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

DSSD 2011 AMERICAN COMMUNTY SURVEY MEMORANDUM SERIES ACS11-R-06

MEMORANDUM FOR ACS Research and Evaluation Team

From: David C. Whitford ***signed 6/22/11***
Chief, Decennial Statistical Studies Division

Prepared by: Alfredo Navarro
Assistant Division Chief, ACS Statistical Design
Decennial Statistical Studies Division

Karen E. King
Chief, ACS Variance Estimation and Statistical Support Branch
Decennial Statistical Studies Division

Subject: Simulating the Effect of Filtering on 5-Year Estimates

The document attached gives some results of the third and final assessment requested by Director Steve H. Murdock in the Fall of 2008. A series of assessments were proposed to look at the potential quality or reliability of ACS 5-year data. The only 5-year ACS estimates available at the time were from the Multiyear Estimate Study conducted in 2006, which included data for 34 of the ACS test counties and covered the period from 1999 through 2005. The assumption was that the patterns of quality seen in these data would be consistent with the quality of the first 5-year estimates released in 2010. A report titled "Quality Rating Classification of 5-year ACS MYES Estimates by Population Size Groupings and in Comparison with Census 2000 Long Form Estimates" summarizes results of the other of the assessments.

If you have questions regarding it, please contact Karen King at (301) 763-1974.

Attachment

cc:
A. Tersine (DSSD)
J. Hartman
J. Karberg
M. Starsinic
D. Griffin (ACSO)
S. Baumgardner
J. G. Robinson

American Community Survey Research and Evaluation Program

June 22, 2011

Simulating the Effect of Filtering on 5-Year Estimates

FINAL REPORT

KAREN KING
ALFREDO NAVARRO
DECENNIAL STATISTICAL STUDIES DIVISION

Simulating the Effect of Filtering on 5-Year Estimates

by Karen King and Alfredo Navarro
Decennial Statistical Studies Division

Introduction

In 2010, the ACS released all 5-year estimates for census tracts, census block groups, and small governmental units without any data quality filtering. However, concerns were raised about the reliability of many of the estimates likely to be included in the data products in 2008. Based on a requested by Director Steve H. Murdock in the Fall of 2008, a series of assessments were proposed to look at the quality or reliability of ACS 5-year data. The only 5-year ACS estimates available at the time were from the Multiyear Estimate Study conducted in 2006, which included data for 34 of the ACS test counties and covered the period from 1999 through 2005. The assumption was that the patterns of quality seen in these data would be consistent with the quality of the first 5-year estimates released in 2010. A report titled "Quality Rating Classification of 5-year ACS MYES Estimates by Population Size Groupings and in Comparison with Census 2000 Long Form Estimates" summarizing results of two of the assessments undertaken that were presented to the director on November 20, 2008. This report gives some results of the third and final assessment. This assessment focused on what would happen if standard ACS filtering rules were applied to the 5-year MYES data and a second set of estimates generated using adjusted 2005 – 2007 3-year weighted data to simulate the 5-year data.[1]

Overview of Filtering Rules

Data quality filtering is applied to all 1-year and 3-year ACS data products in an attempt to reduce the number of low-quality estimates that are released. The coefficient of variation (CV) is the measure of quality used in filtering rules. The CV is the standard error of the estimates divided by the estimate itself.

For ACS base (or detailed) tables, filtering is applied by finding the median CV of all detailed lines in a table, excluding total and subtotal lines. If the median CV is higher than 0.61, then the entire table is "filtered out", or not published. Entire tables are either in or out; there is not partial table filtering of base tables. Zero estimates have undefined CV so are assigned a CV of 1 when calculating the median CV. This is under the assumption that a zero estimate is unstable. Setting the CV to 1 in the filtering rule increases the chances of the tables failing the filtering rules.

Filtering for ACS data profiles is based on the filtering for the base tables described above. Each estimate in the profile is "sourced" from estimates one or more base tables. If the base table providing the profile estimate fails filtering, then the estimate is not published in the profile and it receives an "N" value on American Fact Finder. If a profile estimate is sourced

[1] The 2005-2007 3-year data is closer to the sample design of the 5-year data than the 5-year MYES. We wanted to see if the results were similar.

from more than one table, then all sourced tables must pass filtering, or the profile estimate is filtered out.

