ASSESSING THE QUALITY OF THE CENSUS 2000 MASTER ADDRESS FILE USING EARLIER EVALUATION DATA

Joseph Burcham, U.S. Census Bureau
Washington, D.C. 20233

Key Words: Block Canvassing, Quality Improvement Program, Master Address File, coverage, geocoding, Delivery Sequence File, Local Update of Census Addresses

Introduction

The goal of this paper is to describe early work in our attempt to evaluate the usefulness of three major address sources in improving coverage and geocoding on the Master Address File (MAF) based on earlier evaluation results.

The MAF is a file of residential addresses that the U.S. Census Bureau is maintaining. The MAF is a source for the Decennial MAF (DMAF), which the Census Bureau used to conduct Census 2000. The MAF will also be maintained as a sampling frame throughout the next decade.

Geocoding is the assignment of the assignment of a residential address on the MAF to a census block. It is important for all addresses to be geocoded because the Census Bureau must have the ability to geographically locate each address.

The Census Bureau conducted a study called the 1998 MAF Quality Improvement Program (QIP) to measure the coverage and coding errors on the Initial MAF as of April 1, 1998 within Census 2000 mailout/mailback enumeration areas. The Initial MAF was composed of the 1990 Address Control File (ACF) and the November 1997 Delivery Sequence File (DSF) from the U.S. Postal Service. Mailout/mailback enumeration areas are areas in which predominantly city-style addresses (i.e. house number and street name) are used for mail delivery.

In the QIP study, one of the estimates we produced was called 'undercoverage,' which was an estimate of existing addresses that appeared to be missing from the MAF.

Undercoverage from 1998 QIP:

Weighted Estimate	Standard Error
9.08%	0.77%

This entire paper focuses on analysis of these addresses that appeared to be missing from the Initial MAF but that we have now located on the MAF.

All of the analysis explained is exploratory and is based on using 1998 QIP data to evaluate the MAF in 2000.

Background

Several different address sources were used to update the Initial MAF before March 2000. However, this evaluation focuses on three major operations: the September 1998 DSF, the Block Canvassing operation, and the 1998 Local Update of Census Addresses (LUCA).

The DSF is a file of residential and non-residential addresses representing mail delivery points. The U.S. Postal Service provides updated versions of the DSF to the Census Bureau on a regular basis. Because the DSF only includes mail delivery points (does not necessarily have a record for every housing unit), and because there is a time lag of units getting onto the DSF, this file must be used in conjunction with other sources to meet the needs of the Census Bureau.

Block Canvassing consisted of field representatives canvassing 100% of the mailout/mailback areas. The field representatives took address lists with them and were required to update the lists based on situations they observed on the ground.

The 1998 LUCA program made it possible for local tribal governments to participate in the development of the

---

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

MAF. The Census Bureau delivered address lists to participating governments with a requirement that the governments maintain confidentiality of the addresses. The governments updated the lists, and then field representatives verified the suggested updates. Whenever the Census Bureau did not accept a suggested address list change from a LUCA participant, the participant could appeal to an independent review body for resolution.

The MAF is linked to the Topologically Integrated Geographic Encoding and Referencing System (TIGER). This system allows for automatic assignment of most city-style addresses to census blocks. Some addresses on the Initial MAF were ungeocoded (missing block codes) or were geocoded in error. The ungeocoded addresses were delivered to a clerical operation called the MAF Geocoding Office Resolution (MAFGOR) for the assignment of geocodes. MAFGOR may have incidentally corrected the geocodes of addresses that were geocoded in error, but this is not a specific purpose of the MAFGOR operation.

Methodology

The 1998 QIP study was based on a two-stage nationally representative sample within the mailout/mailback areas. The first stage was a sample of counties; the second stage, a sample of blocks.

