Identifying the Metro-Nonmetro Status of Public Use Microdata Areas

September 22, 2016

For more information contact Tom Hertz at THertz@ers.usda.gov

Summary

Much of the analysis of rural areas relies on the distinction between metropolitan (or metro) and nonmetropolitan (nonmetro) counties. The U.S. Bureau of the Census publishes thousands of tables of socioeconomic statistics, based on household surveys, that permit users to compare metro and nonmetro results. However, many possible cross-tabulations cannot be found in published tables, and must instead be calculated directly from the survey microdata; more elaborate analyses, such as person- or household-level regression modeling, also require the use of person- or household-level microdata.

For researchers interested in rural areas, such analyses are often hampered by a lack of geographic detail in the micro-level datasets that the Census Bureau makes available for public use: detailed geographic identifiers are omitted in order to protect the anonymity of respondents. In the 1980 and 1990 Decennial Census 1% Public Use Microdata files (PUMS-B), the most detailed geographic identifiers are called Public Use Microdata Areas (PUMAs), and these are defined in ways that respect the metro/nonmetro definitions of that era. In 2000, however, the practice of adhering to metro/nonmetro boundaries in defining PUMAs was discontinued. As a result, in the PUMS files for both the 2000 Decennial Census, and the American Community Survey (ACS) which has replaced the Census long form, it is not possible perfectly to identify metropolitan and non-metropolitan areas: some PUMAs are a mix of metro and nonmetro counties or county fragments.[1]

[1] Much more geographic detail is available in the five-year average ACS *summary* files, beginning with 2005-2009, which provide averages of statistics at even the census block group level. But these files do not contain individual person-or household-level records, which are needed for many analyses.

To address this limitation, ERS has created two data files that specify the share of each PUMA’s population that resides in a metro area. One file applies to the Census 2000 PUMA codes, and reflects the 2003 OMB metro classification system. The second is for use with the 2010 PUMA codes, and allows users to choose between the OMB 2003 and OMB 2013 metro definitions. In most situations, designating those PUMAs that are *majority* nonmetro, in terms of population, as “nonmetro PUMAs” yields a fair approximation of nonmetro areas.

This document describes the contents of these files, assesses their validity, and explains how to merge them with ACS PUMS files. Details on sources and methods are found in the comments in the Stata program files “puma2000.do” and “puma2010.do” (plain text format).

File to match to 2010 PUMA Codes (ACS 2012 microdata and later)

Note: The current (vintage-2010) PUMAs appear in ACS PUMS files from 2012 to the present, and should be applicable through 2021.

Metro Def'n: Users may choose to apply either the OMB's June 2003 definition of metropolitan status, or the OMB's June 2013 definition.[2] Use of the 2003 definition will permit comparisons with nonmetro results in ACS files from 2005-2011.

File names: Metro_Nonmetro_PUMA_2010.dta & Metro_Nonmetro_PUMA_2010.csv

File types: Stata (compatible with Stata version 11 and later) & Comma Separated Values

File locations: [link to data product webpage]

Description: Metro population shares for each PUMA are provided, along with a flag (=1) for PUMAs that are majority metro. For the 2013 OMB metro definition, these population shares are from 2010, and are available for the 50 States, the District of Columbia, and Puerto Rico. For the 2003 metro definition, population shares are from 2000, and are not available for Puerto Rico. The mapping of the 2000 population shares to the 2010 PUMAs was accomplished using Brown University's Longitudinal Tract Database, which provides a crosswalk between the 2000 and 2010 Census tracts.

Records: 2,378 PUMAs including PR; 2,351 excluding PR.

Contents:

```
Variable name   Type      Description
----------------------------------------------------------------------------
statefip        byte      State FIPS code (numeric)
puma            str5      2010 PUMA ID (not unique without statefip prefix)
mpop2000        double    Metro03 population in PUMA in 2000
pumapop2000     double    Total population in PUMA in 2000
mpopshare2000   float     Share of PUMA that is Metro03
metropuma03     byte      PUMA Majority Metro by 2003 Definition
mpop2010        double    Metro13 population in PUMA in 2010
pumapop2010     double    Total population in PUMA in 2010
mpopshare2010   float     Share of PUMA that is Metro13
metropuma13     byte      PUMA Majority Metro by 2013 Definition
----------------------------------------------------------------------------
Sorted by: statefip  puma
```

To match this file to the ACS PUMS files from 2012 to the present, merge them using the variables *statefip* and *puma*, which are also found in the ACS.