Results and Narrative

Table 7a shows the impact of the standard filtering, when applied to all 10,906 MYES geographic areas. After applying the filtering to the 139 tables produced across all areas, table 7a shows that for areas under 1K and between 1K and 5K, about 75% and 60% respectively of the base tables would be filtered out (not published). Table 7b shows roughly similar results for the simulated 5-year data. Large number of tables are not released, but as the population size increases, the percent filtered out drops until only about 6% of tables would be filtered out among areas of 65K or greater.

Tables 8a and 8b displays what happens to the estimates in the base tables analyzed in table 7 when the standard filtering (using the 0.61 cutoff) is and isn't applied. Here we see the resulting CV distribution of the estimates. With no filtering, more than 60% of estimates for areas with population less than 1K are zeroes, and an additional 15% to 20% have CVs above 0.61; only about 8% to 10% of the estimates have CVs less than 0.3. As the population size increases, the percent of zero estimates and extreme CVs decreases, until combined they are about 16% for areas above 65K for both sets of estimates. After filtering is applied, the distributions look much better, but the number of estimates filtered out is very high. Below 1K, over 90% of the estimates would be filtered out, and even at 10K-20K, about half of the estimates would be filtered out. Filtering improves the distributions but is not a cure-all - even for areas 65K and higher, about 10 percent of the estimates that would be published (after filtering) still have an extremely high CV or are zero.

Tables 9a and 9b we look at the profile estimates instead of the estimates from the base tables. Distributions start better, but still filtering would remove about 87% for MYES estimates and 77% for the simulation estimates in areas below 1K.

Tables 10a and 10b look at 16 profile estimates (lines). For each estimate, it shows the number of geographic areas for which the estimate would be published, the percent of areas where the estimate would be filtered out, and the resulting CV distribution for estimates in *published areas only*. For example, the results for Age 21+ and Age 65+ for MYES estimates (Table 10a) are below:

		% Not		CV	CV		
	# Published	Published	CV<0.1	0.1-0.3	0.3-0.61	CV>0.61	Est=0
SEX AND AGE							
21 years and over	4,249	61.0%	93.9%	6.1%	0.0%	0.0%	0.0%
65 years and over	4,249	61.0%	29.5%	65.9%	4.4%	0.1%	0.0%

For both these estimates, their value would be published for 4,249 areas, but they would be filtered out for 10,906 - 4,249 = 6,657 (61.0%) of all MYES areas. For the areas where the estimate would be published, the CV for both estimates is good. About 100% of the 21+ estimates have CVs less than 0.3, and about 95% of the 65+ estimates have CVs less than 0.3. Simulated estimates have similar results.

Note that several of the MYES estimates in Table 10a, including "Not in Labor Force", "Carpooled", and "Utility gas" would be filtered out in more than 90% of all areas. In Table 10b, for these simulated estimates, the percent not published were lower 83%, 68%, and 80% respectively. In each of these cases, the estimate is taken from a table with a large proportion of estimates that could be small or zero, leading to a high chance of failing the filtering rules.

Tables 11a and 11b go one step further with the data profile filtering and applies a rule that has never needed to have been applied to published ACS data. A secondary filtering rule for data profiles (and other products derived from base tables) states that if more than half of the individual lines in a profile are filtered out, then the entire profile is not published. Although this rule has so far never triggered for any published data product, it would be very relevant here. All profiles for MYES areas below 1K, and nearly all for those below 5K, would be filtered out by this rule. For the simulation estimates, below 1K the results are the same, but for below 5K about 70% would be filtered out.

This assessment shows that applying current ACS filtering rules would prevent a vast amount of data from being published. This is due in part to the large number of small and zero estimates in the smallest areas.

