Comparing the Quality of the Master Address File and the Current Demographic Household Surveys' Multiple Frames

Xijian Liu[1]
US Census Bureau
4600 Silver Hill Road, Washington, DC 20323

Background

The U. S. Census Bureau currently uses a costly multiple frame system to support several major demographic household surveys. These surveys include the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), Consumer Expenditure Surveys (CE), American Housing Survey (AHS), National Crime and Victimization Survey (NCVS), and the State Child Health Insurance Program (SCHIP). The existence of the Master Address File (MAF) provides a potentially more cost efficient frame option for these surveys. The Census Bureau is currently conducting evaluations to support the goal of using a MAF-based frame to replace the current multiple frames for the 2010 Demographic Survey Redesign.

The current frames include a unit frame, a group quarters (GQ) frame, a permit frame, and an area frame. The unit frame and the GQ frame cover addresses of housing units and group quarters that existed during the time of Census 2000 in blocks that had a high percent complete addresses, i.e., addresses that had a house number, a street name, and a zip code, and had building permit coverage. The permit frame obtains addresses from building permit offices (BPOs) to capture new construction after Census 2000 in blocks that had building permit coverage. For the remaining universe that is primarily in rural areas, an area frame approach is used. The current area frame covers approximately 12% of the addresses in the U.S. The field operations used to collect addresses for the permit frame and the area frame are expensive.

The MAF is an inventory of addresses for all known living quarters in the United States and Puerto Rico and is maintained by the Census Bureau. The MAF is updated by operations related to decennial censuses, by the Delivery Sequence File (DSF) from the United States Postal Service every six months, and also by area listing operations from the Demographic Area Address Listing (DAAL) operation. The American Community Survey (ACS) uses the MAF as the source of its sample addresses.

The Frame Assessment for Current Household Surveys (FACHS) is an evaluation program developed to support the goal to use a MAF-based frame for current household surveys for the 2010 Demographic Survey Redesign. The primary focus is to compare the coverage and content quality of the current multiple frames and the MAF extracts in the entire universe and in the two sub-universes corresponding to the current area frame and permit frame. The current unit frame and GQ frame were extracted from the MAF after Census 2000. This paper gives an overview of three evaluations of the FACHS, the area frame study, the new construction study, and the national evaluation study.

In this paper, the MAF-based frame means an address frame constructed using addresses from the MAF according to current ACS eligibility criteria or filtering rules. Filtering rules have an impact on coverage measures in this paper. The MAF is an inventory of addresses of living quarters for the nation, including some known bad addresses. To construct an address frame from the MAF, one must use a set of criteria, or a set of filtering rules to screen out certain unwanted addresses. One simple example of an unwanted address is an address being flagged as commercial delivery point on the DSF. Most FACHS evaluations use filtering rules developed for ACS to determine which addresses are eligible to be included in a MAF-base frame. There are ongoing efforts by ACS to improve its filtering rules. FACHS will conduct research and will recommend filtering rules for current surveys. In addition, the MAF-based frame considered in

---

[1]Xijian Liu is a mathematical statistician 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. Any views expressed on statistical and methodological issues are those of the author and not necessarily those of the U.S. Census Bureau.

this paper does not take into account of the major MAF improvements done during the 2010 Census operations and possible future improvement operations by current surveys.

The Area Frame Study

The Area Frame Study compares the address lists obtained from area listing operations for current household surveys to the MAF extracts prior to the operations. The primary focus of this study is the coverage in the rural areas. Area frame blocks were mostly outside the mail out and mail back enumeration area in Census 2000. The 2000 Housing Unit Study (Barrett et al, 2003) found that the gross undercoverage rate was about 6.5% for the decennial MAF in those areas.

Some results from this study have been published by Kennel and Corlett (2005). These results were derived from comparing the addresses on the January 2004 MAF extracts to the addresses from area frame listing for the 2004 panel of SIPP and assuming the latter were very close to the ground truth. These area listings for SIPP were conducted between September 2003 and February 2004. The January 2004 MAF extracts included the DSF updates in September 2003. The study employed ACS filtering rules to extract valid addresses from the MAF. The major findings are summarized in Table 1.

Table 1. Coverage Rates for the Jan 2004 MAF extracts for the Area Frame Sub-Universe

Percent of addresses from area listing that were not found on the MAF (undercoverage)	7.6% (0.3%)
Percent of addresses from area listing that were found on the MAF but excluded by ACS filter (undercoverage)	4.4% (0.2%)
Percent of addresses on the ACS MAF that were invalidated by area listing (overcoverage)	10.7% (0.4%)

Data provided by Kennel and Corlett (2005). Standard errors are in parenthesis

The two undercoverage components shown on Table 1 contributed to a total gross undercoverage of 12.0% of the area frame sub-universe for the January 2004 MAF extracts.

