# Dual System Estimates of Housing Units Based on Administrative Records

James E. Farber
U.S. Census Bureau, 4700 Silver Hill Road, Stop 9200, Washington, DC 20233, USA

**Keywords:** Master Address File; Census; Coverage

## 1. Introduction[1]

The Master Address File (MAF) is the Census Bureau's primary list of addresses. It was designed to meet the needs of both the decennial census and some periodic surveys, such as the American Community Survey. As its name implies, the MAF seeks to be a master list of every address in the United States, including intelligence about the presence and type of structure, whether residential, commercial, group quarters or even demolished or otherwise nonexistent. Bureau of the Census (1999) contains more detail about the MAF.

During Census 2000, the MAF received frequent and intensive updates from processes like national address canvassing and housing unit unduplication. These operations resulted in a MAF that begins the decade with a high degree of accuracy. Indeed, the net coverage of the MAF housing unit universe after Census 2000 was estimated to be 99.4 percent (Barrett *et al*, 2001).

But maintaining and improving the accuracy of the MAF over the decade will be a challenge because the costly and intensive field work of the census will not always be available. If the quality of the MAF deteriorates from its currently high level, the negative effects will be great. The sampling frame for periodic surveys will be poor, which could weaken important survey estimates. The address list for the 2010 census would also be deficient, creating coverage errors. And fixing the problem would likely be costly, especially if the solution was to canvass the entire nation.

Resource constraints require the development of tools to identify small areas in need of MAF improvement, where intensive efforts like field work could be targeted. One potential targeting tool is administrative records, data compiled by Federal agencies that administer programs such as Medicare. Administrative records have the advantages of being inexpensive and geographically comprehensive. They may provide a means of identifying small areas, like ZIP Codes or census blocks, where MAF coverage is poor.

The specific targeting methodology we proposed and tested in this research is dual system estimation, with the MAF as the first system and administrative records the second. Dual system estimation provides a convenient one-number summary of MAF coverage for small areas. This use of dual system estimation differs from its traditional application at the Census Bureau, which is to improve estimates or evaluate census counts. The goal in this context is not to provide an alternative set of housing unit estimates but rather to identify small areas potentially in greatest need of MAF updating. If successful, targeting based on dual system estimation and administrative records would enable more efficient use of scarce field and budgetary resources.

## 2. Background on Administrative Records

The Census Bureau has a long history of use of administrative records from agencies like the Social Security Administration and the Internal Revenue Service (Long, 1993). For this MAF research, we used a database of addresses from six administrative records sources:

- Internal Revenue Service (IRS) Individual Master file
- IRS Information Returns Master file[2]
- Medicare enrollment database
- Selective Service System registration file
- Department of Housing and Urban Development Tenant Rental Assistance Certification System file
- Indian Health Service patient registration file

The addresses on these files were combined and unduplicated during the creation of the Statistical Administrative Records System (StARS) 1999. The StARS 1999 is a national census-like database of address and person records collected solely from the administrative records listed above.

The addresses underwent a number of processes to produce an independent and unduplicated MAF-like file. One process involved geocoding to census blocks, with about 75 percent of the administrative records addresses successfully geocoded. For city-style addresses, those

[1]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.

[2]The Individual Master file is the file of tax returns and is often referred to as the 1040 file. The Information Returns Master file includes income reports from sources such as W-2 and 1099 forms.

with a house number and street name, geocoding was even more successful at about 85 percent. This high geocoding rate enables comparisons of the MAF and administrative records in very small areas, such as census blocks. Moreover, nearly all of the administrative records addresses have at least a five-digit ZIP Code, so we are able to place even non-geocoded addresses into reasonably small areas.

Farber and Leggieri (2002) give more information about StARS 1999, including details of the source files and the steps of address processing.

