ASSESSING THE QUALITY OF THE INITIAL MASTER ADDRESS FILE FOR CENSUS 2000

Joseph Burcham, Diane Barrett, U.S. Census Bureau, Planning, Research, and Evaluation Division
Joseph Burcham, U.S. Census Bureau, Room BH118-2, Washington, D.C. 20233

Key Words: QIP, MAF, coverage errors, geocoding errors

Introduction

The Master Address File, or MAF, is a file of residential addresses that the Census Bureau is maintaining. The MAF is a source for the Decennial MAF (DMAF), which the Bureau will use to conduct Census 2000. The MAF will also be maintained as a sampling frame after Census 2000.

As of August 1999, the Bureau has used four different sources of addresses to update the MAF in areas where the Census Bureau will use mailout/mailback enumeration in Census 2000. The four sources are:

- The 1990 Address Control File (ACF)
- The November 1997 Delivery Sequence File (DSF) from the U.S. Postal Service
- The September 1998 DSF
- The address files from the Block Canvassing operation

The ACF is the file of addresses collected during the 1990 Census. The DSF is maintained by the U.S. Postal Service and contains more up-to-date information about residential and non-residential addresses that receive mail. In the Block Canvassing operation, field representatives traveled to the mailout/mailback areas to provide additional updates to the MAF.

The Initial MAF, which is the version of the MAF we evaluated in this study, consisted of the ACF and the November 1997 DSF.

Because the Census Bureau must have the ability to geographically locate each address, each address on the MAF is assigned, or geocoded, to a census block.

Goal of the Evaluation

Census Bureau staff designed the 1998 Quality Improvement Program (QIP) to measure the effectiveness of the Initial MAF in accurately reflecting existing housing units as of April 1, 1998.

The MAF will eventually become the single source of addresses that the Bureau will use to conduct Census 2000 as well as other surveys. By evaluating an early version of the file, we can determine the impact of the two initial sources to the file. By understanding the impact of these two sources, we can get an indication of the amount of coverage improvement needed for Census 2000 and other surveys.

We accomplished the goal by producing national level and census division level ratio estimates of coverage errors and coding errors on the MAF.

We had data suggesting that the quality of the DSF varied among postal service areas. During early stages in planning, we explored the possibility of producing estimates for each postal service area. Due to the fact that postal service areas cross county boundaries, we were unable to assign our sample counties to a specific postal service area. Therefore, we produced estimates for the level of census geography that is closest to postal service area, which is census division.

The five errors that we were specifically interested in measuring were:

- *Undercoverage error* - an existing residential address is missing from the MAF
- *Overcoverage error* - a non-existing residential address is included on the MAF
- *Geocoding error* - an existing residential

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

address is coded to the wrong census block on the MAF

- *Ungeocodable error* - an existing residential address is on the MAF, but not coded to a census block at all

- *Non-residential coding error* - an existing residential address is incorrectly coded "non-residential" on the MAF

The 1998 QIP Operation

QIP methodology is modeled after the Census Bureau's 1996 Integrated Coverage Measurement (ICM) methodology. ICM is currently known as the Accuracy and Coverage Evaluation, and was designed to measure the coverage of people and housing units in the Census. To test the operational feasibility of using the ICM methodology for QIP, a pilot study was conducted in 1997 in six counties. With a few modifications, the methodology proved to be effective in measuring the coverage of housing units on the MAF.

The 1998 QIP operation consisted of the following steps:

- selecting a stratified, two stage cluster sample of areas to be used in the study (the first stage being a sample of counties and the second stage being a sample of blocks)

- creating the Independent Listing (IL), which was a current list of all housing units existing in the blocks.

- Matching the IL to the MAF (to evaluate the MAF)

- computing estimates of MAF errors using the match codes

- computing standard errors using stratified jacknife replication

Sample Design

We designed our sample to give us coefficients of variation of about 10 to 15 percent.

The universe consisted of all counties that contained areas classified as mailout/mailback enumeration areas.

Mailout/mailback enumeration areas are areas consisting of households that will receive their Census 2000 forms in the mail. These areas consist of primarily city-style addresses.

To stratify the universe of counties, we first grouped the counties by Census Division. Within each Division, we assigned counties in the universe to one of four "growth" cells:

- Low/Low
- Low/High
- High/Low
- High/High

We created these cells by comparing the Census Bureau's housing unit growth estimates (as projected by the Bureau's Population Division) to housing unit growth estimates on the DSF. Sometimes the estimates agreed and sometimes they did not. We set up the four cells to reflect the level of agreement. For example, the low/low cell identifies low housing unit growth on the DSF and low housing unit growth according to the Census Bureau, etc.

