Evaluating the Use of the Master Address File for the American Community Survey

Jennifer Olejownik & Cliff Loudermilk, U.S. Bureau of the Census, DSMD, Washington, DC 20233

KEY WORDS: Master Address File, Delivery Sequence Files, Coverage.

Abstract The period of 1996-98 marked the developmental phase of the American Community Survey (ACS), which collects housing and socio-economic data typically collected on the long form of the decennial census. The major strength of ACS is that it provides current annual data. Decennial sample data, in contrast, are less frequent because estimates are published once every ten years. The American Community Survey uses the Master Address File (MAF), a complete listing of all residential addresses in the country, for sample selection. When using the MAF to create a sample universe, several instances of probable undercoverage and overcoverage have been identified. This paper describes coverage concerns, which are relevant to ACS. The intent of this paper is to provide an overview of issues relating to the decision to use an "unfiltered" MAF for ACS.

Introduction The goal of this paper is to describe some of the unforeseen quirks adopted as a result of the criteria used by the ACS to process the MAF. While the intent is to describe some these phenomenon, the scope of the paper is limited in terms of discussing the magnitude of the effects of the criteria selected. Possible implications and ideas for future research are exploratory. Work groups have been created to investigate these issues. The figures included in this paper are not exhaustive, nor are they indicative of a comprehensive evaluation of the MAF.

The American Community Survey The ACS began in 1996 and was created to collect social and economic data traditionally obtained through the decennial census long form sample. The ACS was conceived in order to provide accurate data in a more timely fashion through a process known as continuous measurement, where data are collected and reported in a continual manner. The ACS collects standard demographic data such as income, marital status, ethnicity, housing, transportation, and occupation. The survey gathers data through mailout/mailback procedures, telephone and face-to-face interviews.

Data for decennial operations are gathered once every 10 years and its usefulness diminishes as the decade progresses. The ACS allows database users to feel more confident about quality since data are refreshed each year.

The Census Bureau is implementing ACS in several phases. In 1996, the demonstration period extracted data from four counties: Multnomah, OR; Brevard, FL; Rockland, NY; and Fulton, PA. The population of these areas was sampled at 15 percent and small governmental units, or areas that have their own functioning government and is less populous, were sampled at twice that rate since it is difficult to generate meaningful statistics in these areas. This design allows researchers to capture profiles of local areas or neighborhoods. The demonstration period expanded to nine counties in 1998. Intentionally, two of these counties overlapped with the 1998 Decennial Census Dress Rehearsal in order to understand the dynamics and complexities of conducting two different surveys at the same time and location.

With three year aggregates, the second phase of ACS implementation from 1999-2002 allows researchers to examine or compare differences between Census 2000 long form estimates and ACS estimates for 1999-2002. In this phase, additional counties are added to increase sample size. In 2000, the sample will increase to over 1,200 counties, indicating the third phase which expands representativeness and constructs the framework for national comparisons. As counties are gradually added each year, data from the ACS can be compared with data from the decennial long form for all states, large cities, and large substate areas.

In 2003, ACS will be sampling from every county in the U.S., which will signify the beginning of a nationwide continual sample of addresses each year. After five years of sampling 3 million addresses each year, the American Community Survey will be able to generate statistics for all 50,000 census tracts in the U.S., a special feature of the ACS. Until now, data have been unavailable for census tracts as blocks were the smallest unit available for analyses. Moreover, every community's demographic composition will be refreshed each year instead of once per decade.

The American Community Survey is unique for a variety of reasons. First, continuous measurement allows for small scale analyses, that is, neighborhoods or local areas can be explored just as metropolitan statistical areas have been studied in the past. Obtaining micro-level data has many advantages for small communities. Data of this kind are an excellent asset for making decisions for planning, constructing, or evaluating the effectiveness of public programs. Secondly, researchers can immediately uncover any

microscopic change in the configuration of data in a remote area. Because data are collected annually instead of once every ten years, ACS data are up-to-date and changes are easily recognized when comparing the data from year to year. Thirdly, data from the American Community Survey will be updated twice a year and it has the ability to accommodate specific customer needs since the survey can be augmented with supplemental questions related to special topics or interests. Lastly, the ACS is unique for yet another reason, it is the first survey to utilize the MAF for sample selection.

