Assessing the Filter Rules For Extracting Addresses From the Master Address File To Construct a Housing Unit Frame for Current Demographic Surveys

Joel M. Martin *, Clifford L. Loudermilk
U.S. Census Bureau, Washington, DC 20233

Abstract

The Master Address File (MAF) is a national inventory of addresses for living quarters in the United States that the Census Bureau continually updates. One of the goals of the 2010 Demographic Household Survey Redesign is to switch to using a MAF-based frame for current demographic household surveys. To create the best possible current household survey frame from the MAF, a set of filter rules must be established to determine which MAF records should be accepted into the frame. The quality of the filter can have a major impact on the coverage of the resulting frame. We assess the effectiveness of various filter criteria by comparing the resulting frames with a benchmark address list collected from the field listings of a nationally representative probability sample.

Keywords: Sampling Frame, Filter, Surveys, Master Address File

1. Introduction

In the past, the Census Bureau has used multiple sampling frames that each focus on different types of housing units (HUs) to redesign the sample for the current demographic household surveys following each Decennial Census. Instead of using multiple frames that are created once every decade, the goal of the 2010 Sample Redesign is to reduce costs, maintain or improve coverage, and increase flexibility by selecting samples from one continually updated MAF-based frame. The Master Address File (MAF) is a national inventory of addresses for living quarters in the United States that is updated by many sources, including the previous decennial census, the Delivery Sequence File (DSF) from the United States Postal Service (USPS), and other small-scale operations.

Because the MAF is created from different sources and contains historical addresses and duplicated addresses, the MAF should not be considered a sampling frame in and of itself. Instead, the MAF is a source from which a sampling frame can be built. *Filter rules* must be developed to extract records from the MAF to build the sampling frame.

The filter rules for extracting records from the MAF have a significant influence on the coverage of the resulting sampling frame. The establishment of effective filtering rules will be critical to the success of the switch to a MAF-based sampling frame. This paper assesses a set of important filter rules that are under consideration for the current demographic household surveys when we convert to a MAF-based sampling frame after 2010. We identify MAF addresses within the filter subclasses in question and estimate the percent valid/invalid within each subclass using the outcomes from field listings of a nationally representative probability sample of census blocks.

2. Methodology

To conduct our analysis we compared the following two lists of addresses:

- The results files from field listings conducted in a nationally representative sample of Census blocks from October 2006 through August 2007
- The January 2007 MAF extracts, which reflect the MAF addresses prior to the field listings

---

* Joel Martin and Clifford Loudermilk are mathematical statisticians in the Demographic Statistical Methods Division of the U.S. Census Bureau. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

The field listings are based upon the National Evaluation Sample (NES), a nationally representative sample of 5,700 Census tabulation blocks. The addresses in the tabulation block sample are representative of the approximately 135 million valid American Community Survey (ACS) addresses in the nation. The NES design uses strata of region, block cluster size, percent of DSF adds in block cluster, permit status, and percentage of E-911 addresses in block cluster. The sample was selected with probability proportional to size (number of HUs in block clusters) with the exception of strata in the small block category. We selected the sample in small block strata using equal selection probability to avoid having higher weights for the smallest block clusters in these strata.

After we selected the sample, field representatives canvassed each NES block, using the existing MAF addresses as their starting point. Each address on the dependent list was either validated (as an existing, habitable HU) or invalidated (as nonexistent, uninhabitable, nonresidential, or a duplicate of another HU address).

With the information from the NES listings, we knew the "ground truth" for all addresses from the January 2007 MAF extracts that fell within the sample blocks. This allowed us to derive estimates of the percentage of invalid addresses within the filter subclasses we targeted.

The measure for addresses in filter subclass $\boldsymbol{X}$ is:

$V_{\boldsymbol{X}}$ = Percent of housing units in category $\boldsymbol{X}$ that were invalidated during NES field listings
= $A_{\boldsymbol{X}} \div B_{\boldsymbol{X}}$

where

$B_{\boldsymbol{X}}$ = The weighted sum of all housing units on the unenhanced list (i.e., the Jan-07 extracts) that were (1) in category $\boldsymbol{X}$, (2) within NES blocks, and (3) valid according to the filter rules

and

$A_{\boldsymbol{X}}$ = The weighted sum of housing units in $B_{\boldsymbol{X}}$ that were invalidated by field representatives (according to the results files from the NES listings)

We identified four important filter issues that will confront the current surveys when they convert to a MAF-based frame after 2010. The filter issues all deal with the creation of a housing unit (HU) frame. Group quarters (GQs) were not considered. For each filter issue, we identified the relevant subclasses of MAF addresses, and then calculated $V_{\boldsymbol{X}}$ for each subclass. The filter issues are discussed in the following four sections.