In 1998 field representatives traveled to the sample blocks and created a listing of all existing residential addresses. This listing was matched to the November 1997 MAF to identify deficiencies that existed on the Initial MAF. Specifically, this matching process consisted of first a computer match. Cases that did not get resolved by the computer match were sent to clerical matching and/or field followup for resolution. (See Burcham, Joseph and Diane Barrett (1999), "Assessing the Quality of the Initial Master Address File for Census 2000" for more details).

A new match was performed for this study. This new match was between the current MAF and addresses that appeared to be missing from the November 1997 MAF. This match occurred from February 25 to March 6, 2000.

Results

All estimates in this presentation are based on existing addresses that appeared to be missing on the Initial MAF but that we have since located on the MAF.

Geocoding Statistics

Universe #1:

Table 1 shows the percentage of addresses once missing from the MAF but now on the MAF that were ungeocoded as of March 2000.

Table 1. Percentage Ungeocoded

Weighted Estimate	Standard Error	Sample Size
1.07%	0.56%	7855

The specific universe in Table 1 includes addresses that:

- were not found on the MAF during QIP
- were coded to a block by QIP, and
- were matched to addresses on the MAF that were added by at least one operation between the development of the Initial MAF and March 2000

Universe #2:

Table 2 shows the percentage of geocoded, MAF addresses that were geocoded to the QIP block in March, 2000.

Table 2. Percentage Geocoded to QIP Block

Weighted Estimate	Standard Error	Sample Size
87.1%	2.64%	7729

This statistic is of interest because if an address currently on the MAF is geocoded to the same block that the QIP operation geocoded it to, we have more confidence that the address is geocoded to the correct block.

For the 13% of addresses that have not been geocoded to the QIP block, we will not have enough evidence, without conducting additional field work, to determine the correct block.

The specific universe in Table 2 includes addresses that:

- were not found on the MAF during QIP,
- were coded to a block by QIP, and
- were matched to addresses on the MAF that were added and geocoded by at least one operation between the development of the Initial MAF and March 2000.

Universe #3:

The rest of the results apply to the individual address sources.

All addresses from the QIP study were geocoded to mailout/mailback areas. It is possible that some addresses originally classified as mailout/mailback have since been moved to other enumeration areas. Because the address sources that we are examining only occurred inside mailout/mailback areas, we limit the analysis of these sources to addresses that remained in the mailout/mailback areas.

The following table shows a breakdown of TIGER block code agreement for addresses that have been geocoded to the QIP block.

Table 3. TIGER Block Code Agreement

Characteristic	Weighted Estimate	Standard Error
Same as QIP block	91.74%	1.93%
Previously same as QIP block (now a different block)	2.50%	1.05%
Different than QIP block	5.19%	2.06%
No block	0.57%	0.24%

Sample size = 5063

We estimate that TIGER geocoded about 92 percent of addresses in the universe to the QIP block. It geocoded around eight percent to a different block than the QIP block. It did not provide a block code for less than one percent.

Table 4 shows the extent that Block Canvassing agreed with the QIP block code.

Table 4. Block Canvassing Block Code Agreement

Characteristic	Weighted Estimate	Standard Error
Same as QIP block	88.62%	3.38%
Different than QIP block	2.96%	1.62%
No block	8.41%	2.63%

Sample Size = 5063

We estimate that Block Canvassing geocoded addresses in the universe to the QIP block almost 89 percent of the time. Block Canvassing did not geocode about 8.5 percent of the addresses.

Note that the estimate of "no block code" is limited by the fact that Block Canvassing may have provided addresses in a different form than the form we searched for when matching. In other words, an address in the 8.5% group may have appeared to be missing from Block Canvassing but Block Canvassing provided a duplicate address referring to the same unit.

The specific universe for tables 3 and 4 included addresses that:

- were not found on the MAF during QIP,
- were coded to a block by QIP,
- were matched to addresses on the MAF that were added and geocoded by at least one operation between the development of the Initial MAF and March 2000,
- have a current block code equal to the block code assigned by QIP, and
- are currently inside of mailout/mailback areas on the MAF

Universe #4:

The next table is also based on addresses that have been geocoded to the QIP block on the MAF. It is also limited to areas that participated in 1998 LUCA. The table shows the percentage of addresses receiving a block code from 1998 LUCA that were geocoded to the QIP block by LUCA.