Table 7a

Impact of Filtering on Number of Base Tables Published, by Size of Area,

Using Current and Alternate Filtering Rules

All MYES Geographic Areas

Pop Range	Total # of Tables	# Published	# Not Published	% Not Published	% Not Pub (0.50)	% Not Pub (0.40)
< 1K	441,109	108,203	332,906	75.5%	81.9%	85.5%
1K-5K	830,108	341,939	488,169	58.8%	68.3%	75.0%
5K-10K	139,695	87,070	52,625	37.7%	47.6%	57.1%
10K-20K	35,584	26,344	9,240	26.0%	32.9%	41.3%
20K-65K	45,592	38,762	6,830	15.0%	19.4%	24.9%
> 65K	23,769	22,359	1,410	5.9%	7.7%	10.3%

Table 7b

Impact of Filtering on Number of Base Tables Published, by Size of Area,

Using Current and Alternate Filtering Rules

5-Year Simulation Geographic Areas

Pop Range	Total # of Tables	# Published	# Not Published	% Not Published	% Not Pub (0.50)	% Not Pub (0.40)
< 1K	1,929,553	588,865	1,340,688	69.5%	77.2%	82.4%
1K-5K	1,423,916	766,115	657,801	46.2%	56.3%	65.8%
5K-10K	392,953	252,657	140,296	35.7%	45.4%	55.0%
10K-20K	272,579	205,922	66,657	24.5%	31.6%	40.1%
20K-65K	276,610	236,599	40,011	14.5%	19.3%	24.9%
> 65K	92,574	87,179	5,395	5.8%	7.9%	10.7%

Table 8a

CV Distribution of Base Table Estimates, by Size of Area,
Without and With Filtering
All MYES Geographic Areas

No Filtering

			CV	CV		
Pop Range	# Est	CV<0.1	0.1-0.3	0.3-0.61	CV>0.61	Est=0
< 1K	9,452,925	0.4%	7.5%	11.0%	21.0%	60.2%
1K-5K	17,790,588	3.3%	11.9%	16.4%	22.9%	45.6%
5K-10K	2,993,895	9.0%	16.8%	21.5%	20.9%	31.8%
10K-20K	762,624	13.4%	23.7%	21.2%	17.6%	24.1%
20K-65K	977,112	20.5%	30.4%	19.0%	14.1%	15.9%
> 65K	509,409	39.9%	30.3%	13.6%	8.4%	7.9%

Standard Filtering

			CV	CV			% Filtered
Pop Range	# Est	CV<0.1	0.1-0.3	0.3-0.61	CV>0.61	Est=0	Out
< 1K	427,942	5.3%	41.0%	31.5%	12.2%	10.0%	95.5%
1K-5K	2,667,708	13.0%	32.8%	31.5%	14.7%	8.1%	85.0%
5K-10K	1,019,768	18.5%	31.6%	29.9%	13.5%	6.5%	65.9%
10K-20K	367,071	22.6%	37.3%	24.3%	10.2%	5.7%	51.9%
20K-65K	647,393	28.5%	39.2%	18.8%	8.2%	5.3%	33.7%
> 65K	447,540	44.7%	32.6%	12.8%	5.9%	3.9%	12.1%

Table 8b

CV Distribution of Base Table Estimates, by Size of Area,
Without and With Filtering
5-Year Simulation Geographic Areas

No Filtering

			CV	CV		
Pop Range	# Est	CV<0.1	0.1-0.3	0.3-0.61	CV>0.61	Est=0
< 1K	40,534,734	0.7%	9.5%	14.1%	14.8%	60.9%
1K-5K	29,932,968	4.9%	15.8%	20.5%	14.2%	44.6%
5K-10K	8,260,494	8.5%	18.9%	21.7%	12.7%	38.2%
10K-20K	5,730,042	13.5%	23.8%	21.4%	11.3%	29.9%
20K-65K	5,814,780	20.5%	29.8%	19.5%	9.2%	21.0%
> 65K	1,946,052	37.3%	30.8%	14.8%	6.0%	11.2%

Standard Filtering

			CV	CV			% Filtered
Pop Range	# Est	CV<0.1	0.1-0.3	0.3-0.61	CV>0.61	Est=0	Out
< 1K	3,421,166	5.2%	37.3%	34.1%	11.1%	12.3%	91.6%
1K-5K	7,573,261	12.5%	34.0%	31.8%	10.4%	11.3%	74.7%
5K-10K	2,952,148	17.6%	34.1%	28.3%	9.0%	11.1%	64.3%
10K-20K	2,853,569	22.7%	36.3%	23.9%	7.5%	9.7%	50.2%
20K-65K	3,885,880	28.4%	38.3%	19.1%	5.9%	8.4%	33.2%
> 65K	1,716,799	41.6%	33.3%	14.1%	4.6%	6.5%	11.8%