2.1 Excluded from Delivery Statistics (EDS) Addresses

The major source of new addresses for the MAF in the years between Decennial Censuses is the DSF, a national inventory of addresses maintained by the USPS. Addresses on the DSF are designated as either "Included in Delivery Statistics" (IDS) or "Excluded from Delivery Statistics" (EDS). The distinction is that IDS addresses are considered to be current mail delivery points, while EDS addresses are not. Should the HU sampling frame for current surveys include or exclude EDS addresses?

We believe that many EDS residential addresses are "housing units-to-be". They are planned residential housing units that have not yet been built. Further, since DSF updates are made from the ground up (that is, they are provided to USPS headquarters by the many thousands of local postal carriers and post offices), the EDS status of a DSF address on the MAF may lag considerably behind its true status at the time the surveys are creating their sampling frames (or conducting their interviews). Therefore, EDS addresses from the DSF could be a critical source of timely new construction updates for the current surveys.

We estimated the percentage of invalid HUs for all EDS addresses. We also calculated estimates for those subclasses of EDS addresses defined by the Delivery Point Type on the DSF.

2.2 DSF Start Date

To determine which DSF addresses from the MAF to include in a HU sampling frame, we must differentiate between those DSF addresses that were captured by the census versus those that were not. For those DSF addresses that were included on inputs to census field operations, this is not a problem since we know whether or not the census validated the address. But for the DSF addresses that were never acted upon by the census, the question is more difficult. The filter must define a DSF "start date", which will be the date of the DSF version from which new addresses will begin to be accepted into the sampling frame as "post-census adds". All DSF adds from earlier versions will be rejected by the filter, since it is assumed the census captured the addresses.

We estimated the percentage of invalid DSF adds from the DSF version used as the start date of November 1999 for the American Community Survey (ACS) filter, as well as some alternative DSF dates.

2.3 Census Deletes that Persist on the DSF

Another major challenge in constructing MAF filter rules is deciding what to do with those addresses that were deleted during Census 2000 field operations, but which persist as residential addresses on the DSF. Are they DSF errors that were correctly deleted by the census enumerators? Are they valid HUs that were mistakenly deleted by the census? Or were they invalid at the time of the census, but now valid?

We calculated a national estimate of the percentage of invalid addresses among the "census deletes that persist on the DSF". Then, we further analyzed the data for possible ways to target the valid addresses within this category of addresses.

2.4 Post-Census DSF Adds in the "Duplication Zone"

In developing a filter, we also must take into consideration the potential duplication of addresses that exists on the MAF. Should our filter reject DSF adds in those areas where the risk of duplication is thought to be most severe? The ACS filter does just that, defining a "duplication zone" in which DSF adds are not accepted into the ACS HU frame. Addresses in this "duplication zone" tend to be rural with a relatively high probability of duplication with census addresses because of E-911 readdressing activities. We were limited in our ability to analyze the addresses in this duplication zone, but did identify one subclass of the rejected DSF addresses for which we produced estimates.

3. Results

3.1 Excluded from Delivery Statistics (EDS) Addresses

Each address on the DSF is designated as either "included in delivery statistics" (IDS) or "exclude from delivery statistics" (EDS). The distinction is that IDS addresses are considered current mail delivery points, while EDS addresses are not. The IDS addresses are obvious candidates for a housing unit sampling frame, but what about the EDS addresses?

We believe that many of the EDS addresses are planned new construction---future HUs that are added to the DSF in the planning stages before they receive mail. On average, about 35% of the new addresses added to the MAF during each DSF update are EDS addresses. Should we exclude this substantial share of the new DSF addresses from our frame because they were not yet mail delivery points when added to the DSF? Or should we include them because they may be HUs by the time they are contacted for interview by one of the demographic surveys?

First, we estimated the percent invalid by the most recent IDS/EDS status for all the post-census DSF adds in our sample blocks that were valid according to the ACS filter:

Table 1: Percent of Invalid Housing Units by Most Recent DSF Delivery Status for Post-Census DSF Adds

Delivery Status On Most Recent DSF	Percentage of Invalid Housing Units (SE)
IDS	14.8 (1.2)
EDS	49.1 (3.9)

The EDS addresses, while more often "bad" than the IDS addresses, were still found to be valid, habitable HUs approximately 50% of the time. This relatively high percentage of invalids for current EDS records is a concern, though, so we looked further to see if we could find a way to separate "good" EDS records from "bad" EDS records. The estimates in Table 2 below are based upon the initial IDS/EDS status for these addresses.

