed that a block was correct. We could use the results from this evaluation and block flag information from the MAF to determine if certain operations contributed more to geocoding error than others. However, this research might be limited by the fact that not all census operations concentrated on the placement of a unit in the correct block.

References

Barrett, Diane F., *The 1998 Master Address File (MAF) Quality Improvement Program (QIP)*, Bureau of the Census, June 4, 1999.

Childers, Danny R., *The 1990 Housing Unit Coverage Study*, 1990 Decennial Census Preliminary Research and Evaluation Memorandum No. 193, Bureau of the Census, October 21, 1992.

Childers, Danny R., *Accuracy and Coverage Evaluation: The Design Document*, DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, Revised, Bureau of the Census, January 26, 2001.

Hogan, Howard, *Specification of the Decennial Master Address File Deliverability Criteria for Census 2000*, DSSD Census 2000 Procedures and Operations Memorandum Series #D-1, Bureau of the Census, June 30, 1999.

Hogan, Howard, *Specifications for Eliminating Duplicate Records on the Hundred Percent Census Unedited File*, DSSD Census 2000 Procedures and Operations Memorandum Series #D-10, Bureau of the Census, November 7, 2000.

Hogan, Howard, *Specification for Reinstating Addresses Flagged as Deletes on the Hundred percent Census Unedited File*, DSSD Census 2000 Procedures and Operations Memorandum Series #D-11, Bureau of the Census, November 7, 2000.

Hogan, Howard, *Specification of the Kill Universe on the Decennial Master Address File for Census 2000*, DSSD Census 2000 Procedures and Operations Memorandum Series #D-13, Bureau of the Census, December 21, 2000.

Hogan, Howard, *Specifications for Assigning the Housing Units Status and Population Count on the Hundred-Percent Unedited File Prior to the Imputation for Unclassified Units*, DSSD Census 2000 Procedures and Operations Memorandum Series #D-14, Bureau of the Census, January 19, 2001.

Killion, Ruth Ann, *Quality of Geocodes Evaluation Computer Processing Specifications*, PRED TXE/2010 Memorandum Series, Chapter MAF-GC-S-04, Bureau of the Census, May 1, 2002.

Killion, Ruth Ann, *Before Followup Clerical Matching Specifications for the Quality of Geocodes Evaluation*, PRED TXE/2010 Memorandum Series, Chapter MAF-GC-S-08, Bureau of the Census, November 20, 2001.

Killion, Ruth Ann, *After Followup Clerical Matching Specifications for the Quality of Geocodes Evaluation*, PRED TXE/2010 Memorandum Series, Chapter MAF-GC-S-09, Bureau of the Census, March 5, 2002.

Kostanich, Donna, *Accuracy and Coverage Evaluation (ACE) Survey: Block Cluster Sample*

Selection Specification, DSSD Census 2000 Procedures and Operations Memorandum Series #R-3, Bureau of the Census, March 29, 1999.

Kostanich, Donna, *Census 2000 Overview of Unclassified Estimation*, DSSD Census 2000 Procedures and Operations Memorandum Series #Q-2, Bureau of the Census, August 19, 1999.

Kostanich, Donna, *Accuracy and Coverage Evaluation Survey: Computer Specifications for Housing Unit Dual System Estimation (U.S.) - Re-issue of Q-56*, DSSD Census 2000 Procedures and Operations Memorandum Series #Q-80, Bureau of the Census, June 25, 2002.

Miskura, Susan M., *Overview of the Duplicate Housing Unit Operations*, Census 2000 Informational Memorandum No. 78, Bureau of the Census, November 7, 2000.

Miskura, Susan M., *Results of Reinstatement Rules for the Housing Unit Duplication Operations*, Bureau of the Census, November 21, 2000.

Vitrano, Frank A., *Determining Original Source for the November 2000 Master Address File for Evaluation Purposes*, Planning, Research, and Evaluation Division TXE/2010 Memorandum Series: MAF-EXT-S-01, Bureau of the Census, March 5, 2001.

Vitrano, Frank A., *Determining Address Classification for Master Address File (MAF) Evaluation Purposes*, Planning, Research and Evaluation Division TXE/2010 Memorandum Series: MAF-EXT-D-01, Bureau of the Census, September 26, 2001.

Vitrano, Frank A., *Overview of the Quality of Geocodes Evaluation*, Planning Research, and Evaluation Division TXE/2010 Memorandum Series: MAF-GC-O-01-R1, Bureau of the Census, October 22, 2001.