There are nine census divisions. Nine divisions times four "growth" cells resulted in 36 "growth" strata nationwide.

Within each stratum, we selected three counties systematically proportional to the size of the county. Selecting three counties in each stratum resulted in 108 counties.

We required about 170,000 residential addresses in sample. We allocated that sample size to the counties in order to achieve a self-weighting design within each "growth" stratum.

Within each sample county, we selected a sample of blocks. The blocks were grouped into four "size" strata. A block is assigned to one of these strata based on the estimated number of units in the block. The four size strata were:

- 0-2 HUs
- 3-19 HUs
- 20-79 HUs
- 80+ HUs

We allocated the county sample size proportional to the count of HUs in the bottom three size strata. Then, in each of the three strata, we selected the required number of blocks with equal probability.

For the 0-2 strata, we wanted to select as few of these

blocks as possible but also minimize the potential impact on variance of high housing unit growth in these blocks. So, we selected twelve blocks with equal probability in each of these strata.

The Independent Listing

To create the IL, field representatives traveled to the blocks and listed all residential units that existed on April 1, 1998. It is assumed that this listing is more current than the MAF. So, we evaluated the MAF by comparing, or matching, the IL to the MAF.

Matching

In matching, whenever a residential IL address in a particular block matched to a residential MAF address coded to the same block, we were confident that the MAF address represented an existing housing unit.

Whenever an address on one list did not match an address on the other list, or whenever two addresses matched but the MAF address was coded in error, this identified one of the errors we were measuring.

The first type of matching we did was computer matching. All IL addresses were residential and geocoded to one of the sample blocks. But on the MAF, we matched to addresses regardless of whether they were residential or non-residential or geocoded or ungeocoded. Also, on the MAF we matched to addresses coded to the sample block but also to addresses in the zip codes surrounding the sample blocks. The point of matching to a larger number of addresses on the MAF is that we have a better chance of identifying MAF coding errors.

Because the computer match was not perfect, we had several followup operations for the purpose of finding more matches and verifying the existence of units.

These followup operations were:

- *Before Followup Review* - which was clerical matching that occurred right after the computer match
- *Field Followup* - where field representatives traveled back to the sample blocks for the cases that were still unresolved
- *After Followup Review* - where the clerical matchers assigned final match codes

In field followup, we required field representatives to determine whether or not housing units existed for the addresses in question. There were some situations where we had an address that corresponded to an existing unit, but the address was incorrect. In these situations, some field representatives answered "the address does not exist," which was later interpreted as "the unit does not exist" in After Followup Review.

In future studies of this nature, we should allow corrections to addresses on the followup forms.

Estimation

We used the final match codes to produce ratio estimates for each of the five MAF errors that we were interested in.

When final match codes were being assigned, some addresses were still unresolved. An address could be unresolved because:

- It is not known if the address refers to an existing residential unit
- The correct block is unknown

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

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

For the most part, each worst-case scenario estimate was worse than its corresponding no-unresolved estimate by about half of a percentage point. We decided to be conservative and present only the worst-case scenario estimates in our reports.

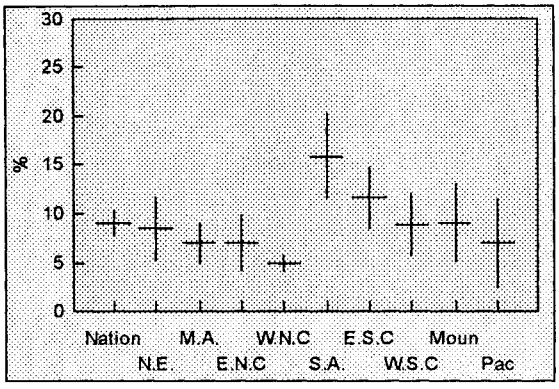
The Undercoverage Estimate

Undercoverage simply means a unit was missing from the MAF but we found it on the ground.

Specifically, this estimate is defined as:
The percentage of existing housing units in the sample blocks that are missing from the MAF.

Figure 1 shows 90% confidence intervals for the national and census division level undercoverage estimates.

Figure 1. Undercoverage



The national undercoverage estimate (on the far left) is 9.1%. The confidence interval ranges from 7.8% to 10.3%.

The census divisions are abbreviated on the graph. The abbreviations of the divisions, with their names, are:

N.E. - New England
M.A. - Middle Atlantic
E.N.C. - East North Central
W.N.C. - West North Central
S.A. - South Atlantic
E.S.C. - East South Central
W.S.C. - West South Central
Moun - Mountain
Pac - Pacific

The undercoverage estimate ranges from about 5% in the West North Central division to about 16% in the South Atlantic division. The undercoverage rate in the South Atlantic division is significantly higher than the undercoverage rates in four other divisions.