[2] We use the original June 2003 and June 2013 OMB classifications. Minor changes to county status implemented between major decennial revisions are not reflected here.

Validity of Metro Classification for 2010 PUMAs

There are 2,351 PUMAs in the United States under the 2010 geography. Of these, 206 (9% of the total) contain only nonmetro areas. Antoher 1,874 (80% of PUMAs) contain only metro areas, according to the OMB 2013 definition. Each of the remaining 271 PUMAs, which together contained 12.2% of the population in 2010, is a mix of metro and nonmetro areas (first table, below). Results are similar when the vintage-2010 PUMAs are mapped onto the OMB 2003 metro definition and evaluated according to their populations in 2000 (second table).

2010 PUMAs **2010 Population** **2013 Metro Def'n**	Number of PUMAs	Population in Metro Counties	Total Population	Percent of US Population
Pure Nonmetro	206	0	26,339,051	8.5%
1 to 50% Metro	143	4,861,069	19,275,702	6.2%
51 to 99% Metro	128	12,992,860	18,532,582	6.0%
Pure Metro	1874	244,598,203	244,598,203	79.2%
Total	2,351	262,452,132	308,745,538	100.0%

2010 PUMAs **2000 Population** **2003 Metro Def'n**	Number of PUMAs	Population in Metro Counties	Total Population	Percent of US Population
Pure Nonmetro	244	0	30,476,243	10.8%
1 to 50% Metro	141	4,577,381	18,288,774	6.5%
51 to 99% Metro	116	10,499,616	15,162,994	5.4%
Pure Metro	1850	217,536,057	217,536,057	77.3%
Total	2,351	232,613,054	281,464,068	100.0%

The variable *metropuma13* is set equal to 1 when more than 50% of the 2010 population in that PUMA resides in a metro county under the 2013 definition. The variable *metropuma03* is set equal to 1 when more than 50% of the 2000 population in that PUMA resides in a metro county under the 2003 definition. This correctly classifies 97% of the U.S. population, this figure being the sum of the percentages in the cells on the diagonal of either of the tables below.

Population percentages in actual and estimated metro areas: 2010 PUMAs

2013 Metro def'n, 2010 Populations:

	Actual (County)			
Estimated (PUMA)	Metro	Nonmetro	Total	Error Rate
Metro	257,591,063	5,539,722	263,130,785	
	83.4%	1.8%	85.2%	2.1%
Nonmetro	4,861,069	40,753,684	45,614,753	
	1.6%	13.2%	14.8%	10.7%
Total	262,452,132	46,293,406	308,745,538	
	85.0%	15.0%	100.0%	

2003 Metro def'n, 2000 Populations:

	Actual (County)			
Estimated (PUMA)	Metro	Nonmetro	Total	Error Rate
Metro	228,035,673	4,663,378	232,699,051	
	81.0%	1.7%	82.7%	2.0%
Nonmetro	4,577,381	44,187,636	48,765,017	
	1.6%	15.7%	17.3%	9.4%
Total	232,613,054	48,851,014	281,464,068	
	82.6%	17.4%	100.0%	

About 2% of people in designated-metro PUMAs are actually residents of nonmetro counties, under either metro definition (see column headed “Error rate”). But between 9.4% and 10.7% of people in designated-nonmetro PUMAs are actually residents of metro counties. The higher classification error rates for designated-nonmetro PUMAs are unavoidable, but these figures are still low enough to allow the designated-nonmetro PUMAs to serve as a reasonable proxy for nonmetro America. Note also that the total population living in designated-nonmetro PUMAs is very close to the total population actually living in nonmetro counties (as identified by the 2000 or 2010 Census figures available through the American FactFinder), under either definition. Researchers who wish to focus only on those PUMAs with the highest nonmetro population shares (i.e. the lowest metro shares) can select these using the variables *mpopshare13* and *mpopshare03.*

File to match to 2000 PUMA Codes (ACS 2005-2011 microdata)

Note: The vintage-2000 PUMAs appear in the 2000 Decennial Census PUMS and in ACS PUMS files for 2005-2011.