## 3. Targeting Methodology

In this research, we use dual system estimation to summarize the MAF quality for small areas. Dual system estimation has long been used to estimate the net population coverage of the decennial census based on the results of a post-enumeration survey (Kostanich, 2001). In the MAF context, the MAF is like the census and administrative records are the survey. We use administrative records to indicate where the MAF may be deficient. But unlike a survey, administrative data are available nearly everywhere. There is no sample and hence no sample size consideration that limits the detail of the comparisons. We can compare the MAF and administrative records directly in very small areas.

Ranking the areas by their dual system estimates (DSEs) then enables targeting. The areas with extreme DSEs are those that should be targeted for field work or other intensive updates. Large DSEs indicate potential MAF undercoverage, and small DSEs indicate MAF overcoverage.

If we use the absolute DSEs, we will generally target the areas with the most housing units and miss smaller areas where relative coverage may be worse. To avoid this, we calculate and rank areas by their relative DSE, which is the absolute DSE for the area divided by the number of MAF housing units.

Dual system estimation is our proposed methodology because it is relatively simple, and because it accounts for omissions from and erroneous inclusions in the MAF. Both types of errors are important because both can affect the accuracy of the census or surveys. For example, the American Community Survey uses the MAF as its sampling frame and also as a weighting control. Undercoverage can lead to bias in the American Community Survey estimates, while overcoverage can increase variance and squander valuable field resources on non-existent addresses. Therefore we want to target small areas where the gross MAF coverage appears most erroneous.

From Barrett *et al* (2001), a simplified version of the DSE in a post-stratum is $C\times(\frac{CE}{N_E}\times\frac{N_P}{M})$, where, in the MAF context,

- $C$ is the total number of MAF addresses
- $CE$ is the number of MAF addresses that truly exist
- $N_E$ is also the total number of MAF addresses
- $N_P$ is the number of addresses in administrative records
- $M$ is the number of matching addresses between the MAF and administrative records

In the traditional DSE based on a survey, the $N_E$ term applies to the total number of cases only in the sample areas. With administrative records, there is no survey hence $N_E = C$. Cancelling these terms reduces the DSE to $CE\times\frac{N_P}{M}$.

We estimate $CE$ within each post-stratum using the MAF housing unit coverage results from Census 2000 (Barrett *et al*, 2001). The post-strata in the Census 2000 housing unit coverage study were:

- occupancy status
- race/Hispanic origin domain of householder
- size of structure
- Metropolitan Statistical Area type/Type of Enumeration Area
- Census region

The sample size of the Census 2000 housing unit coverage study precluded the use of any geography below the national level. Our research involves DSEs within very small areas, such as census blocks. Thus we make the synthetic assumption that the MAF national correct enumeration rates apply uniformly within post-strata to smaller geographic areas.

A second assumption is that non-matched administrative records addresses physically exist and should be on the MAF. We can attenuate this assumption by post-stratification, placing MAF and administrative records addresses into groups based on their characteristics. For example, our administrative data come from programs administered to people. Therefore, most of the administrative records addresses correspond to occupied housing units. We would not want to broadly apply dual system estimation results in an area with many vacant MAF addresses.

Another assumption required to use dual system estimation correctly is independence between the two systems: the MAF and administrative records. This assumption is likely satisfied for occupied housing units. Intuitively, we believe there is no correlation between someone's likelihood to file taxes, for example, and the likelihood of their address being captured through one

of the MAF update operations. For vacant units, the independence assumption is violated because of the nature of administrative records. A vacant unit on the MAF has a low probability of occurring in administrative records. For example, very few tax returns are filed from vacant housing units.

There are other assumptions related to dual system estimation that may be violated in this context. The post-strata we use in this research were designed to minimize heterogeneity in the census housing unit universe. It is unknown if this post-stratification design also minimizes heterogeneity in the administrative records addresses. Wolter (1986) describes other assumptions required for dual system estimation that we will not cover here.

The goal of this research is to determine if DSEs can be a targeting tool to use in conjunction with other targeting methods, such as comparison of the MAF and U.S. Postal Service files. We want to compile a preponderance of evidence about which small areas have MAF coverage problems to enable more efficient allocation of scarce and expensive resources like field work. Because they are not the sole targeting tool, the DSEs do not require the level of precision needed for census adjustment, for example. Hence it is unnecessary to satisfy all of the DSE assumptions in this context.