Master Address File The Master Address File is the Census Bureau's perpetual housing unit inventory. The vision of the Master Address File is two-fold. First, the MAF aims at improving the accuracy of censuses and surveys since it will document every living quarters in the U.S., therefore producing a stronger and more representative sample. Secondly, retaining an up-to-date repository of addresses embodies a certain cost-saving advantage. Because locating new construction is coordinated with U.S. Postal efforts, the Master Address File eventually will streamline some of the canvassing procedures used for enumeration which can be quite costly.

Although, the American Community Survey is the first survey to use the Master Address File (MAF) to select sample, decennial census is partially using the MAF for the Census 2000 operations. In the future, the MAF is expected to be integrated into all other statistical operations at the bureau. Developing a frame that incorporates every household or residential address in the U.S., however, is quite an extensive operation.

The MAF is comprised of two principle components, addresses that appeared on the 1990 Address Control File (ACF) and addresses from the Delivery Sequence File. The Address Control File is a listing or inventory of housing units contained within a specific geographic area for 1990 Census operations. Delivery Sequence Files (DSF) come from the Post Office and are lists of addresses which include a carrier route number. These files supply the Master Address File with updated addresses. Postal carriers provide residential information which is included to supplement or modify the MAF as needed. The Census Bureau receives updates from the Post Office twice a year, primarily in areas with traditional or city-style addresses (e.g. 8401 Main Street).

By using matching software, addresses from the ACF and DSF with ZIP+4 files are merged, standardized and unduplicated. These addresses are also linked with

TIGER, or the Topologically Integrated Geographic Encoding and Referencing system used by Geography Division, for referencing and producing consistent products for field operations. A census block is assigned when a unit is matched to TIGER, which provides a valuable link between databases. Conversely, the addresses that do not code to the ZIP+4 file (a list of valid street names and house number ranges with ZIP Code) are subjected to a TIGER geocoding process where addresses are linked to TIGER segments and then unduplicated.

Rural areas, or those with noncity-style addresses, are more challenging. Automated matching requires specific location information and, typically, noncity-style addresses are deficient in this regard. In these areas, mail is delivered to post office boxes rather than to physical or locatable addresses. Because carriers do not deliver mail in these rural areas, the DSF are not particularly helpful in rural areas as new neighborhoods often go undetected.

Constructing a Universe for Sampling from the Master Address File Because the MAF is the only source of its sampling frame for the ACS, constructing a universe from which to select sample is not a simple task. Using the MAF for selecting ACS sample is very different than using the MAF as a base for Census 2000 operations.

For decennial operations, MAF undercoverage, or the effect of not including every housing unit, is obviously undesirable, but it is not catastrophic. The MAF is a starting point for Census 2000. The MAF contains a fairly complete store of geocoded addresses, which will be updated and changed by an extensive series of field operations.

Where the MAF is most complete, Decennial will perform block canvassing everywhere in the nation. Block canvassing is a dependent listing check, or using a list of already documented addresses (such as the MAF) to compare what is observed on the ground.

Rural residences, as described earlier, are problematic for the ACS since the MAF is not an exhaustive list for sampling in these areas. Decennial operations use the MAF as a skeletal framework and fill in any non-city style coverage gaps through an independent listing, which involves creating an original list of addresses by writing down either the addresses or descriptions of living quarters.

Thus, undercoverage for the Master Address File is more of a concern for the American Community

Survey. Coverage gaps in the MAF are not fatal for Census 2000. The "holes" will be plugged by decennial field operations.

Comprehensive strategies have been devised to maintain the MAF. The Community Address Updating System, or CAUS, is the main initiative that will keep the MAF accurate, complete and up-to-date. Essentially, field representatives will visit county or other state, local, or tribal governmental offices to gather information about newly constructed housing units or other units that may be missing from the MAF. They will target areas based on national administrative

for the census. Therefore, they prefer to accept only geocoded MAF records, assuming that other addresses will be picked up in block canvassing, address listing or other field operations.

For ACS, dropping ungeocoded housing units is not necessarily the best course of action. First, one must decide what these ungeocoded records really represent. Are they "bad" addresses that do not exist? Are they duplicates of other records on the MAF that do not have geocodes? Or are they addresses that are "good" and not created elsewhere on the MAF?

Table 1.
Comparison of Housing Unit Counts

County	# valid housing units on MAF (excluding ungeocoded housing units)	# ungeocoded housing units	# housing units from 1990 Census
DeSoto, LA	7,894	3,540	10,919
Schuylkill, PA	61,581	4,941	66,457

sources to determine where the MAF needs to be improved. Listing operations will be conducted to verify and correctly locate address modifications. Until CAUS is implemented, the ACS will rely exclusively on the MAF and will be deficient wherever the MAF misses housing units.

