



UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

May 24, 2012

2012 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES
#ACS12-RER-20

DSSD 2012 AMERICAN COMMUNITY SURVEY MEMORANDUM SERIES #ACS12-UC-05

MEMORANDUM FOR ACS Research and Evaluation Steering Committee

From: David C. Whitford/*Signed*/
Chief, Decennial Statistical Studies Division

Prepared by: Dianne Aubuchon
Community Address Updating Systems Branch
Decennial Statistical Studies Division

Subject: Use of Ungeocoded Delivery Sequence File Records in the American
Community Survey Sampling Frame

Attached is the final American Community Survey Research and Evaluation report for the Use of Ungeocoded Delivery Sequence File Records in the American Community Survey Sampling Frame. The current method to determine the areas where the American Community Survey excludes otherwise valid United States Postal Service Delivery Sequence File records because of the risk of duplication depends on the Census 2000 blueline status. The American Community Survey wanted to redefine this method with more up-to-date information.

If you have any questions about this report, please contact Larry Bates (301) 763-5926 or Dianne Aubuchon (301) 763-7706.

Attachment

cc:

ACS Research and Evaluation Team
Tony Tersine (DSSD)
Larry Bates
Cliff Loudermilk (DSMD)
Matt Zimolzak (ACSO)

Use of Ungeocoded Delivery Sequence File Records in the American Community Survey Sampling Frame

FINAL REPORT

Executive Summary

The American Community Survey has been defining the area in which we use ungeocoded Delivery Sequence File records using information from Census 2000. We wanted to update this criterion to rely on more recent information. We evaluated criteria based upon the percentage of city-style addresses in a county.

We recommended that the American Community Survey use ungeocoded United States Postal Service Delivery Sequence File records in counties with at least 80 percent city-style addresses¹. We found that only about 1.5 percent of the ungeocoded records would have a higher risk of duplication; this would be only about 23,600 of the approximately 1.5 million ungeocoded records in the 2011 American Community Survey supplemental sampling frame. With this criterion we will be using ungeocoded records in 2,865 counties, with a net gain of 33,000 records that we did not previously use.

This recommendation was implemented beginning in the 2012 Main Phase Unit Frame Universe creation.

1. Introduction

The American Community Survey (ACS) excludes otherwise valid new United States Postal Service (USPS) Delivery Sequence File² (DSF) records in some counties because of the possibility that those DSF addresses may duplicate existing noncity-style³ addresses. For example, ACS only uses geocoded⁴ DSF addresses in blocks that, for the most part, contain 80 percent or more city-style addresses or blocks with no residential addresses. Areas that typically contain a high percentage of noncity-style addresses are considered to be inside the Duplication Zone (DZ), a term given to areas where ACS excludes otherwise valid new DSF records because of the risk of duplication. Ungeocoded addresses cannot be evaluated at a block level, because they are not assigned to a block. The only geographic information available for them is the county.

Between 2000 and 2011, ACS used ungeocoded DSF addresses in all 2,126 counties that were either completely or partially inside the “blueline” (i.e., mailout/mailback) for the 2000 Census. Beginning in 2008, ACS began using ungeocoded DSF addresses in a select group of 69 counties that had no mailout/mailback areas in Census 2000. These counties had a high percentage of city-style addresses (over 95 percent), or a combination of a high percentage of city-style addresses (90-95 percent), some coverage deficiency (at least 10 percent undercoverage), and some growth between 2000 and 2007 (≥ 10 percent, per Population Estimates Program (PEP) housing unit (HU) estimates). Since each county had a high

¹ An address is city-style if it has a house number and street name; otherwise it is non city-style.

² A file containing all mailing addresses serviced by the USPS.

³ Examples of non-city-style addresses are PO Boxes, rural routes, and description only addresses

⁴ When an address is assigned to a location that is identified by one or more geographic codes such as a block code.

percentage of city-style addresses there was little chance that the ungecoded DSF records were duplicates of existing noncity-style addresses.

The purpose of this research is to redefine the DZ using more current information than the Census 2000 blueline status. We will determine whether to include or exclude all ungecoded DSF addresses for an entire county.

2. Methodology

In determining the counties where ungecoded DSF records can be used, ACS evaluated two criteria in defining the duplication zone:

- The percentage of records with city-style addresses in each county, and
- The Address Characteristic Type (ACT) code⁵ of the blocks where previously ungecoded DSF addresses were placed by any Census operation

To compute the percentage of city-style addresses, we used all valid, geocoded records on the most current ACS extract⁶.