## 4. Simulating Targeting in Census 2000

The MAF underwent a number of discrete update operations before Census 2000 that enables us to test the targeting potential of dual system estimation via simulation. The administrative records addresses in StARS 1999 were current as of about April 1999. In mid to late 1999, the MAF received updates from:

- block canvassing
- address listing
- Local Update of Census Addresses 1998
- Local Update of Census Addresses 1999 Relisting

Details of these MAF update operations are given in Hogan (1999).

We removed updates from those operations from a late 1999 MAF to simulate a MAF that was roughly concurrent with the StARS addresses. We then computed relative DSEs for blocks, ZIP Codes and counties. Finally, we ranked the areas based on their relative DSEs.

Our proposed method is to target those areas with the largest and smallest relative DSEs. We tested this method by examining where most of the updates occurred. If they were in the areas with extreme relative DSEs, then our targeting method is accurate.

For this research, we simulated targeting in the state of New Jersey, which had a large number of MAF updates in 1999. We excluded areas that received updates predominantly during address listing because the intent of address listing was to build the MAF in these areas. The pre-address listing MAF was known to be poor in these areas and hence targeting them was a foregone conclusion.

A graphical summary of the results for ZIP Codes is in Figure 1 on the last page of this paper. We omit the results for blocks and counties because they parallel the results for ZIP Codes. The figure demonstrates that larger relative DSEs generally are associated with higher percentages of MAF updates. The correlation is 0.96, showing a strong relationship between the relative DSEs and the percentage of MAF updates. This correlation is clearly affected by the few very large relative DSEs, but not to the extent one might think. Even for ZIP Codes with relative DSEs less than 5, the correlation between the percent of updates and the relative DSE is 0.75.[3]

Some of the relative DSEs are very large because even in non-address listing areas the MAF often began with few housing units. These large DSEs are clearly indicators that the MAF was missing addresses, as a large percentage of updates occurred in ZIP Codes with relative DSEs greater than 1000.

Table 1 demonstrates the DSE calculation for one ZIP Code in New Jersey. For simplicity, we compute the DSE using all addresses with no post-stratification. In the research, we post-stratified to compute DSEs.

*Table 1. Comparison of all addresses in ZIP 08317*

| | | MAF Tallies | | |
|---|---|---|---|---|
| | | In MAF | Not in MAF | Total |
| Admin. Records Tallies | In Admin. Rec. | 242 | 416 | 658 |
| | Not in Admin. Rec. | 33 | ??? | ??? |
| | Total | 275 | ??? | ??? |

The DSE for all addresses in ZIP Code 08317 is $CE \times \frac{N_P}{M} = (275 \times 0.9769) \times (\frac{658}{242}) = 730$, where 0.9769 is the overall adjustment for MAF erroneous enumerations (Barrett *et al*, 2001). The relative DSE is about 2.7. In this ZIP Code, there is evidence that the MAF was missing some addresses before the 1999 update operations.

Indeed, following the 1999 updates, this ZIP Code gained a large number of addresses. Block canvassing

[3] A single outlier was removed from the group of ZIP Codes with relative DSEs less than 5 when computing the correlation of 0.75.

and the other operations added 294 addresses and deleted 1 for a net total of 568 addresses on the MAF after updates. The MAF did not grow to 730 addresses, the value of the absolute DSE in this ZIP Code. But the relative DSE nonetheless indicated potential MAF undercoverage that turned out to be real. The MAF had coverage problems in this ZIP Code before the update operations, and administrative records provided an early indicator of the undercoverage.

## 5. Factors Associated with Accurate Targeting

The DSE method proposed in this paper appears to enable targeting, but it is not a perfect predictor of which small areas have poor MAF coverage. It will be important to use this method in conjunction with other targeting methods, some based on administrative records and some not, to maximize the efficiency of the MAF improvement process. It is worthwhile, though, to identify the factors associated with accurate DSE targeting. When the DSE method targets an area that the other methods do not, we want to assess whether the DSE is faulty or whether it is identifying a real coverage problem that the other methods simply miss.