An important question in creating a universe from the MAF is "Which piece of the MAF should be used?" The MAF itself is a very large database of housing units from many different, overlapping sources. Included are both residential and nonresidential addresses, duplicates, ungeocoded addresses, demolished units, vacant units, etc. Users of the MAF must think about a way to subset or filter the MAF in order to get a product that suits their needs.

Both Decennial and ACS have customized criteria to process the MAF, but they differ in important ways. For instance, Decennial requests that the Geography Division remove all ungeocoded records before delivering their MAF extract (called the "DMAF"). Ungeocoded records are MAF addresses that could not be assigned a census tract and block because they did not match to a line segment in TIGER. These addresses, whether "*good*" addresses or not, are not used by Decennial. They must assign every address to a tract and block in order to do meaningful tabulations

The answer is probably that the ungeocoded housing units are a mix of these situations. Retaining ungeocoded records in the universe may fill coverage gaps, but it could introduce double coverage. ACS needs to evaluate the overall effect of keeping ungeocoded records in its universe.

Certain MAF counts suggest ungeocoded housing units can be important to coverage. For example, consider the housing unit counts from the MAFs delivered in August 1998:

Refer to Table I.

The MAF counts (excluding ungeocoded housing units) are considerably below the 1990 Census counts, indicating potentially serious undercoverage. Notice, however, the counts of ungeocoded records seem to plug the gaps very nicely.

While suggestive, this may also be circumstantial. Two counties, DeSoto, LA and Schuylkill, PA were studied.Ungeocoded addresses were randomly selected and checked against several sources—the U.S. Postal Service website, Maps-On-Us, and Mapquest. Almost all of the addresses were locatable in DeSoto and none of the addresses were duplicates of any of the geocoded records on the MAF. Also, similar results were found

for Schuylkill county. For DeSoto, keeping the ungeocoded records would avoid some major undercoverage.

Hillburn Village is a small incorporated place in the southeastern corner of the county and is located along the New York/New Jersey border. Hillburn Village is one of the first areas where "holes" in the MAF were observed. These "holes" or pockets of missing data, are generally indicative of a coverage problem which, in

Table II: MAF Counts in Hillburn Village, Rockland County, NY		
ACS Sample Year	Number of Valid HUs on MAF	Change from 1990 Census HU Count
1996	137	-55.1%
1997	69	-77.4%
1998	47	-84.6%
1999	90	-70.5%

Extensive research has not been conducted on ungeocoded records for the ACS. Although, research from ACS staff suggests that ungeocoded records are typically converted to geocoded records as time lags. However, given the strong possibility that dropping the ungeocoded records would cause undercoverage, the decision was made to keep all ungeocoded records in the ACS universe because undercoverage is a more serious problem for ACS than overcoverage.

In the pre-census period, the ungeocoded records are a greater issue for the ACS than they are likely to be in the immediate post-census period, when the MAF should be in good shape (with a high degree of geocoding) because of decennial address updating operations. But there will always be ungeocoded records introduced onto the MAF, and the numbers will likely increase as time elapses after the Census 2000. Additional research into how ungeocoded records affect MAF coverage will help ACS and any other demographic surveys that wish to use the MAF as an exclusive frame for sampling.

Example of a Coverage Issue: A number of problems have been identified as a result of requesting ungeocoded units for the ACS. For the purpose of this paper, one problem is described in detail. Coverage implications are discussed and suggestions are made for future research.

Hillburn, NY

Description of the Problem: Rockland County, New York, has been a comparison site for ACS since 1996.

turn, could affect the accuracy of sampling and data quality. On every Rockland MAF received, the count of housing units in Hillburn has been consistently and significantly below the count of 305 housing units (HUs) reported in the 1990 Census:

Refer to Table II.

Except in a few rare circumstances, the number of housing units in an incorporated place should not be significantly lower than the count from the 1990 Census. Such a difference probably represents MAF undercoverage. It is unclear as to why the records on the 1996 Master Address File only contained some of the city-style addresses located on the 1990 Address Control File. Furthermore, why did addresses on the MAF in subsequent years have even fewer addresses than the 1996 MAF for the same area?

While Hillburn was the first place to exhibit such a discrepancy, persistent undercounts have also been discovered for other incorporated places in ACS comparison counties.

Refer to Table III.

All of the areas where undercounts were found are small incorporated places that are believed to contain "good" addresses. While it is possible that the coverage gaps only occur in small places, it may be that the gaps also exist in larger places but are masked by growth in these areas.