In general, southern areas of the United States tend to display higher undercoverage than northern areas.

As stated before, all estimates that we present are worst-case scenario estimates. Most of the worst-case scenario estimates are worse than their corresponding no-unresolved estimates by only about half of a percentage point. The South Atlantic division is an exception. The undercoverage estimate in this division ranges from a no-unresolved estimate of 14.8% to a worst-case estimate of 15.9%.

The Overcoverage Estimate

Overcoverage means a unit was on the MAF but we did not find it on the ground.

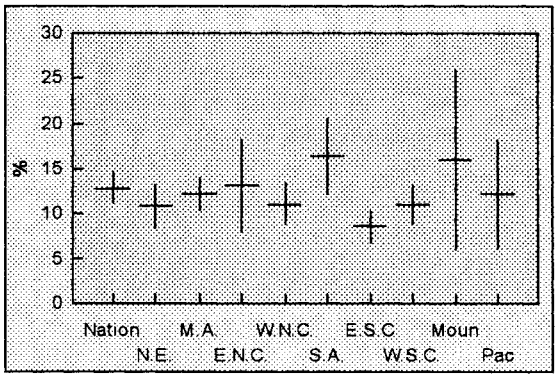
Specifically:
This estimate is the percentage of MAF addresses coded to the sample blocks that should not be coded to the sample blocks.

An overcoverage addresses could be:

- an address that refers to a housing unit that exists outside of the sample blocks
- an address that does not refer to an existing housing unit at all, or
- a duplicate of another residential MAF address

Figure 2 shows 90% confidence intervals for overcoverage.

Figure 2. Overcoverage



The national overcoverage estimate is 12.8%, with a confidence interval ranging from 11.1% to 14.6%.

Overcoverage ranges from about 8.5% in the East South Central division to about 16% in the South Atlantic division. These two rates are the only rates that are significantly different.

The Geocoding Error Estimate

Geocoding error means we found the unit on the ground, but it was coded to the wrong block on the MAF.

Specifically, this estimate is the percentage of MAF housing units existing in the sample blocks that are geocoded in error.

When conducting a study of geocoding error based on sample blocks, there are different types of geocoding errors to consider:

- *Geocoding error of exclusion* - a housing unit exists inside a sample block but is incorrectly excluded from the sample block on the MAF (it is coded to the wrong block on the MAF)

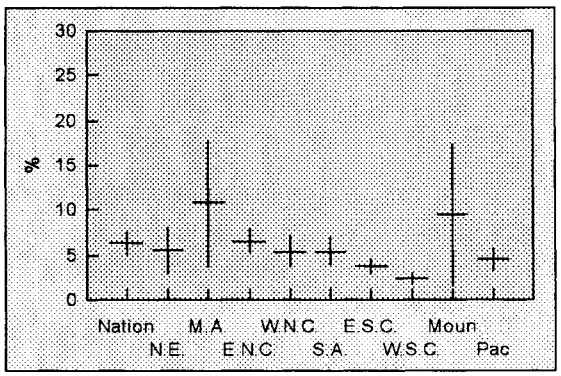
- *Geocoding error of inclusion* - a housing unit is incorrectly included in a sample block on the MAF, but exists outside of the sample block (it is coded to the wrong block on the MAF)

Another factor to consider when studying geocoding errors is the area in which one searches for them.

Because of limited resources, the only type of geocoding error we measured was geocoding error of exclusion. During the computer match, we searched for these errors within the sample blocks but also within zip code on the MAF. These types of geocoding errors are the only types reflected in our geocoding error estimate.

Figure 3 shows 90% confidence intervals for geocoding error.

Figure 3. Geocoding Error



Our national estimate of geocoding error is 6.2%, with a confidence interval ranging from 4.9% to 7.5%

Geocoding error ranges from about 2.5% in the West South Central division to about 11% in the Middle Atlantic division. The West South Central estimate is significantly lower than the estimate in five other divisions.

The geocoding error estimate in the South Atlantic division is also an exception to the 0.5% rule. Geocoding error in this division ranges from a no-unresolved estimate of 4.7% to a worst-case estimate of 5.4%.