Table 2: Percent of Invalid Housing Units by Initial DSF Delivery Status for Post-Census DSF Adds

Delivery Status On Initial DSF	Percentage of Invalid Housing Units (SE)
IDS	20.4 (1.5)
EDS	12.7 (1.4)

DSF addresses that started out as EDS were actually a little better than those that were initially IDS. This surprising result may lend support to the notion that EDS addresses are planned new construction HUs that do become valid HUs, while the IDS classification may include a wider variety of addresses, including some of poorer quality.

Table 3 below shows the estimated percentage of invalid housing units for post-census DSF adds, classified by IDS/EDS status at first DSF appearance versus most recent DSF appearance. The estimates exclude the most recent DSF additions. Table 3 does not contain information from DSF additions that first appeared after the Sep-2005 DSF version. This means that each address included in the analysis appeared on at least three DSF versions.

Table 3: Percent of Invalid Housing Units by First vs. Last DSF Appearance IDS/EDS Status

IDS/EDS Status on First vs. Last DSF Appearance for Post-Census DSF Adds	Percentage of Invalid Housing Units (SE)
IDS -> IDS	19.2 (1.5)
IDS -> EDS	49.6 (4.6)
EDS -> IDS	6.4 (1.1)
EDS -> EDS	48.9 (5.3)

The category of housing units that went from EDS on their initial DSF delivery to IDS on the most recent DSF delivery has the smallest percentage of invalid housing units. At 6.4% invalid, the EDS -> IDS category is considerably better than the category of housing units that were initially IDS and remained IDS (19.2 %).

While perhaps surprising, a logical case can be made for this result. A change from EDS to IDS status is an affirmative action---the USPS has evaluated the address sometime after its inception on the DSF and validated it as a current mail delivery point. For addresses that have always been IDS, we have no evidence that they have been through any evaluation since they were first put on the DSF. If we assume that some have never been evaluated, and some portion of these are truly "bad" addresses, then the result makes sense.

The final question about EDS addresses that we considered: Are there any subsets of the current EDS addresses that are considerably worse than others? The DSF provides a sub-classification of EDS addresses by Delivery Point Type. The current ACS filter accepts all EDS addresses (subject to other criteria) except the Delivery Point Type="X" addresses. Table 4 shows our estimates for the ACS-valid addresses in the other Delivery Point Type categories.

Table 4: EDS Addresses: Percent of Invalid Housing Units by Delivery Point Type (for ACS-Valid HUs)

Delivery Point Type	Percentage of Invalid Housing Units (SE)
R – (Residential Curbline)	52.5 (3.7)
S – (Residential Neighborhood Delivery Centralized Box Unit)	42.4 (9.4)
T – (Residential Central)	29.5 (12.6)
U – (Residential Other)	59.1 (15.2)
All EDS Housing Units	46.8 (3.6)

Due to the large standard errors, there doesn't appear to be a definitive case for excluding any EDS records based on Delivery Point Type. The "U" addresses seem most problematic and deserve a closer look, but we do not have enough information from the NES data to propose any changes to existing filter rules.

A fifth Delivery Point Type, Delivery Point Type="X" addresses, could not be included in Table 4. These "X" type delivery points are addresses that ACS has already marked as invalid and thus were not included in the analysis. We looked at all "X" addresses on the latest DSF that were not in the census. Our estimate is that 78.6% of these "X" addresses are invalid, which supports the idea that these records should be excluded from our HU sampling frame. While the "U" and "R" addresses seem problematic, there doesn't appear to be a definitive case for excluding them because doing so would exclude the approximately 50% that are valid housing units. Addresses that are initially EDS and remain EDS on the most recent DSF are approximately 49% invalid while EDS addresses that changed to IDS addresses are approximately only 6% invalid. Therefore, the most promising filter change to separate "good" from "bad" EDS addresses might be based on the length of time that an address has been EDS on the DSF/MAF.

3.2 DSF Start Date

The DSF "start date" is an important element of any MAF filter. In effect, it represents the hypothetical dividing line between pre-census DSF addresses and post-census DSF addresses on the MAF. Any addresses appearing on the DSF before the start date, which were not linked to addresses captured by the census, are discarded by the "start date" filter condition. We presume that these addresses are old enough that they probably duplicate census addresses (but in a form that did not allow us to match them) or are simply erroneous addresses (because the census did not find them). Addresses appearing on the DSF on or after the start date are accepted because we believe it more likely that they represent new addresses that didn't come into existence until after the census.