A similar study has recently been completed using addresses from area listing for the Current Population Survey (CPS) and yielded very similar coverage results (Kennel, 2007). This subsequent study also performed more detailed analysis of the addresses that were not found on the MAF. It found that about 40% of the missing addresses were incomplete addresses. The current MAF update process does not use incomplete addresses from the DSF because of concerns about duplication. It also found that about 40% of the missing addresses were addresses of manufactured homes. Another natural question is whether the missing addresses were new construction built after Census 2000. Year build information was collected by survey listing operation only for conventional housing units (i.e., not manufactured homes) in permit issuing blocks. Kennel (2007) found that of the missing addresses that were conventional housing units in permit issuing blocks, about half were built after Census 2000 and the other half were older addresses missed by Census 2000. Unfortunately, we did not have year built information for blocks not covered by building permits.

The results from the Area Frame Study also showed that filtering rules could have an impact on address coverage. Improving the filtering rules can reduce the amount of addresses that were on the MAF but were erroneously excluded by the filtering rules.

The size of the area frame of the current household surveys has decreased over time. The percentage of addresses in the area frame for current household surveys was 100% in pre-1960 designs, about 30% in 1990 design, and about 12% in 2000 design. We expect this national trend of more and more areas being converted to using city-style addresses to continue. But the situation could be quite different in different parts of the nation. For the 2000 design, over 50% of the area frame addresses are in the South Census Region and two states have over 50% of their addresses in the area frame.

The New Construction Study

In this paper, we refer to addresses of new construction as addresses built after Census 2000. There are two aspects of the new construction address coverage. One is how completely an address list contains all new

addresses since Census 2000 for a specific time, e.g., on December 31 2006. The other is how soon new addresses appear on an address list after their "birth", e.g., building permit authorization or construction completion.

One project aimed at comparing the new construction coverage rates of the MAF and the coverage rates of the permit frame (Johnson and Hakanson, 2007) and focused on the first aspect above. It established three address lists: the MAF extracts, a proxy for the permit frame, and the ground truth. The MAF extracts used by this study were the extracts for ACS and included updates by the DSF in Sept 2005. The most difficult list to establish was the proxy for the permit frame because current household surveys collected only sample permit addresses but not all permit addresses issued after the start date of the permit frame. The project team had to collect complete lists of permit addresses since 1999 from selected BPOs. The study used a probability sample of 31pairs of BPOs. In 22 of the 31 pairs, at least one BPO was able to give a complete list of permits issued from 1999 to October 2005. To establish the ground truth, Census Bureau field representatives conducted area listing in a sub-sample of 1,146 blocks selected within the areas covered by the these permit offices between April and August 2006. The addresses from area listing were matched to the permit addresses and to addresses on the MAF. Then, the addresses from area listing were determined as the ground truth for post Census 2000 new construction by an algorithm using matching results or field observations.

The main results are shown in Table 2.

Table 2. Percent of new construction addresses from area listing that did not match to the MAF or the permit frame

	MAF	Permit
Including manufactured homes	34.4 (6.8)	43.5 (10.1)
Excluding manufacturer homes	30.5 (7.2)	32.1 (8.7)

Data provided by Johnson and Hakanson (2007). Standard errors are in parenthesis.

The standard errors shown in Table 2 were quite large and none of the comparison of coverage rates between the MAF and permit frame show statistical differences at the 10% level. However, the current permit frame is known not to have very good coverage of new manufacturer homes because many building permit offices do not issue permits for placement of new manufacturer homes.

This study had several limitations that any project of measuring new construction address coverage may have. These limitations may have contributed to the larger than expected gross undercoverage measures for the MAF and for the permit frame. This project had small sample sizes and was designed as a test for a larger scale study. Because of various reasons, we will not conduct the larger study. Despite these limitations, this was a very interesting project.

The two other New Construction Study projects use a different approach and rely on matching existing permit frame sample addresses to MAF address. The Permit-DSF Lag Study, a second project, measured the lag time from when a building permit was authorized to when we received the address through the DSF, or the permit-DSF lag (Flanagan and Loudermilk, 2006). For this project, we matched more than one million permit sample addresses collected from the Permit Address Listing field operation between 1999 and 2002 for current surveys to the January 2006 MAF extracts. For the matched addresses, we compared the month of permit authorization and the month we first received the address from the DSF, usually in March and September of each year. The main findings of this project are shown in Table 3.