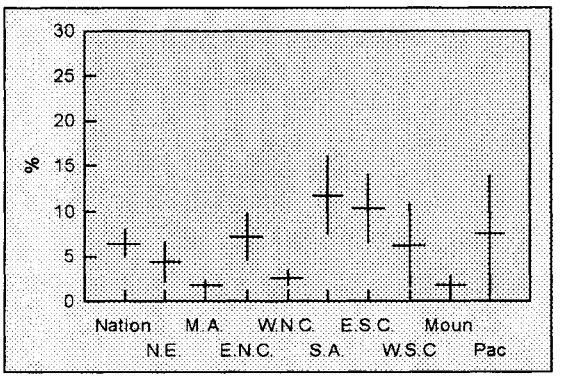
The Ungeocodable Match Rate Estimate

This estimate is an indication of the amount of units we found on the ground that were ungeocodable on the MAF.

Specifically:
This estimate is the percentage of MAF housing units existing in the sample blocks that are ungeocoded.

Figure 4 shows 90% confidence intervals for ungeocodable match rate.

Figure 4. Ungeocodable Match Rate



The national estimate is 6.4%, with a confidence interval ranging from 4.9% to 7.9%.

Ungeocodable Match Rate ranges from about 2% in the Middle Atlantic division to about 12% in the South Atlantic division. The South Atlantic estimate is significantly higher than the estimate in four other divisions.

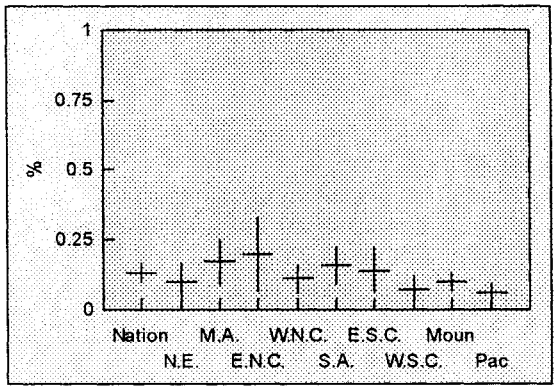
The Non-residential Coding Error Estimate

Non-residential coding error means we found a unit on the ground but it was incorrectly coded non-residential on the MAF.

Specifically:
This estimate is the percentage of residential MAF units in the sample blocks that are incorrectly coded non-residential.

Figure 5 shows the 90% confidence intervals for non-residential coding error. Notice that all graphs shown previously had a y-axis ranging from 0 to 30%. This graph has a y-axis ranging from 0 to 1%.

Figure 5. Non-residential Coding Error



The national estimate is only 0.13%, with a confidence interval ranging from 0.1% to 0.16%.

Non-residential coding error is less than a fourth of a percent in every census division.

Relationship Between the Estimates

One relationship between the estimates that is worth mentioning is that of coding errors vs. coverage errors. We were successful in distinguishing between coverage errors (undercoverage/overcoverage) and coding errors, to the extent that we located coding errors within zip code on the MAF. Without this distinction, cases that were actually on the MAF but coded in error would appear to us as missing from the MAF, or in other words, undercoverage. So, our undercoverage estimate is lower and more accurate than it would have been without accounting for these coding errors.

Conclusions

In theory, the MAF should contain all residential units in the nation. Because most of our coverage estimates are fairly high, this confirms the need for significant coverage improvement on the MAF prior to Census 2000.

The South Atlantic division appears to contain more MAF deficiencies than any other division. It shows the highest undercoverage rate, overcoverage rate, and ungeocodable match rate. These deficiencies may be due to the quality of the Delivery Sequence File in the Southeast and Mid-Atlantic postal service areas.

The lowest undercoverage rate in any division is 7% and lowest overcoverage rate is 8.5%. Because even the lowest estimates are relatively high, every census division shows a need for coverage improvement.

Because all of our estimates of non-residential coding error were so low, we do not believe this error is a major concern for MAF building.

As we approach Census 2000, one of the most important address building operations is the Block Canvassing operation. Again, the purpose of Block Canvassing is to improve the quality of the MAF in all mailout/mailback areas of the nation. Because of our relatively high estimates of errors on the Initial MAF, we believe the Block Canvassing operation is critical in ensuring the highest possible quality of the MAF.

In past censuses, field work has always been essential in creating the address file. The Initial MAF for Census 2000 was created without any field operations. So, it may not be a big surprise that we measured such a high number of errors. The Block Canvassing operation and several other field operations have been developed to update the MAF prior to Census 2000.

References

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

Barrett, Diane F. (1999), "The 1997 Master Address File Quality Improvement Program Pilot Study," internal memorandum for Robert Marx, Bureau of the Census.

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

Killion, Ruth Ann. (1993), "Coverage of Housing in the 1990 Decennial Census," internal memorandum for Thomas Walsh, Bureau of the Census.

Morrison, Joel L. (1997), "Draft Sections of the Census 2000 History," internal memorandum, Bureau of the Census.