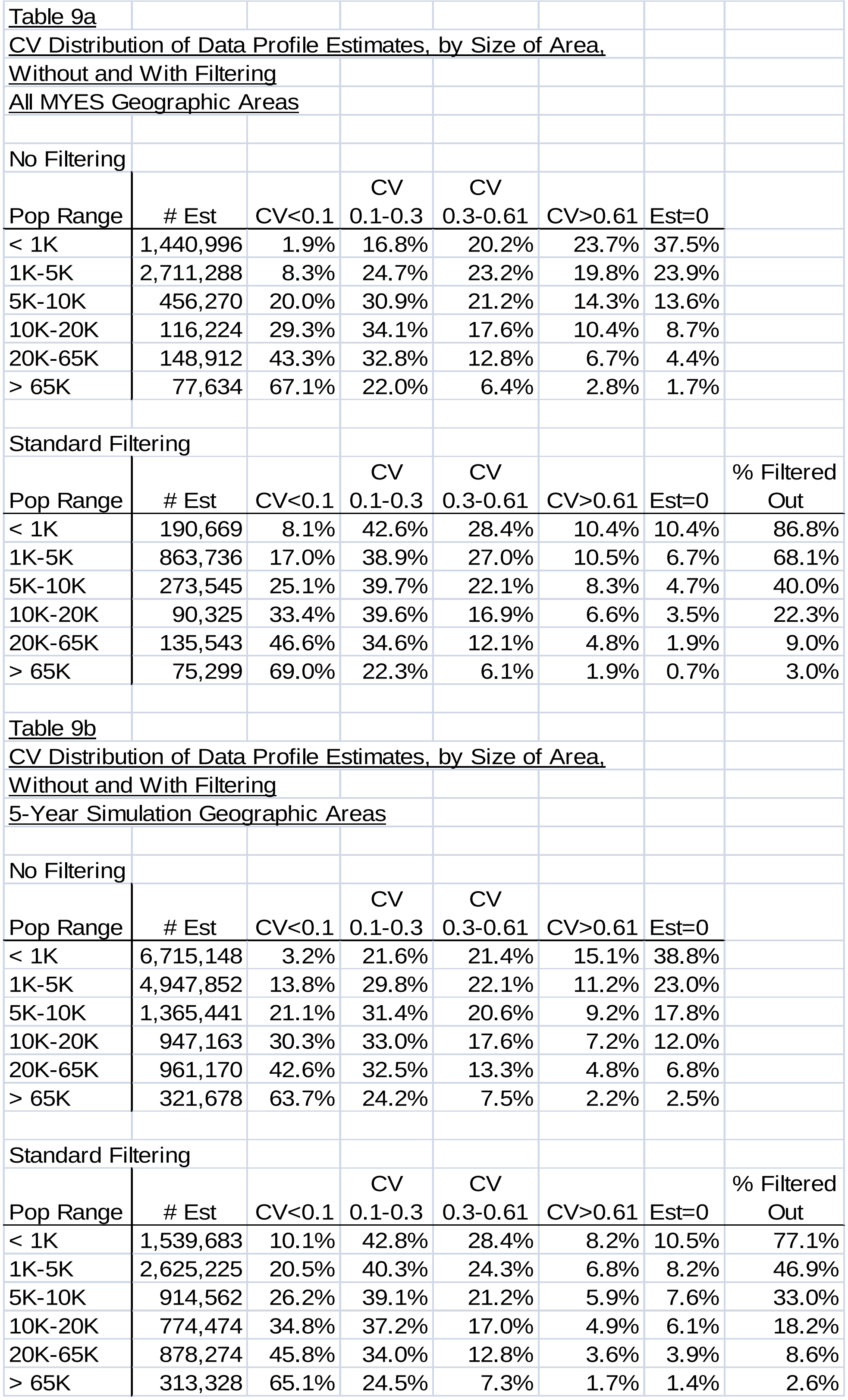

Table 9a

CV Distribution of Data Profile Estimates, by Size of Area,
Without and With Filtering
All MYES Geographic Areas

No Filtering

Pop Range	# Est	CV<0.1	CV 0.1-0.3	CV 0.3-0.61	CV>0.61	Est=0
< 1K	1,440,996	1.9%	16.8%	20.2%	23.7%	37.5%
1K-5K	2,711,288	8.3%	24.7%	23.2%	19.8%	23.9%
5K-10K	456,270	20.0%	30.9%	21.2%	14.3%	13.6%
10K-20K	116,224	29.3%	34.1%	17.6%	10.4%	8.7%
20K-65K	148,912	43.3%	32.8%	12.8%	6.7%	4.4%
> 65K	77,634	67.1%	22.0%	6.4%	2.8%	1.7%

