MASTER ADDRESS FILE: UPDATE METHODOLOGY AND QUALITY IMPROVEMENT PROGRAM*

Philip M. Gbur, Machell Kindred, and Michael L. Mersch, U.S. Bureau of the Census
Philip M. Gbur, U.S. Bureau of the Census, DSSD, Room 3763-3, Washington, DC 20233

Key Words: Coverage, Decennial Census, Census Test

I. INTRODUCTION

An accurate and complete address list is a critical ingredient in all U.S. Census Bureau surveys and censuses. The complex and costly nature of the process used to compile this list has led outside interests, as well as internal experts, to suggest that the Census Bureau create and maintain a national Master Address File (MAF) rather than prepare a new list for each program need. The initial goal of creating and updating the MAF is to meet the requirements for supporting the 1995 Census Test and 2000 Decennial Census activities.

The starting point for the MAF, in urban areas, will be the 1990 address list. The system for revising and updating the MAF will allow for four types of updates: 1) adding new and missed addresses, 2) correcting errors or incorporating changes to existing addresses, 3) deleting addresses, and 4) adding new information (such as telephone number).

The primary source for adding and correcting MAF addresses will be the United States Postal Service's (USPS) Delivery Sequence File. This is a periodically updated file maintained by the USPS. Information collected by local mail carriers, who identify changes on their routes such as new or destroyed units, is collected by the USPS in a central location where the Delivery Sequence File is maintained. The USPS file will be matched to the MAF on a periodic basis to obtain new addresses, correct existing ones, and identify potential nonexistent addresses.

A MAF Quality Improvement Program has been designed to identify geographic areas with housing unit coverage problems based on specific characteristics and to improve the overall MAF process. A statistical sample of areas from across the country will be selected to obtain data on characteristics associated with coverage problems. Selection of the areas will be based on criteria such as the percentage of "city-style" addresses, the percentage of nonmatches between the MAF and Delivery Sequence File, the number of multi-unit versus single unit structures, and similar data.

The following sections describe the creation of the MAF for the urban 1995 Census Test sites and associated evaluations; the general methodology planned for the MAF creation; MAF updates from the Delivery Sequence File; and the MAF Quality Improvement Program.

II. 1995 CENSUS TEST MAF

The Census Bureau is planning to conduct a 1995 Census Test in three urban sites: 1) Oakland, California; 2) Paterson, New Jersey; and 3) New Haven, Connecticut and a rural site consisting of six parishes in northwest Louisiana (DeSoto, Red River, Bienville, Jackson, Natchitoches, and Winn). The MAF process is being used to construct the address lists for the urban sites. For the rural site, the address list will be compiled through a prelist operation in which Census Bureau representatives will canvass the site and list all residential units. The remainder of this paper will focus on the urban areas. The address lists compiled and updated through the 1995 Census Test operations will become the MAF for the test site areas at the conclusion of the Census Test.

A. Creation

1. 1990 Address File - USPS File Match

The starting point for the urban address lists was the 1990 Decennial Census address list - the 1990 Address Control File. Reasons for starting with the 1990 address file were: 1) the Census Bureau already had it and 2) extensive operations were conducted to improve it during the 1990 Decennial Census. As a result, the 1990 file was arguably the best available list of U.S. residential addresses.

The addresses corresponding to the ZIP Codes in the test sites, and a surrounding ring of adjacent ZIP Codes, were extracted from the 1990 Address Control File. ZIP Codes were used since they were the only reasonably sized geographic identifier on both the census and postal files. The extract was computer matched, using a probability match, to a February 1994 vintage Delivery Sequence File received from the USPS for the same ZIP Codes. The USPS address file was selected because it: 1) provides a national coverage of addresses; 2) is expected to improve deliverability of mail pieces using the addresses on the MAF; and 3) builds on previous cooperation with the USPS in address list development.

After matching, the address file contained matched, Address Control File nonmatched, and Delivery Sequence File nonmatched addresses. This file was then computer geocoded (assigned geographic codes such as tract and block) using the Census Bureau's Topologically Integrated Geographic Encoding and Referencing (TIGER) System.