Table 5. 1998 LUCA Block Code Agreement

Weighted Estimate	Standard Error	Sample Size
88.73%	4.35%	732

The specific universe for table 5 included addresses that:

- were not found on the MAF during QIP,
- were coded to a block by QIP,
- were matched to addresses on the MAF that were added and geocoded by at least one operation between the development of the Initial MAF and March 2000,
- have a current block code equal to the block code assigned by QIP, and
- are currently inside of mailout/mailback areas on the MAF

Coverage Statistics

Universe #5 (same as universe #3):

Table 6 below shows the percentage of addresses geocoded to the QIP block on the MAF that existed on the Initial MAF:

Table 6. Addresses on the Initial MAF

Weighted Estimate	Standard Error	Sample Size
22.65%	4.24%	5063

All addresses we examined for this study appeared to be missing from the Initial MAF. However, we estimate that almost 23 percent of addresses in the universe were, in fact, on the Initial MAF. This finding confirms a limitation from the QIP study: that we could not locate all existing addresses on the Initial MAF, especially if they were ungeocoded or geocoded in error at the time of QIP. We have evidence that the nine percent undercoverage estimate from QIP was overstated.

Table 7 below shows the breakdown of the September 1998 DSF flag and is based on the universe of interest.

Table 7. 9/98 DSF Coverage Statistics

Characteristic	Weighted Estimate	Standard Error
Not on the DSF	29.77%	4.79%
Residential on DSF	69.76%	4.88%
Non-res. On DSF	0.46%	0.16%

Sample Size = 5063

We estimate that about 70 percent of units in the universe were on the DSF as residential units and almost 30 percent were not on the DSF. The estimate for "not on DSF" is subject to the duplicate limitation discussed under universe #3. Another point worth mentioning is the fact that the DSF has a record for every delivery point and not necessarily for every housing unit. Some housing units considered "missing" from the DSF may be in multi-unit structures that are accounted for by delivery point records.

Non-residential coding is less than one half of a percent. This error does not appear to be a major problem on the DSF.

The next table shows a breakdown of all actions that Block Canvassing took on addresses on the universe of interest.

Table 8. Block Canvassing Coverage Statistics

Characteristic	Weighted Estimate	Standard Error
Added	31.85%	4.45%
Verified, Corrected, or Moved	58.02%	6.60%
Deleted or changed to Non-residential	1.91%	1.26%
No action	8.22%	2.57%

Sample Size = 5063

Block Canvassing added or verified the existence of close to 90 percent of addresses in the universe. The operation deleted (or changed to non-residential) about two percent of addresses. It provided no action for about eight percent of the addresses. Block Canvassing is supposed to provide an action for all addresses sent to be verified. So, we suspect that this eight percent estimate represented addresses that were not sent to Block Canvassing to be verified. This estimate is also subject to the duplicate limitation.

Another limitation with the Block Canvassing operation is that sometimes when an address received no action from Block Canvassing then it received a default Block Canvassing code of "verify." We do not know how often this occurred. This default code is unfortunate for evaluation purposes because we cannot distinguish between true verifications and defaulted verifications.

Table 9 shows estimates of discrepancies between sources in providing addresses to the MAF. Specifically, it shows the different combinations where one particular source provided an address that the other sources missed. A "+" on the graph means the particular source recognized an existing unit as "existing." A "-" on the graph means the source did not recognize an existing unit as "existing." The universe of addresses used in Table 8 applies here also.