Table 3. Permit-DSF lag time by type of structure

Type of structure	Median lag (in month)	Average lag (in month)
Single	6	4
Multi	13	12
All	7	6

Data provide by Flanagan and Loudermilk (2006)

According to data released for 2005 from the Survey of Construction, the average number of months it took from permit authorization to construction completion was about 7 for single unit and 13 for multi-units. This suggested that the DSF seemed to be able to capture new construction rather quickly after construction completion, at least at the national level.

Another interesting finding from this lag study is that about one quarter of the addresses we received on the DSF arrived in the same month or before the month the permit was authorized. People may commonly expect the DSF to capture a new address after the tenants moved in to the new house. It is somewhat surprising that it is not unusual for the DSF to include addresses before permit authorization.

One limitation of this study was that it was not able to produce standard errors because it was very difficult to determine the sampling probability for sample addresses selected for multiple surveys. The lag measures were also derived from matched addresses that ware about 83% of all permit sample addresses include in the study.

The third study we conducted was a Permit Sample Address Matching Study. We matched the CPS sample addresses for November 2005 interviewing to the January 2007 MAF extracts that were updated by the September 2006 DSF. Then we determined the first time an address was received from the DSF as non-commercial and derived the percentage of permit sample addresses that were on the MAF for various points of time. This study also incorporated additional information from time of interviewing to screen out addresses that were out-of-scope for CPS, e.g., addresses of construction not yet started or completed, vacant housing units, and addresses that had only persons out-of-scope for CPS. More detailed information on CPS design can be found in the CPS Technical Paper 66 (Census Bureau, 2006).

The matching results are shown in Table 4. This result indicates that a high percent of the permit addresses that responded to CPS eventually appeared on the MAF. It is interesting to notice that the MAF was able to capture 93-94% of the sample addresses around the time of interviewing. For surveys, operational lag time does exist. Although in November 2005 the MAF would have high percent of the permit sample addresses on the ground, surveys operations require processing time and can only select samples from earlier MAF extracts. From this point of view, the permit frame does have some advantage of generally obtaining new addresses earlier and allows more time for processing. For a MAF-based frame, it will be critical to design an efficient sampling system that minimizes the lag time from when new addresses become available from the DSF to when interviewing is conducted.

Table 4. Percent of permit sample addresses for November 2005 CPS interviewing that were matched to addresses on the MAF

September 2004	82.8 (1.06)
March 2005	89.4 (0.79)
September 2005	93.0 (0.65)
March 2006	94.0 (0.62)
September 2006	94.5 (0.62)

Standard errors are shown in parenthesis

CPS final household weights and associated replicate weights were used to derive the measures for the above table. It should be pointed out that the definition of a match on Table 4 required that a permit sample address be matched to a MAF addresses came from the DSF as a non-commercial address. Regardless of commercial status, 95.8% of the permit sample addresses matched to an address on the September 2006 MAF.

One limitation of this study is that it did not account for the possible "gain" of a MAF-based frame over the permit frame when the MAF captures some new construction addresses before permits were issued.

Characteristics of addresses of MAF undercoverage

We are interested in knowing whether the current surveys' sample addresses that may not be on a MAF-based frame have different characteristics than other sample addresses. For this purpose, we did the

following analysis using the CPS sample addresses for November 2005 interviewing and using results from the area frame study and the permit sample address matching study.

We define permit frame sample addresses that may not be on a MAF-based frame as those addresses that did not appear on the DSF for the first time by September 2004 (17.2%, see Table 4). We define area frame sample addresses that may not be on a MAF-based frame as those addresses determined as MAF undercoverage in the second area frame study (Kennel, 2007). Kennel (2007) reported a 12.7% gross undercoverage rate for the area frame. These area frame "undercoverage" addresses included addresses built after April 2000 in permit issuing blocks that were excluded from area sampling and addresses determined as out-of-scope for CPS at time of interviewing such as vacant addresses, vacant trailer pads, and construction not yet completed. The weighted percent of area frame gross undercoverage rate reduced to 8.0% when CPS household weights were applied.

Table 5. Weighted percent of CPS sample addresses not on the MAF-based frame

	Number of responding households	Weighted percent of CPS sample addresses not on a MAF based frame
Unit/GQ	44,479	0
Area	7,080	8.0 (0.44)
Permit	3,476	17.2(1.06)
All frames	55,035	2.0 (0.08)

The next table shows major labor force characteristics collected for CPS in November 2005 interviewing at addresses that may or may not be on the MAF based frame. CPS final person weights and associated replicate weights were used. The percent of population in combined labor force is the weighted number of persons in combined labor force divided by the weighted total of persons in-scope for CPS. The unemployment rate is the weighted number of unemployed persons divided by the weighted number of persons in combined labor force.