Most Hillburn residents receiving mail use post office boxes since the U.S. Post Office in Hillburn does not provide residential delivery. Post office boxes, however, cannot be coded by geography and are not included on the MAF. The addresses located in Hillburn were assigned city-style addresses as a result of the Emergency 911 system that was implemented many years ago. This complicated matters for the creation of the MAF since many of the post office boxes in the village of Hillburn were omitted simply because they could not be geocoded.

within Hillburn, 259 of these addresses were geocoded to blocks in other tracts within the county. Using Internet sites, it was discovered that several of these addresses in different tracts had similar characteristics when compared to the addresses in Hillburn. Many of these streets share a common name, such as First Street or Second Street, and can be found in almost every county in the U.S.

Refer to Table IV.

		# HUs in 1990 Census	HU Counts from MAFs		
State	Place Name		1996	1997	1998
NY	Hillburn village	305	137	69	47
OH	Brice village	49	---	6	22
SC	Eastover town	355	---	---	148
TX	Pleak village	250	---	182	193

Table III:
Small Incorporated Places in 1998 ACS Test Sites
With Significantly Fewer HUs on MAF
Than Were Reported in 1990 Census

The records that were dropped or missed from previous years did not appear on the most current MAF. These addresses were verified, that is, the physical location of each address was determined. The majority of the addresses were locatable using various Internet sites. The addresses did not appear on the U.S. postal website, suggesting that the DSF did not pick up these records, possibly because they were P.O. boxes.

Furthermore, most street segments were located in the TIGER database, so the addresses from ACF should have been geocoded without any problem. Although this information is helpful, it does little to clarify why the records from the Address Control File were dropped at a decreasing rate each year from the MAF.

Tracking addresses from the 1990 Address Control File revealed that some blocks previously geocoded inside Hillburn Village were geocoded to blocks in other tracts within the county. For example, of the 305 units

It is unclear what prompted such a change since there were no changes to the basic street address and ZIP code.

The Hillburn problem demonstrates one of the most pressing concerns to date in regard to use of the Master Address File. The most obvious and immediate problem affects the American Community Survey, because coverage gaps undermine sampling strategy. For future users of the Master Address File, the situation in Hillburn could create additional complications as this problem might be indicative of similar occurrences in other counties.

Coverage Implications: For those small incorporated places within ACS comparison sites that have coverage gaps similar to those observed in Hillburn, a problem exists for the ACS. Unlike the current demographic surveys which do not publish estimates for governmental units, the ACS publishes reliable estimates for small governmental units within counties.

Table IV: Results of Matching the 1990 ACFIDs in Hillburn, NY to the 1999 ACS MAF		
Outcome	**# of housing units**	**% of total housing units**
ACFID appears on the MAF and is geocoded correctly to Hillburn	90	29.5%
ACFID appears on the MAF and is geocoded incorrectly to a non-Hillburn tract in Rockland County	141	46.2%
ACFID appears on the MAF and is geocoded incorrectly to Orange County, NY	28	9.2%
ACFID does not appear on the MAF	46	15.1%
Total Records from 1990 ACF in Hillburn	305	100.0%

The undercoverage problem would have been more problematic for current demographic surveys. The "missing" units that are merely geocoded in error to a different part of the county would not present a very serious problem--there is no county-level undercoverage. The actual undercoverage would consist of the units that are geocoded outside the county or are missing altogether.

It is expected that these Hillburn-type "holes" in the MAF will be filled once the MAFs are updated with block canvassing results from the Census 2000.

Ideas for Further Research Since block canvassing will presumably plug the Hillburn-type "holes", there is probably no advantage pursuing what happened to the dropped 1990 ACF records. These units should be represented on the MAFs coming out of the Census 2000.

However, it may be beneficial to investigate whether the geocoding error rate on the MAF is significantly worse in small incorporated places than in other areas. The result could have a negative impact on the ability of ACS to provide accurate data for small governmental units.

Conclusion: While the Hillburn problem symbolizes undercoverage, other problems, such as bad zip code clusters and false duplicates, also persist. Due to the confines of formatting, some of these problems have been researched and are documented in a previous version of this paper. It can be concluded that the ACS will continue to adopt some coverage quirks when requesting ungeocoded housing units. In the years to come, coverage issues can be explored and better understood as sample size increases for the ACS. Until additional research is conducted, results from MAF evaluation groups will be eagerly anticipated to clarify issues beyond the scope of this paper.

Disclaimer: This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.