Table 9. Discrepancies Between Sources

Characteristic	Weighted Estimate	Standard Error
LUCA + / BC -	4.26%	1.53%
LUCA + / DSF -	7.66%	2.11%
LUCA + / BC - / DSF -	4.00%	1.43%
BC + / DSF -	22.64%	3.22%
DSF + / BC -	2.53%	1.34%

Sample Size = 5063

As shown in the table, LUCA recognized about four percent of addresses as existing that Block Canvassing did not recognize. It recognized about eight percent as existing that the DSF did not. And, it recognized about four percent as existing that were not recognized by Block Canvassing or the DSF. The DSF recognized about 2.5 percent that Block Canvassing did not.

Block Canvassing recognized about 23 percent of addresses as existing that the DSF did not recognize as existing. In other words, if the DSF was used but Block Canvassing was not used, this 23 percent of addresses in the universe might have been lost. This estimate could be as high as it is partly because of the duplicate limitation, because addresses provided by Block Canvassing which appear to be missing from the DSF could really be in a different form on the DSF. If most of these addresses were not on the DSF as individual units, this would confirm the need for conducting the Block Canvassing operation to update the MAF.

The specific universe for tables 6-9 included addresses that:

- were not found on the MAF during QIP,
- were coded to a block by QIP,
- were matched to addresses on the MAF that were added and geocoded by at least one operation between the development of the Initial MAF and March 2000,
- have a current block code equal to the block code assigned by QIP, and
- are currently inside of mailout/mailback areas on the MAF

Future Research

As mentioned earlier in the paper, in this study we only focused on addresses that appeared to be missing from the Initial MAF and have since been located on the MAF. Our original goal was to determine the percentage of addresses originally missing that still have not been added. However, because of a matching limitation, we cannot accomplish the original goal at this time.

When examining results from the match between QIP addresses and the MAF, we began to suspect that a lot of addresses were not matching because of slight misspellings. This non-matching could lead us to conclude that a lot of addresses are missing from the MAF when they actually are on the MAF.

In the future, we will be researching ways to accomplish the original goal.

This paper focused only on the "undercoverage" estimate from the QIP study. But, there were four other estimates we computed in the QIP study. In the future, we hope to conduct additional analysis on all four of these.

1. *Overcoverage* - these were MAF addresses coded to the sample blocks that did not belong in the sample blocks. We will examine whether or not these addresses eventually were deleted or moved to different blocks on the MAF.

2. *Geocoding errors* - these were MAF addresses geocoded in error. We will examine whether or not these addresses eventually were geocoded correctly.

3. *Ungeocoded errors* - these were existing MAF addresses that were ungeocoded. We will examine whether or not these addresses eventually were geocoded, and if they were geocoded correctly.

4. *Non-residential coding errors* - these were existing MAF addresses that were incorrectly coded "non-residential." We will examine whether or not these addresses eventually received a corrected status of "residential."

Conclusions

Based on the information we have, it appears that geocoding error is a bigger problem than cases not receiving geocodes. As stated in the results section, only one percent of addresses were ungeocoded but of the geocoded ones, 13 percent were geocoded in error.

No individual source accounted for correctly geocoding

more than 92 percent of correctly geocoded addresses on the MAF. Also, each source provided at least 2.5 percent of addresses that any other individual source missed. Based on the results, each source seems to have made improvements above and beyond what the other sources accomplished.

According to the results, the Block Canvassing operation added more addresses to the Initial MAF than any other updating source. This results provides good evidence that the Block Canvassing operation was needed for MAF development.

The MAF will be maintained as a sampling frame after Census 2000. Based on the results from this evaluation, it may be reasonable to assume that more than one source of addresses will be necessary to keep the MAF as accurate as possible throughout the next decade.

References

Barrett, Diane F. (1999), "The 1998 Master Address File Quality Improvement Program," internal memorandum for Robert Marx, Bureau of the Census.

Burcham, Joseph and Diane Barrett (1999), "Assessing the Quality of the Initial Master Address File for Census 2000," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Bureau of the Census (1999), "Program Master Plan: Census 2000 Block Canvassing Operation," internal memorandum.

Pennington, Robin A., et al. (2000), "A Preliminary Examination of the Address List for Census 2000," Proceedings of the Section on Survey Research Methods, American Statistical Association.