Table 6. Percent in combined labor force and unemployment rate, Nov. 2005

	Percent of population in combined labor force	Unemployment rate
Unit/GQ	66.2 (0.17)	5.0 (0.12)
Area on MAF	61.8 (0.69)	4.2 (0.32)
Area not on MAF	62.0 (2.16)	5.6 (1.29)
Permit on MAF	72.2 (0.92)	3.1 (0.32)
Permit not on MAF	73.1 (1.83)	3.3 (0.66)
Full sample	66.1 (0.14)	4.8 (0.11)

Standard errors are shown in parenthesis

As Table 6 shows, there are clear differences of labor force characteristics for persons in the unit/GQ frame and persons in the permit frame. Standard errors for not on the MAF-based frame categories are generally high and no statistical difference can be detected for the comparisons of labor force characteristics between on MAF based frame and not on MAF based-frame categories within the area frame and within the permit frame. Although not statistically different, the larger unemployment rate point estimate for area frame addresses not on the MAF-base frame may require further investigations.

One limitation of this analysis, like some of the other studies, is that it only accounts for the potential "loss" and not for the potential "gain" of sample addresses of using a MAF based frame. Further study may also be conducted to include certain vacant addresses. Deficiency associated with the matching process may inflate the amount of MAF undercoverage as well.

The National Evaluation

This study intends to derive national level and census region level gross coverage estimates of the MAF. It uses a methodology similar to that of the Area Frame Study and compares the MAF extracts to addresses obtained from area listing to derive MAF coverage measures. This study collects addresses using area

listing in a probability sample blocks in the U. S. Then these addresses will be used as the ground truth and are compared to the MAF extracts to derive gross undercoveage and overcoverage measures. By August 2007, we have completed listing operations for the 2007 sample of about 5,800 blocks. We also plan to repeat this study in 2008 and to produce a report using two-year worth of data by the end of 2008.

Summary

Most FACHS evaluations are underway and will be completed by the end of 2008. We will then make a formal decision as to whether we should use a MAF-based frame. The preliminary results available so far do not give us any warning signs for switching to a MAF-based frame, at least for the quality at the national level. We do recognize that there may be needs for MAF improvements, e.g., in states that had a large proportion of rural areas. Developing an efficient sampling operation that minimizes the lag time from when the DSF is available to when the interviewing is conducted could also be a challenge. FACHS also plans to conduct filtering rule research to help develop criteria to determine which addresses on the MAF will be included in a MAF-based frame. This last research project will go beyond 2008.

Acknowledgement

The results of this paper come from several evaluation projects. As a leader of the FACHS program, I only contributed some of the ideas. The following Census Bureau employees participated in one or more of these projects that provided the results for this paper: Robert Colosi, Kelly Dixon, Tina Flanagan, Amanda Hakanson, Nancy Johnson, Timothy Kennel, Aliza Kwiat, Mei Li, Clifford Loudermilk, Joel Martin, and Julie McCune. Marjorie Corlett, a former Census employee, also made significant contribution to the planning and execution of the FACHS. Tina Flanagan, Mei Li, and Julie McCune produced the results for table 4, 5, and 6.

References

Barrett, D. et al, Census 2000 housing unit coverage study, Census 2000 Evaluation Report O.3., 2003

Flanagan, T. and Loudermilk, C., How quickly do permit addresses appear on the Delivery Sequence Files? Proceedings of the Join Statistical Meeting, 2006

Johnson, N. and Hakanson, A., Report on the Frame Assessment for Current Household Surveys: New Construction Study, U.S. Bureau Internal Memorandum, 2007

Kennel, T., Second analytical report for Frame Assessment for Current Household Surveys (FACHS) Area Frame Study, U.S. Bureau Internal Memorandum, 2007

Kennel, T. and Corlett, M., Comparing the US Census Bureau's Master Address File to the Survey of Income and Program Participation area listings. Proceedings of the Join Statistical Meeting, 2005

Tersine, T. et al, Estimating new construction lag in the Master Address File: results from the 1999 Permit Address Matching Study, Proceedings of the Join Statistical Meeting, 2001.

U. S. Census Bureau, Design and Methodology, Current Population Survey Technical Paper 66, 2006