The ACS filter uses a start date of Nov-1999, so any address appearing for the first time on the Nov-1999 version of the DSF or later is accepted into the ACS HU frame (subject to other filter criteria). However, there were also two earlier versions of the DSF used to construct the MAF: Nov-1997 and Sep-1998. Should addresses appearing on either of these DSF versions, but not captured by Census 2000, be accepted by the current surveys?

Table 5: Percent of Invalid Housing Units by DSF Start Dates

DSF Start Date	Percentage of Invalid Housing Units (SE)
DSF 1 November – 1997 Adds (and earlier)	68.6 (3.2)
DSF 2 September – 1998 Adds	42.4 (9.3)
DSF 3 November – 1999 Adds	27.3 (3.1)
DSF 4 February - 2000 Adds (adds only)	20.3 (6.9)
DSF 5 March - 2000 Adds	28.7 (5.7)

Approximately 69% of the earliest DSF addresses (those that appeared on the very first DSF in Nov-1997) were found to be "bad". Presumably, these addresses are duplicates of census addresses or erroneous addresses never deleted from the DSF. The Sep-1998 adds are better, with approximately 42% of them being invalid. Only 27.3% of the Nov-1999 adds are deemed to be "bad". The Nov-1999 start date seems reasonable but the Sep-1998 start date might also be acceptable.

To further demonstrate the distinction made by the November 1999 start date we compare all the DSF adds after November 1999 to those that first appeared on the DSF before November 1999 in Table 6 below.

Table 6: Percent of Invalid Housing Units by DSF Start Dates

DSF Start Date	Percentage of Invalid Housing Units (SE)
DSF 1 and **DSF 2** (before November 1999)	66.1 (3.3)
DSF 3, **DSF 4**, and **DSF 5** (after November 1999)	26.2 (2.6)

The results in Table 5 and Table 6 may not be directly applicable to the process of creating a post-2010 filter for the current surveys. The timing of the DSF updates to the MAF in the period before the 2010 Census may be very

different than what is shown in Table 5. The results, though, should be helpful when considering the "start date" question for the current surveys filter.

3.3 Census Deletes that Persist on the DSF

In early 2008, there were more than 1.5 million addresses on the MAF that were residential addresses on the latest DSF, but flagged for deletion by Census 2000 operations. Should the current surveys' filter reject these census deletes, even though they persist on the DSF as residential addresses? Table 7 shows our estimate for this class of DSF addresses.

Table 7: Percent of Invalid Housing Units among Census Deletes that Persist on the DSF

Census Deletes that Persist on the DSF	Percentage of Invalid Housing Units (SE)
All Census Deletes that Persist on the DSF	35.6 (4.5)
Urban Blocks	38.3 (5.2)
Rural Blocks	25.0 (9.0)

As expected, the estimate reveals a substantial mix of both "good" and "bad" addresses. Presumably, the 36% that are "bad" are DSF errors that were correctly deleted during the census, but have never been cleaned up on the DSF. While some of the 64% that were "good" may have been deleted in error by the census, we suspect that many of them may have been new construction HUs with long construction lags. That is, they may have been planned HUs added to the DSF in advance of having been built (see the discussion of EDS addresses in 3.1), but when construction still had not begun by the time of the census, the addresses were deleted. Later, when construction was finally completed, the addresses remained flagged as census deletes on the MAF.

Is there any way to target addresses within this category that are substantially better than others? Our research yielded several possibilities. Table 7 also shows the results of breaking down the census deletes that persist on the DSF into Urban and Rural subcategories. If the "bad" addresses are primarily DSF errors, then the errors are more pronounced in urban areas than in rural areas.

"Blue line status" is another variable we used to discriminate between valid and invalid addresses among the census deletes that persist on the DSF. "Blue line" is a Census Bureau concept used to separate areas according to the method in which census data is captured. In areas "inside the blue line", census forms are mailed out and mailed back, while other collection methods, including personal visits are required in "outside the blue line" areas. The areas outside the "blue line" are mostly rural in nature.

Table 8: Census Deletes that Persist on the DSF by Blue Line Status

Census Deletes that Persist on the DSF	Percentage of Invalid Housing Units (SE)
Inside the Blue Line	37.5 (4.7)
Outside the Blue Line	15.2 (14.2)

It appears that the percentage of invalid addresses outside the blue line (15.2%) is lower than the percentage of invalid addresses inside the blue line (37.5%). However, the fairly large standard errors make the distinction much less certain.

Additional research into the census deletes that persist on the DSF may unearth other subclasses that have even lower percentages of "bad" addresses than those in rural blocks or outside the blue line. This information would be helpful in refining the current survey filter rules for this category of addresses if the estimate of 35.6% invalid is deemed to be unacceptably high.