Our goal is to create a model to quantify the effects of these factors on the accuracy of the DSE targeting method. We have not yet built this model, but we have analyzed the administrative records in areas where our simulated targeting was accurate. From this simple review of the data, the following factors appear to predict accurate targeting:

- a large proportion of addresses from IRS tax returns
- a large proportion of addresses from multiple administrative records sources
- a large proportion of geocoded addresses

Most of the population files a tax return, and most people list their home address on their tax return. Areas with many addresses from the IRS tax return file generally have good address coverage from administrative records. Therefore differences between the MAF and administrative records in these areas, reflected by very large or very small DSEs, tend to reflect real MAF coverage error.

Similarly, areas with administrative records addresses that come from multiple source files tend to have good coverage by administrative records. There are 63 different combinations of administrative records sources. Research continues on identifying the specific combinations that produce the greatest targeting accuracy.

Areas with many geocoded administrative records addresses also appear to have good coverage, and hence differences with the MAF indicate true MAF coverage problems.

A factor associated with poor DSE targeting is a large proportion of addresses from the IRS Information Returns Master file. These addresses often create extremely large DSEs that do not reflect large MAF undercoverage. A large part of the problem is that these are often business addresses, such as a bank or an accountant's office. We attempted to identify commercial addresses when building StARS, but were not always successful. Their inclusion in the administrative records address database biases the DSEs upward.

## 6. Conclusions

The accuracy of the MAF must be maintained over the decade to ensure reliable survey estimates and census enumeration. Because constant blanket field work is not possible, other methods are required to allocate resources to where they are most needed. This paper has demonstrated that administrative records have the potential to assist in targeting small areas for MAF updating operations. Specifically, DSEs based on the MAF and administrative records appear to identify areas, like ZIP Codes or census blocks, where MAF coverage is poor. Certain factors of administrative records appear to lead to more accurate targeting, while other factors actually detract from the DSE targeting method. We will continue to examine these factors and attempt to quantify their effects on the DSEs. But regardless of how accurate this DSE targeting method appears, it cannot be the sole source of targeting data. The accuracy of the MAF will be maximized by the development of a number of targeting methods that provide a preponderance of evidence about which small areas require the most attention.

## 7. References

Barrett, D.F., Beaghen, M., Smith, D. and Burcham, J. (2001), "ESCAP II: Census 2000 Housing Unit Coverage Study," Executive Steering Committee for A.C.E. Policy II Report 17, Washington, DC: Bureau of the Census.

Bureau of the Census (1999), "The Census Bureau's Master Address File (MAF) Census 2000 Address List Basics," <www.census.gov/geo/mod/maf_basics.pdf>.

Farber, J.E.and Leggieri, C.A. (2002), "Building and Validating the Statistical Administrative Records System 1999," Administrative Records Research Memorandum Series, Washington, DC: Bureau of the Census.

Hogan, H. (1999), "Specification of the Decennial Master Address File Deliverability Criteria for Census 2000," DSSD Census 2000 Procedures and Operations Memorandum Series #D-1, Washington, DC: Bureau of the Census.

Kostanich, D. (2001), "Accuracy and Coverage Evaluation: Dual System Estimation Results," DSSD Census 2000 Procedures and Operations Memorandum Series #B-9*, Washington, DC: Bureau of the Census.

Long, J.F. (1993), "Postcensal Population Estimates: States, Counties, and Places," Population Division Working Paper No. 3, Washington, DC: Bureau of the Census.

Wolter, K.M. (1986), "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association*, 81, 338-346.