Table 1 provides the results of the computer matching and geocoding results. The computer match rate was slightly lower for Paterson than for the other two sites. The geocoding rate was slightly higher for the Oakland site. Additional analyses are underway to determine the reason for these differences.

The Census Bureau is assuming that the Delivery Sequence File address is the most mailable (current and deliverable) address for a given unit. Therefore, the USPS address was used whenever there was a difference between the Address Control File and USPS computer matched addresses.

2. Review

Geographers in the Census Bureau's regional offices reviewed the addresses which were not geocoded by TIGER. Locally available reference sources were used to determine the location and/or existence of the unit. Based on this review, the geographers made changes to TIGER such as the addition of a street name or expansion of an address range to allow the address to be geocoded using TIGER. Addresses for which changes could not be identified which would allow TIGER geocoding were kept with the lowest level of geography with which they could be identified. Table 2 presents the results of this review. After the review, the geocoding rate was similar across sites. Thus, the differential between Oakland and the other two sites seen after initial computer geocoding was eliminated as a result of the review.

After the geographical resolution operation was complete, nonmatched Address Control File and Delivery Sequence File addresses were reviewed. Statisticians at Census Bureau headquarters determined whether any additional matches could be identified. Table 3 presents the results of the manual matching review. For both 1990 and USPS address file origin addresses, the match rate for the reviewed addresses was higher for Paterson. Since the initial computer match rate was lower for Paterson than the other sites, the overall match rates were more similar across sites. The balance of the nonmatched addresses remained on the address files. The resulting files will be used as the initial address files for the 1995 Census Test urban sites.

B. Evaluation

Address coverage of the initial address files for the 1995 Census Test will be improved using several operations. The first two of these operations, and those expected to generate the most address updates (adds, corrections, moves, and deletes), are Precanvass and the Local Update of Census Addresses.

For Precanvass, enumerators will canvass the entire site and verify the address lists of the urban sites. Results from the Precanvass will be captured and incorporated into the address lists.

For the Local Update of Census Addresses, the address lists will be provided to local government officials for review and updating. The local officials will be sworn in by the Census Bureau and subject to the same restrictions as Census Bureau employees to safeguard the confidentiality of the address lists. The address lists will be the same lists used for the Precanvass operation in the urban sites. After the local officials return the lists, adds to the list will be verified by Census Bureau field personnel. Verified adds, as well as corrections, moves, and deletes will be incorporated into the address lists.

By evaluating the extent of the revisions to the address lists generated by these concurrent operations, a measure can be made of the completeness and accuracy of the initial address lists produced as a result of the MAF process. Updates from the operations will be analyzed by geography and address type to determine whether any systematic change to the MAF process could improve the address lists.

III. 2000 CENSUS MAF

A. Creation

1. 1990 Address File - USPS File Match and Clerical Review

The national MAF creation process will mirror the process described above for the 1995 Census Test. Again, the starting point will be the 1990 Address Control File which will be computer matched to a national USPS Delivery Sequence File. However, due to limited resources and the vast amount of work required for a national MAF, the clerical review will be completed in phases. The entire national MAF will be clerically reviewed over several years. Selected areas will be eligible for review in the first year. For the second and subsequent years, the current year(s) areas and the previous year(s) areas will be eligible for review. The areas to be reviewed the first year will be those which meet some selection criteria.

Professional geographers, supported by clerks, in the Census Bureau's regional offices will clerically review the addresses which were not geocoded by TIGER. Once again, locally available references will be used and the geographers will make any required changes to TIGER which will allow the addresses to geocode. After the clerical geocoding review operation is complete, the addresses will be geocoded by processing them through the revised TIGER system. After geocoding, clerks in the regional offices will review nonmatched Address Control File and Delivery Sequence File addresses to determine whether any additional matches can be identified. Addresses which can not be geocoded or matched will remain on the MAF.