3.4 Duplication Zone Post-Census DSF Adds

The potential for duplication of addresses on the MAF is an important concern for anyone who wishes to use the MAF as a source for a sampling frame. There are two major reasons why duplication is created on the MAF:

- Addresses are added to the MAF from many different sources (census operations, the DSF, post-census listing operations, etc.), with unduplication among these addresses accomplished by complicated address matching algorithms. No matter how effective the algorithms, though, duplication is unavoidable.

- Addresses in some areas of the nation have changed since Census 2000 as part of local E-911 readdressing operations (to make the addresses easier to locate for emergency vehicles). These address changes are most common in rural areas which previously had a large number of non-city-style addresses (addresses missing either a house number or street name, such as RR 1 Box 10). The new addresses added by the DSF will duplicate the old addresses from the census unless the addresses are explicitly linked for us by the USPS.

The ACS filter rejects all new DSF addresses that have been added to the MAF in certain geographic areas (termed the "duplication zone") due to the fear of excessive duplication. As a result, the ACS HU universe remains essentially static in these areas. Can we be confident enough of the DSF adds in any portion of the "duplication zone" to accept the adds in our current surveys filter?

Table 9: Percent of Invalid Housing Units for Duplication Zone Post-Census DSF Adds

Duplication Zone Post-Census DSF Adds	Percentage of Invalid Housing Units (SE)
Single Unit	26.3 (0.3)
Multiunit	43.4 (1.1)
Total	**32.3 (0.4)**

Table 9 shows our estimate for those addresses that are in "duplication zone" blocks "outside the blue line" and have a mix of city-style and non-city-style addresses, with fewer than 100% of the addresses in the block having the DSF as a source. These addresses do not constitute the entire set of "duplication zone" addresses (as defined by the ACS filter), but meaningful estimates for the other types cannot be produced from our data.

It appears that the multiunit housing units in the duplication zone might be more likely to be invalid than the single housing units in the duplication zone. The total percentage of invalid housing units is only approximately 32%. Therefore, we should consider accepting these particular adds in our current surveys filter, unless further research contradicts the current finding.

4. Limitations

In most cases, our analysis provides estimates only of the percentage of invalidated addresses within subclasses of addresses already declared valid by the filter. These estimates help us determine which subclasses of addresses currently pass the filter when, in fact, they should be rejected. But we are more limited in our ability to identify subclasses of addresses that are currently rejected by the filter when they should be accepted. In other words, this paper primarily considers filter inclusion errors - a source of overcoverage error. We also need to find ways to evaluate the filter exclusion errors - a source of undercoverage error.

Our analysis was restricted to records that were assigned to a census block on the January 2007 MAF extracts. We were unable to analyze DSF records that were not assigned to census blocks.

5. Conclusion

Our research yielded these conclusions:

- While current EDS records are more likely to be "bad" than current IDS records, the reverse seems to be true when the initial IDS/EDS status is considered. That is, records that initially appeared on the DSF as EDS are more often valid than those that first appeared as IDS. Given this, we think that a filter rule to exclude DSF records that remain EDS for an extended period of time would likely improve the sampling frame for current surveys.

- Among EDS addresses, those with Delivery Point Types "X" are mostly invalid and should be excluded from a current surveys filter. While other Delivery Point Types seem problematic, we do not feel there is enough evidence to recommend any further exclusions based on Delivery Point Type.

- The DSF start date used in the ACS filter (Nov 1999) seems better than the earlier alternatives, although the Sep 1998 start date may also be acceptable. However, because the timing of the DSF deliveries may vary, these start date results may not be directly applicable to a post-2010 filter. The start date used by the filtering rules of a decennial census may matter as well.

- The census deletes that persist on the DSF are approximately 36% invalid, which presents a challenge for a current surveys filter. If possible, we should do more research to try to target the valid HUs within this category.

- We were able to analyze one subclass of DSF adds rejected because they are in the "duplication zone", finding that roughly 32% are bad. There is not enough evidence to exclude these "duplication zone" adds from the current surveys frame.

6. References

Martin, Joel M. and Clifford L. Loudermilk; "Frame Assessment for Current Household Surveys: Filter Rules Research (FACHS-FRR)."

Li, Mei, Clifford Loudermilk, and Xijian Liu; "Frame Assessment for Current Household Surveys – National Evaluation."

Loudermilk, Clifford L. and Timothy L. Kennel; "Deciphering the DSF: Which Addresses from the Delivery Sequence File Should be Included in the Sampling Frames for Demographic Surveys?", August 7, 2005.

Zimolzak, Matthew and Lawrence Bates; "Customer Requirements Document for American Community Survey Geographic Products (V 1.1)", May 9, 2008.