To evaluate where ungecoded records were found, we included all DSF records on the Master Address File (MAF) /Topologically Integrated Geographic Encoding and Referencing (TIGER) database (MTdb) that were ungecoded prior to receiving any Local Update of Census Addresses (LUCA) updates for the 2010 Census. We examined the ACT codes from the most recent ACS extract for those ungecoded addresses that were later geocoded during the 2010 Census.

We are assuming that the blocks in which new ungecoded units will tend to be placed are similar in characteristics to the blocks where previously ungecoded units were found. We used this information to assess the likelihood that new ungecoded DSF units exist in blocks where ACS currently uses the DSF; which by definition are blocks where there is less likelihood of duplication.

Since ACS uses geocoded DSF records in blocks where there are 80 percent or more city-style addresses, we evaluated an analogous criterion of 80 percent or more city-style addresses in a county for ungecoded DSF records.

To help us in the evaluation we defined several variables:

- percent city-style =
$$\frac{\text{total valid geocoded city-style addresses}}{\text{total valid geocoded addresses}}$$

⁵ A block level categorization of addresses based on the percent of city-style and DSF addresses in a block; see Attachment A.

⁶ At the time of analysis, this was the 2011 supplemental edited MAF extract

- Old DZ – based on the 2000 blueline status plus the 69 counties discussed above, indicates whether a county is inside or outside the current duplication zone. This is the duplication zone used for the 2011 supplemental phase (s11).
- New DZ – based on our chosen new duplication zone criterion (80 percent or more city-style addresses), indicates whether a county is inside or outside this new duplication zone

We also looked at how many records we would gain and lose due to this change, calculating the following:

- Gained – total number of invalid s11 ungeocoded DSF addresses (i.e., in the old DZ) that would have been valid (i.e., outside the new DZ) if we used the new duplication zone definition in s11.
- Lost – total number of valid s11 ungeocoded DSF addresses we would have lost if we used the new duplication zone definition in s11 (i.e., in the new DZ).
- percent gain in a county = $\frac{\textit{gained}}{\textit{total valid addresses}}$
- percent loss in a county = $\frac{\textit{lost}}{\textit{total valid addresses}}$

When calculating these we used a revised universe. The DSF start date⁷ used in 2011 universe creation was changed to reflect the new DSF start date we will use for 2012 universe creation. We applied this new DSF start date to the 2011 universe to more accurately depict what effect this will have on the 2012 universe.

3. Limitations

We encountered the following limitations:

- We needed to make a decision in time to apply the new duplication zone criteria to the filter for the main 2012 (M12) ACS sample. Due to the timing, we had to use Address Canvassing⁸ (AdCan) results instead of final Census results.
- We wanted to separate completely from data that depended upon Census 2000. As a result, we were limited as to the amount of data we had to work with to create the new criteria.
- Because of time and resource constraints we decided to look for one rule to cover all counties instead of getting into county-by-county decisions. So while we looked for an overall good fit, the resulting rule may suit some counties more so than others.

⁷ DSF start date is used in defining what records are considered to be post-Census DSF adds. It is changing so that it will define post 2010 Census DSF adds instead of post 2000 DSF Census adds.

⁸ A country-wide operation that updated address lists in preparation for the 2010 Census

4. Results

The goal of the research was to determine in which counties ACS can use ungeocoded DSF records while minimizing the chance that the ungeocoded records are duplicates of existing non-city style addresses. We wanted to select a cutoff point for the percentage of city-style addresses in a county that could be used to define the DZ; one with a low risk of duplication that did not decrease the coverage rate.

We calculated the following tallies to look at the impact on the frame using the new DZ criterion versus the old DZ criterion, in terms of the number of counties (total, gained, and lost) and the number of units (total, gained, and lost) to evaluate the coverage impact of the new DZ.

- Counties inside the new DZ – subdivided by old (s11) DZ status
- Counties outside the new DZ – subdivided by old (s11) DZ status
- Change in valid address counts from the old to new DZ definition
- The newly geocoded records⁹ – separated by their new and old DZ status

5.1 Coverage: Gains and Losses

Table 1 shows the placement of counties between the old DZ and the new duplication zone. We can see that under a new 80 percent city-style address definition, there would be 2,865 counties outside of the DZ and 278 that would be inside. There are 730 counties that change from being counties where we do not currently use ungeocoded DSF records to counties where we would use them. We would stop using ungeocoded DSF records in 60 counties where we previously would have used them. There is no impact on the remaining 2,353 counties. From a county coverage standpoint this would be good because we would have a net gain in coverage of 670 counties.