2. Local Address Review

Updating and maintaining a complete and current representation of all geographic features across the country is a tremendous undertaking. To enhance the process and to allow local involvement, local jurisdictions and metropolitan planning organizations will be solicited to assist in the TIGER update process. Many of the addresses on the MAF may not geocode as a result of missing features or incomplete address ranges in TIGER. For areas agreeing to participate, the Census Bureau will provide a list of feature names missing from TIGER. The local officials will then enter the features' locations on Census Bureau supplied maps. In addition, the Census Bureau may send maps to be updated by the local areas. Local officials may then revise the maps by adding any missing feature(s) or by annotating incomplete address range(s). Any revisions by the local officials will be incorporated into TIGER allowing geocoding of any corresponding MAF addresses.

B. USPS File Updates

The USPS will periodically provide current versions of the Delivery Sequence File to the Census Bureau. The Census Bureau and the USPS are currently discussing the most beneficial frequency for providing the file, but it will not be anymore frequent than quarterly. Upon receipt, the entire updated Delivery Sequence File will be computer matched to the national MAF and geocoded. The clerical review, discussed above, will be based on the most recent version of the MAF. If possible, the Census Bureau (or USPS) will identify changes from the most recent version of the Delivery Sequence File and only these changed addresses will be matched to the MAF. This may require a much less intensive computer matching/geocoding effort.

C. Quality Improvement Program

1. Overview

To ensure continuous improvement of the MAF, various evaluation programs are planned to:

- Assess the coverage of the MAF
- Evaluate the quality and contribution of the various address sources for the MAF
- Measure the national quality of matching and geocoding the MAF
- Verify that matching rules and geocoding procedures are applied according to specifications and that they are as accurate as possible
- Obtain information which could be used for a proposed targeted precanvass

The MAF Quality Improvement Program is divided into three main areas: 1) independently designed test decks; 2) coverage evaluation; and 3) computer and clerical matching and geocoding rules evaluation. The quality assurance plans for the clerical activities are also a part of the Quality Improvement Program.

2. Independently Designed Test Decks

One or more independently designed test decks will be created to validate that the computer programs perform the matching and geocoding tasks as specified. This will be accomplished by comparing the results of the test deck matching and geocoding to predetermined results.

The independently designed test decks will be passed through the computer matching and geocoding software at various stages in the process: prior to the initial matching and geocoding process; after changes have been made to the software or hardware; and periodically, even if there have been no changes to the software or hardware. The last check is in case there have been changes to the file structure or operating environment.

3. Coverage Evaluation

a. Special Censuses

Between decennial censuses, local areas (such as cities, townships, or counties) may contract for the Census Bureau to conduct a special census for the area. The special censuses are usually requested in high growth areas where the cost of the census would be

offset by gains from the use of higher population counts in state or national funding allocations.

Special censuses provide an opportunity for evaluating the coverage of the MAF for those areas where special censuses have been conducted. The evaluation will consist of comparing housing unit block counts obtained from the special census to the corresponding MAF counts.

b. Census Tests

Census tests leading up to the 2000 Census will provide an opportunity for evaluating the coverage of the MAF for those areas where the tests are being conducted. This will provide measures of MAF quality for census test areas and also provide an opportunity to evaluate the MAF processes.

c. Area Review

The purpose of a coverage evaluation of the MAF is to determine to what extent the MAF identifies all addresses across the country. Addresses obtained during listing reviews for sampled areas and/or the Census Bureau's demographic surveys' field operations will be used to evaluate the national coverage of the MAF. The listings will be compared to the MAF to determine the level of coverage.

1) Area Selection

For the first year, selected areas will undergo geocoding and matching resolution. The areas to be reviewed will be selected by three digit ZIP Code since it is expected that the USPS address quality may vary across these areas. A stratified sample of blocks will be selected from the areas designated for the coverage evaluation. For subsequent years, the sample will be drawn from areas selected in the current year and a sub-sample of the areas selected any previous year.

The sampling will take into account various characteristics through sorting or stratification. The characteristics to be considered may include the following measures, as a count or rate:

a) MAF/USPS file count differential;
b) Nonmatches; and
c) Examination of other characteristics - such as address conversions, address type discrepancies (P.O. Boxes, rural routes), etc.