Standard Filtering

Pop Range	# Est	CV<0.1	CV 0.1-0.3	CV 0.3-0.61	CV>0.61	Est=0	% Filtered Out
< 1K	190,669	8.1%	42.6%	28.4%	10.4%	10.4%	86.8%
1K-5K	863,736	17.0%	38.9%	27.0%	10.5%	6.7%	68.1%
5K-10K	273,545	25.1%	39.7%	22.1%	8.3%	4.7%	40.0%
10K-20K	90,325	33.4%	39.6%	16.9%	6.6%	3.5%	22.3%
20K-65K	135,543	46.6%	34.6%	12.1%	4.8%	1.9%	9.0%
> 65K	75,299	69.0%	22.3%	6.1%	1.9%	0.7%	3.0%

Table 9b

CV Distribution of Data Profile Estimates, by Size of Area,
Without and With Filtering
5-Year Simulation Geographic Areas

No Filtering

Pop Range	# Est	CV<0.1	CV 0.1-0.3	CV 0.3-0.61	CV>0.61	Est=0
< 1K	6,715,148	3.2%	21.6%	21.4%	15.1%	38.8%
1K-5K	4,947,852	13.8%	29.8%	22.1%	11.2%	23.0%
5K-10K	1,365,441	21.1%	31.4%	20.6%	9.2%	17.8%
10K-20K	947,163	30.3%	33.0%	17.6%	7.2%	12.0%
20K-65K	961,170	42.6%	32.5%	13.3%	4.8%	6.8%
> 65K	321,678	63.7%	24.2%	7.5%	2.2%	2.5%

Standard Filtering

Pop Range	# Est	CV<0.1	CV 0.1-0.3	CV 0.3-0.61	CV>0.61	Est=0	% Filtered Out
< 1K	1,539,683	10.1%	42.8%	28.4%	8.2%	10.5%	77.1%
1K-5K	2,625,225	20.5%	40.3%	24.3%	6.8%	8.2%	46.9%
5K-10K	914,562	26.2%	39.1%	21.2%	5.9%	7.6%	33.0%
10K-20K	774,474	34.8%	37.2%	17.0%	4.9%	6.1%	18.2%
20K-65K	878,274	45.8%	34.0%	12.8%	3.6%	3.9%	8.6%
> 65K	313,328	65.1%	24.5%	7.3%	1.7%	1.4%	2.6%

Table 10a

Percent of Selected Profile Estimate Not Published (Filtered Out), and CV Distribution for Those Estimates Published

All MYES Geographic Areas

	# Published	% Not Published	CV<0.1	CV 0.1-0.3	CV 0.3-0.61	CV>0.61	Est=0
SEX AND AGE							
21 years and over	4,249	61.0%	93.9%	6.1%	0.0%	0.0%	0.0%
65 years and over	4,249	61.0%	29.5%	65.9%	4.4%	0.1%	0.0%
RACE							
Asian	1,512	86.1%	10.4%	35.2%	41.7%	10.4%	2.3%
HOUSEHOLDS BY TYPE							
Married-couple families	8,738	19.9%	27.1%	55.8%	16.7%	0.3%	0.0%
EDUCATIONAL ATTAINMENT							
Less than 9th grade	2,245	79.4%	11.5%	53.8%	31.7%	2.9%	0.1%
High school graduate or higher	2,245	79.4%	96.1%	3.9%	0.0%	0.0%	0.0%
Bachelor's degree or higher	2,245	79.4%	42.2%	50.4%	7.1%	0.3%	0.0%
PLACE OF BIRTH							
State of residence	9,820	10.0%	25.7%	64.4%	9.7%	0.2%	0.0%
EMPLOYMENT STATUS							
Not in labor force	607	94.4%	99.8%	0.2%	0.0%	0.0%	0.0%
COMMUTING TO WORK							
Car, truck, or van -- carpooled	860	92.1%	24.2%	64.1%	11.6%	0.1%	0.0%
INDUSTRY							
Retail trade	1,849	83.0%	18.4%	75.5%	6.1%	0.0%	0.0%
INCOME AND BENEFITS (IN 2005 INFLATION-ADJUSTED DOLLARS)							
With Social Security	10,556	3.2%	11.5%	54.5%	29.5%	4.2%	0.3%
PERCENTAGE OF FAMILIES AND PEOPLE WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW THE POVERTY LEVEL							
All people	7,099	34.9%	4.7%	39.5%	49.6%	6.2%	0.0%
UNITS IN STRUCTURE							
Mobile home	2,844	73.9%	6.2%	17.3%	8.1%	14.5%	53.9%
HOUSING TENURE							
Renter-occupied	10,453	4.2%	15.5%	48.6%	26.2%	7.6%	2.1%
HOUSE HEATING FUEL							
Utility gas	769	92.9%	57.6%	26.7%	7.8%	4.3%	3.6%