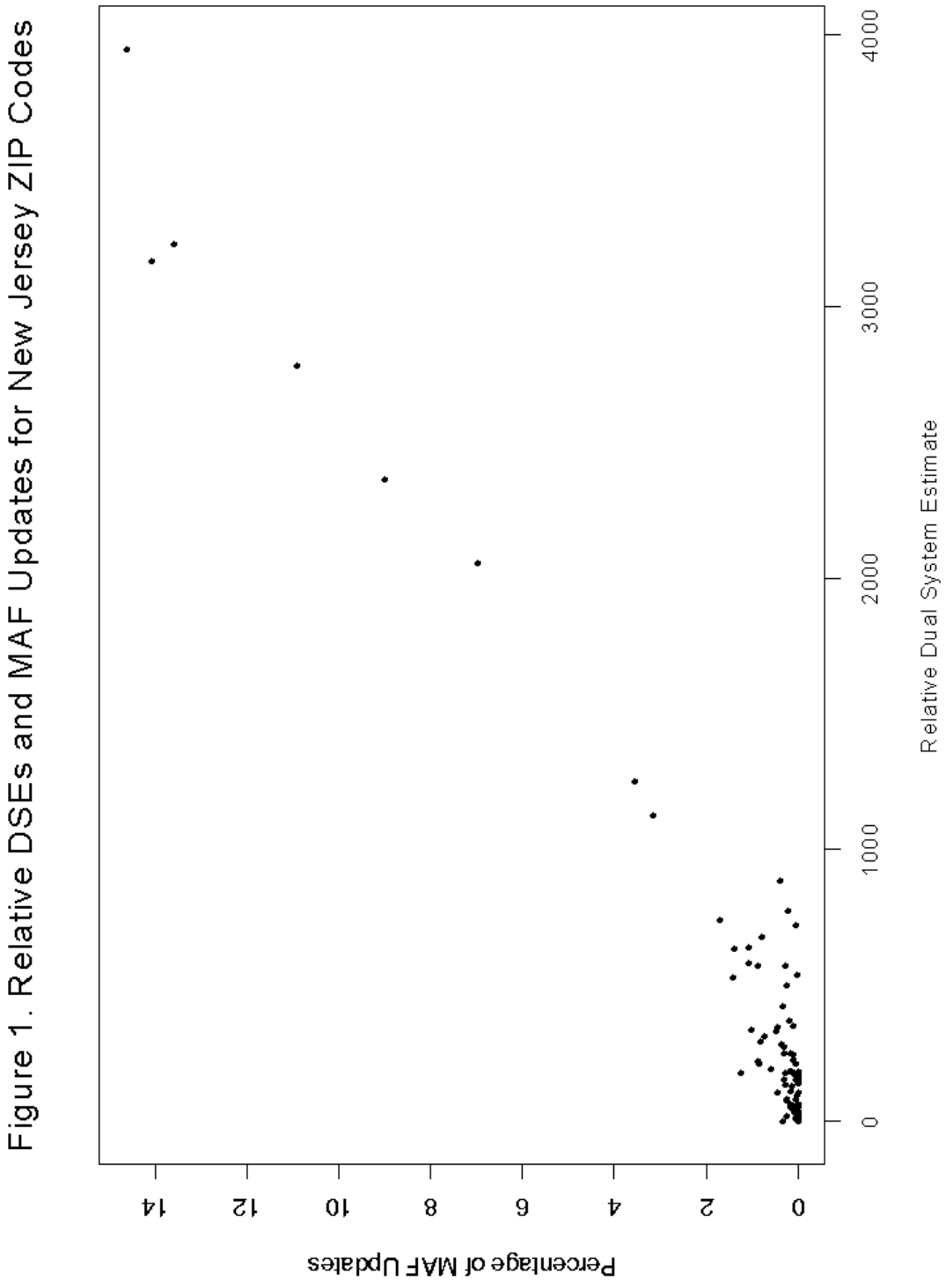
Figure 1. Relative DSEs and MAF Updates for New Jersey ZIP Codes
Percentage of MAF Updates
0
2
4
6
8
10
12
14
0
1000
2000
3000
4000
Relative Dual System Estimate