2) Independent Listing

Address lists will be generated from the MAF for the sampled areas. These lists will be field verified to identify any adds, deletes, and/or corrections. The field revisions will be captured and analyzed to determine if there are systematic (common) causes or specific (special) situations which cause missed or erroneous addresses to be in the MAF.

4. Geocoding and Matching Rules Evaluation

An evaluation sample of addresses will be selected to determine the validity of the matching and geocoding rules and to obtain a national estimate of the matching and geocoding quality of the MAF. The sample will be selected from: 1) computer matched and computer geocoded addresses; 2) addresses geocoded in geocoding resolution and matched in matching resolution; and 3) addresses geocoded in geocoding resolution but not matched in matching resolution. The addresses not geocoded in geocoding resolution and not matched in matching resolution will not be sampled, but the proportion of these cases will be measured.

The accuracy of the assigned geocodes and matching will be determined using the next "higher" level of geocoding and matching. That is, the computer matching will be evaluated using the matching resolution operation and the matching resolution will be evaluated in the field (ground truth). The computer geocoding will be evaluated using the geocoding resolution operation and the geocoding resolution will be evaluated in the field (ground truth).

5. Quality Assurance

The quality assurance operations for the geocoding and matching resolutions will provide continuous feedback to improve the matching and geocoding processes. The particulars of the quality assurance plan have not been determined.

IV. SUMMARY

The entire concept of the MAF is a fundamental change for the U.S. Census Bureau. The 1995 Census Test is the first opportunity to use the MAF in an application and to obtain information on the sources and processes used in MAF creation. A review of MAF processes for rural areas is currently planned for a 1996 census test.

The MAF has the potential to allow better address coverage at a reduced cost for the 2000 Decennial Census and ultimately, for all Census Bureau programs.

It is the first time an effort has been made to maintain a residential address list across time for use by various Census Bureau programs. In addition, the MAF process builds upon past cooperation with the USPS and brings it to a new level with the exchange of address information which will enhance both agencies' ability to perform their missions.

ACKNOWLEDGEMENTS

The MAF concept and development has occurred through much effort of many individuals at the U.S. Census Bureau. The authors wish to thank Ann Vacca, Bill Winkler, Dennis Stoudt, Brian Monaghan, Dan Harding, John Linebarger, and Florence Abramson for comments and suggestions on this paper.

*This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table 1: Results of Computer Geocoding and Matching for the 1995 Census Test Sites

Site	Number of Residential Addresses +	Match Rate (Percent)	Geocoding Rate (Percent)
New Haven, CT	137,031	95.2	90.9
Paterson, NJ	130,870	93.7	90.2
Oakland, CA	324,653	95.1	95.7

+ Note: This number includes matches, Address Control File nonmatches, and Delivery Sequence File nonmatches and is the denominator for the match and geocoding rates.

Table 2: Results of Geocoding Review by Test Site +

Results	New Haven, CT	Paterson, NJ	Oakland, CA
Base	137,031	130,870	324,653
Geocoded (%)	98.3	97.7	98.3
Ungeocoded (%)	0.2	0.2	0.9
Duplicate (%)	1.3	1.7	0.7
Nonexistent (%)	0.0	0.1	0.1
Out-of-Scope (%)	0.1	0.3	NA
Key Geographic Locator (%)	NA	NA	0.0

+ Note: Percentages may not add to 100 due to rounding.

Table 3: Results of Manual Matching Review by Address Source and Test Site

Address Source/ Result	New Haven, CT	Paterson, NJ	Oakland, CA
Address Control File - Base (addresses)	3891	4164	10789
Match (%)	13.4	24.2	1.7
Nonmatch (%)	86.1	75.0	98.2
Delete (%)	0.5	0.8	0.1
Delivery Sequence File - Base (addresses)	4927	5369	11407
Match (%)	10.5	18.8	1.6
Nonmatch (%)	87.9	80.8	98.3
Delete (%)	1.6	0.4	0.1