Table 1: Counties by Old and New Duplication Zone Status

Old DZ Status ¹⁰	New DZ Status ¹¹		
	IN	OUT	Total
IN	218	730	948
OUT	60	2,135	2,195
Total	278	2,865	3,143

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

Table 2 shows the number of records that would be gained and lost by redefining the duplication zone. In the 730 new counties that would be outside the DZ, we would have gained 43,130 records in s11. The largest percent increase in any one county was 17.3

⁹ Records that were ungeocoded in 2009, and are now geocoded in 2011.

¹⁰ This is the status of a unit based on the DZ definition used in 2011.

¹¹ This is the status of a unit based on the new DZ definition that was implemented for 2012.

percent. After some research into this county we discovered evidence of new housing developments, which are the type of records ACS was hoping to include in the frame by making a change to the duplication zone. Additionally, the county has a very high rate of city-style addresses, and with no evidence of address conversions in the county there is low risk that the new HUs are duplicating existing records.

In the 60 counties that we would no longer use, we would lose 10,040 records. The largest percent loss in any one county was 7.4 percent. We are willing to accept a small loss in some counties where there is a higher risk of duplication in order to improve coverage in counties where there is a lower risk of duplication.

Table 2: Change in S11 Valid Counts Using New Duplication Zone Status

Old DZ Status	New DZ Status		
	IN	OUT	Max Gain/Loss
IN	0	43,130	17.3%
OUT	-10,040	0	- 7.4%

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

5.2 Risk of Duplication: Newly Geocoded Records

To assess the risk that these ungeocoded DSF units could be duplicates, we looked at the newly geocoded records in the 2,865 counties outside of the new DZ. We looked at the blocks where these units were placed, and whether those blocks were inside or outside of the 2011 DZ.

Table 3 shows where records that were ungeocoded prior to AdCan, and found by AdCan, were placed. The table is limited to the 2,865 counties outside the new DZ. Approximately 1.5 percent of the newly geocoded records were found in blocks where geocoded DSF records are not used. This indicates only about 1.5 percent of the ungeocoded units would have a higher risk of duplication. For s11, this would have meant that about 23,600 of the approximately 1.5 million ungeocoded records eligible for sample for ACS would have had a higher risk of duplication. As this affects a relatively small number of units, we are willing to accept this risk.

Table 3: Placement of Newly Geocoded Records (Counties Outside the new Duplication Zone)

Placed in blocks where		Total
Geocoded DSF records not used in 2011	Geocoded DSF records used in 2011	
116,615 1.5%	7,722,091 98.5%	7,838,706

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

In Table 4 we see the newly geocoded units that are in the 278 counties inside of the new duplication zone. Almost 40 percent of these were geocoded to blocks where

geocoded DSF records are not used. This confirms that the ungeocoded records in those counties are more likely to be located in areas where there is a higher risk that the DSF records may duplicate existing records on the frame.

Table 4: Placement of Newly Geocoded Records (Counties Inside the new Duplication Zone)

Placed in blocks where		
Geocoded DSF records not used in 2011	Geocoded DSF records used in 2011	Total
25,900 39.2%	40,168 60.8%	66,068

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

We then took a closer look at the newly geocoded records separated by those that we will gain, those that we will lose, those that will remain outside of the DZ, and those that will remain inside of the DZ.

Counties removed from the duplication zone

In Table 5 we show the newly geocoded records from the 730 counties that we would gain from using the new DZ. Almost 89 percent of the newly geocoded records that are outside the new DZ were placed in blocks where we used geocoded DSF records in 2011. This indicates that it would have been relatively safe from a duplicate standpoint to use these records all along.

Table 5: Placement of Newly Geocoded Records for the 730 counties that we gain

Placed in blocks where		
Geocoded DSF records not used in 2011	Geocoded DSF records used in 2011	Total
41,355 11.4%	321,049 88.6%	362,404

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

Counties that will remain outside of the duplication zone

In Table 6 we show the newly geocoded records from the 2,135 counties that were outside the DZ in 2011, and would stay outside of the DZ using the new definition. Only one percent of the newly geocoded records that are outside of the new duplication zone were placed in blocks where we did not use ungeocoded DSF records in 2011; this indicates that overall we were doing well with these areas and should continue using them.