Appendix A: Matches from this evaluation by census status and geocoding status

Table A-1. Matches by census status

Census Status	Count*	Percent*
In Census	1,583,543	54.49 (1.33)
Not in Census	1,322,793	45.51 (1.33)
On the DMAF, not in census	949,036	32.65 (1.14)
On the MAF *only*	373,757	12.86 (0.95)
Total matches	**2,906,335**	**100.00**

* Counts and percentages are weighted; standard errors in parentheses

Table A-2. In-census matches by geocoding error status

Geocoding error status	Count*	Percent*
Block match - no geocoding error	631,540	39.88 (1.82)
Tract match - geocoding error	952,003	60.12 (1.82)
All in-census matches	**1,583,543**	**100.00**

* Counts and percentages are weighted; standard errors in parentheses

Table A-3. In-census block matches by duplicate status

Duplicate status	Count*	Percent*
Reinstated duplicate	505,683	80.07 (2.45)
Not a reinstated duplicate	125,857	19.93 (2.45)
All in-census matches	**631,540**	**100.00**

* Counts and percentages are weighted; standard errors in parentheses

Appendix B: Inside-the-blueline MAF-only matches by DMAF exclusion reason

Table B-1. Inside-the-blueline MAF-only matches with a 1990 ACF original source

DMAF exclusion reason	Count*	Percent*
Block Canvassing duplicate	10,534	16.68 (7.31)
Negative action from Block Canvassing and not residential on Sept. 98 DSF	37,734	59.76 (8.39)
Not in Block Canvassing universe; not geocoded at the time of initial DMAF	14,871	23.55 (7.19)
Total	**63,139**	**99.99**[+]

* Counts and percentages are weighted; standard errors in parentheses

[+] Percentages may not sum to 100 due to rounding.

Table B-2. Inside-the-blueline MAF-only matches with a DSF original source

DMAF exclusion reason	Count*	Percent*
Not residential on Nov. 97 or Sept. 98 DSF	73,909	62.54 (5.89)
Residential on Nov. 97 and non-residential on Sept. 98 DSF	31,465	26.62 (4.60)
Block Canvassing duplicate	1,475	1.25 (0.76)
Not in Block Canvassing universe; not geocoded at the time of initial DMAF	11,334	9.59 (2.74)
Total	**118,183**	**100.00**

* Counts and percentages are weighted; standard errors in parentheses

Appendix C: Units killed in the Kill Process

Table C-1. Killed addresses by reason they were killed

Kill reason	# matches to kills*	% matches to kills*	ALL KILLS	Percent killed in error*
Double Delete, no mail return	43,728	6.70 (1.17)	1,630,860	2.68 (0.51)
Old DSF Address, no mail return	66,179	10.14 (1.05)	756,225	8.75 (0.88)
NRFU delete, not in CIFU, not a CIFU add, no mail return	320,773	49.14 (2.42)	3,243,807	9.89 (0.73)
Update/Enumerate delete, no mail return	2,072	0.32 (0.10)	46,617	4.45 (1.33)
NRFU delete, CIFU delete, no mail return	127,220	19.49 (1.65)	1,430,073	8.90 (0.84)
Not in NRFU, CIFU delete, no mail return	49,187	7.54 (1.39)	528,894	9.30 (1.86)
Field Verification delete or duplicate, not a NRFU add, not a CIFU add	39,365	6.03 (0.74)	427,173	9.22 (1.10)
UHE addresses[8]	4,254	0.65 (0.21)	248,898	1.71 (0.54)
Totals	**652,779**	**100.02**[+]	**8,312,547**	**7.85 (0.41)**

* Counts and percentages are weighted; standard errors in parentheses
[+] Percentages may not sum to 100 due to rounding.

[8] Adds from the July 7, 2000 update of the DMAF which were UHE addresses that were generated from Special Place/Group Quarters which were not allowed to provide a UHE address.

Appendix D: Non-census matches delivered to the DMAF by exclusion process

Table D-1. Non-census matches delivered to the DMAF by exclusion process

Exclusion Process	Count*	Percent*
Killed	652,779	68.78 (1.51)
Deleted as a result of HU status assignment	1,116	0.12 (0.08)
Deleted as a result of unclassified estimation	9,348	0.99 (0.22)
Confirmed Delete in the unduplication process	285,794	30.11 (1.50)
Total	**949,036**	**100.00**

* Counts and percentages are weighted; standard errors in parentheses