Table 10b
Percent of Selected Profile Estimate Not Published (Filtered Out),
and CV Distribution for Those Estimates Published
5-Year Simulation Geographic Areas

		% Not		CV 0.1-	CV 0.3-		
	# Publishe	Published	CV<0.1	0.3	0.61	CV>0.61	Est=0
SEX AND AGE							
21 years and over	19,575	38.0%	85.9%	14.1%	0.0%	0.0%	0.0%
65 years and over	19,575	38.0%	30.5%	67.5%	2.0%	0.0%	0.0%
RACE							
Asian	4,730	85.0%	17.0%	41.6%	30.8%	6.7%	3.9%
HOUSEHOLDS BY TYPE							
Married-couple families	25,445	19.5%	38.9%	57.5%	3.5%	0.0%	0.0%
EDUCATIONAL ATTAINMENT							
Less than 9th grade	24,534	22.3%	2.9%	29.2%	50.1%	12.3%	5.4%
High school graduate or higher	24,534	22.3%	97.1%	2.8%	0.0%	0.0%	0.0%
Bachelor's degree or higher	24,534	22.3%	23.0%	57.8%	18.0%	1.0%	0.2%
PLACE OF BIRTH							
State of residence	27,194	13.9%	44.8%	52.2%	3.0%	0.0%	0.0%
EMPLOYMENT STATUS							
Not in labor force	5,321	83.2%	95.9%	4.1%	0.0%	0.0%	0.0%
COMMUTING TO WORK							
Car, truck, or van -- carpooled	10,262	67.5%	7.8%	65.4%	26.5%	0.2%	0.0%
INDUSTRY							
Retail trade	7,950	74.8%	18.0%	76.6%	5.3%	0.0%	0.0%
INCOME AND BENEFITS (IN 2005 INFLATION-ADJUSTED DOLLARS)							
With Social Security	30,754	2.6%	18.4%	65.4%	14.8%	1.4%	0.0%
PERCENTAGE OF FAMILIES AND PEOPLE WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW THE POVERTY LEVEL							
All people	25,149	20.4%	5.1%	43.7%	46.9%	4.2%	0.0%
UNITS IN STRUCTURE							
Mobile home	8,748	72.3%	4.3%	36.3%	25.2%	11.4%	22.8%
HOUSING TENURE							
Renter-occupied	30,007	5.0%	13.4%	43.8%	32.5%	8.3%	2.1%
HOUSE HEATING FUEL							
Utility gas	6,282	80.1%	49.0%	28.7%	14.0%	3.3%	4.9%

Table 11a
Impact of "Half Rule" on Data Profiles
All MYES Geographic Areas

Pop range	Total # Geo	# Fail "Half Rule"	% Fail "Half Rule"
< 1K	3,174	3,174	100.0%
1K-5K	5,972	5,203	87.1%
5K-10K	1,005	100	10.0%
10K-20K	256	1	0.4%
20K-65K	328	0	0.0%
> 65K	171	0	0.0%

Table 11b
Impact of "Half Rule" on Data Profiles
5-Year Simulation Geographic Areas

Pop range	Total # Geo	# Fail "Half Rule"	% Fail "Half Rule"
< 1K	13,903	13,703	98.6%
1K-5K	10,244	3,511	34.3%
5K-10K	2,827	68	2.4%
10K-20K	1,961	0	0.0%
20K-65K	1,990	0	0.0%
> 65K	666	0	0.0%