Table 6: Placement of Newly Geocoded Records for the 2,135 counties that remain outside the Duplication Zone

Placed in blocks where		
Geocoded DSF records not used in 2011	Geocoded DSF records used in 2011	Total
198,711 1.0%	7,277,591 99.0%	7,476,302

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

Counties added to the duplication zone

Table 7 shows the newly geocoded records from the 60 counties where we would no longer use ungeocoded DSF records under the new duplication zone. Just over 30 percent of the newly geocoded records inside the new duplication zone were placed in blocks where we did not use ungeocoded DSF records in 2011. This is a good indicator that we should be excluding ungeocoded DSF records in these areas.

Table 7: Placement of Newly Geocoded Records for the 60 counties that we lose

Placed in blocks where		
Geocoded DSF records not used in 2011	Geocoded DSF records used in 2011	Total
7,562 31.3%	16,581 68.7%	24,143

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

Counties that will remain inside of the duplication zone

The newly geocoded records from the 218 counties that remained inside the DZ can be seen in Table 8. Almost 44 percent of these units were placed in blocks where we did not use ungeocoded DSF records in 2011, indicating that we should continue to exclude these areas.

Table 8: Placement of Newly Geocoded Records for the 218 counties that remain inside the Duplication Zone

Placed in blocks where		
Geocoded DSF records not used in 2011	Geocoded DSF records used in 2011	Total
18,338 43.8%	23,587 56.2%	41,925

Source: U.S. Census Bureau, American Community Survey 2011 Supplemental Sample Frame

5. Summary

We recommend using ungeocoded DSF records in counties with at least 80 percent city-style addresses. Counties with less than 80 percent city-style addresses would be inside the DZ, and we would not use the ungeocoded DSF addresses in those counties.

With this new criterion coverage will be improved. We will be using ungeocoded records in 2,865 counties. While we will lose 60 counties where we previously used ungeocoded records, we will be gaining 730 counties. This will give us a net gain of about 33,000 records that we did not previously use. Based upon the research into the newly geocoded records, the records we will be adding have a greater chance of being good records for ACS.

We found that among the newly geocoded records in the counties where we will keep using ungeocoded records, less than one percent were placed in blocks where we did not use geocoded DSF records in 2011. Additionally, almost 89 percent of the previously ungeocoded records in the counties where we will now use ungeocoded records were placed in blocks where we used geocoded DSF records in 2011. Both of these indicate that the risk of duplication is low in the counties that will be outside of the DZ under the new definition.

We implemented this recommendation in the ACS 2012 Main Phase Unit Frame Universe creation. Going forward, ACS will recalculate the percent of city-style addresses in a county each year; and new DZ statuses will be assigned where necessary.

6. References

Bates, Lawrence. (2010), "Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey (2011 Supplemental Phase)," DSSD American Community Survey Memorandum Series #ACS11-UC-4, December 29,2010.

Table A1. ACT Codes and Descriptions

Code	Description
C1	All city-style addresses, none with a DSF source
C2	All city-style addresses, some with a DSF source
C3	All city-style addresses, all with a DSF source
R1	All RR/box addresses, none with a DSF source
R2	All RR/box addresses, some with a DSF source
R3	All RR/box addresses, all with a DSF source
P1	All PO box addresses, none with a DSF source
P2	All PO box addresses, some with a DSF source
P3	All PO box addresses, all with a DSF source
D1	All description addresses, none with a DSF source
M1	Mixed city-style and noncity-style addresses, none have a DSF source
MA	95% - 99.99% city-style addresses, some have a DSF source
MB	90% - 94.99% city-style addresses, some have a DSF source
MC	85% - 89.99% city-style addresses, some have a DSF source
MD	80% - 84.99% city-style addresses, some have a DSF source
ME	75% - 79.99% city-style addresses, some have a DSF source
MF	70% - 74.99% city-style addresses, some have a DSF source
MG	0.00% - 69.99% city-style addresses, some have a DSF source
M3	Mixed city-style and noncity-style addresses, all have a DSF source
N1	All assorted noncity-style addresses, none have a DSF source
N2	All assorted noncity-style addresses, some have a DSF source
N3	All assorted noncity-style addresses, all have a DSF source
B1	All nonresidential addresses, none have a DSF source
B2	All nonresidential addresses, some have a DSF source
B3	All nonresidential addresses, all have a DSF source
Z0	No address