Metro Def'n: OMB's June 2003 definition.[3]

File names: Metro_Nonmetro_PUMA_2000.dta & Metro_Nonmetro_PUMA_2000.csv

File types: Stata (compatible with Stata version 11 and later) & Comma Separated Values

File locations: [link to data product webpage]

Description: Metro population shares for 2000 for each PUMA are provided, along with a flag for PUMAs that are majority metro, for the 50 States, the District of Columbia, and Puerto Rico.

Records: 2,101 PUMAs including PR; 2,071 excluding PR.

Contents:

```
Variable name   type       Description
-------------------------------------------------------------------------------
statefip        byte       State FIPS code (numeric)
puma            str5       2000 PUMA (not unique without superpuma or statefip)
superpuma       str5       2000 Super PUMA
mpop2000        double     Metro03 population in PUMA in 2000
pumapop2000     double     Total population in PUMA in 2000
mpopshare2000   float      Share of PUMA that is Metro03
metropuma03     byte       PUMA Majority Metro by 2003 Definition
-------------------------------------------------------------------------------
Sorted by: statefip  puma
```

To add these data to the ACS PUMS files for 2005-2011, merge them using the variables *statefip* and *puma*, which are also found in the ACS. In the ACS PUMS files from 2006-2011, people living in Louisiana PUMAs 01801, 01802, and 01905 were all coded as living in Louisiana PUMA 77777. This is because these three PUMAs no longer had sufficient population to be included as separate entities due the effects of hurricane Katrina. These ACS records will not match with any records in *Metro_Nonmetro_PUMA_2000*. However, the composite PUMA 77777 was 100% metro in 2000, and can be manually recoded as such. For example, in Stata:

```
replace metropuma03=1 if puma=="77777"
```

[3] We use the original June 2003 OMB classifications. Minor changes to county status implemented between major decennial revisions are not reflected here.

Validity of Metro Classification for 2000 PUMAs

There are 2,071 PUMAs in the United States under the 2000 geography. Of these, 225 (11% of all PUMAs) are purely nonmetro; another 1,596 (77%) contain only metro areas according to the OMB 2003 definition. Each of the remaining 250 PUMAs, which together contained 12.5% of the population in 2000, is a mix of metro and nonmetro areas (see table, below).

2000 PUMAs **2000 Population** **2003 Metro Def'n**	Number of PUMAs	Population in Metro Counties	Total Population	Percent of US Population
Pure Nonmetro	225	0	29,206,167	10.4%
1 to 50% Metro	149	5,286,762	20,589,478	7.3%
51 to 99% Metro	101	10,241,799	14,574,882	5.2%
Pure Metro	1,596	217,051,379	217,051,379	77.1%
Total	2,071	232,579,940	281,421,906	100.0%

The variable *metropuma03* is set equal to 1 for PUMAs in the last two rows of the table above, for which 50% or more of the 2000 population resides in a metro county under the 2003 definition. This correctly classifies 96.6% of the U.S. population, this figure being the sum of the percentages in the cells on the diagonal of the table below.

Population percentages in actual and estimated metro areas: 2000 PUMAs

(2003 Metro def'n, 2000 Populations)

	Actual (County)			
Estimated (PUMA)	Metro	Nonmetro	Total	Error Rate
Metro	227,293,178	4,333,083	231,626,261	
	80.8%	1.5%	82.3%	1.9%
Nonmetro	5,286,762	44,508,883	49,795,645	
	1.9%	15.8%	17.7%	10.6%
Total	232,579,940	48,841,966	281,421,906	
	82.6%	17.4%	100.0%	

About 2% of people in designated-metro PUMAs are actually residents of nonmetro counties (see column headed “Error rate”). But 10.6% of people in designated-nonmetro PUMAs are actually residents of metro counties. As already noted, the higher classification error rate for designated-nonmetro PUMAs is unavoidable, but this rate is still low enough to allow the designated-nonmetro PUMAs to serve as a reasonable proxy for nonmetro America. Researchers who wish to focus only on those PUMAs with the highest nonmetro population shares (i.e. the lowest metro shares) can select these using the variable *mpopshare03.*

Comparison with IPUMS Metro Variable

The Minnesota Population Centers' Integrated Public Use Microdata Series (IPUMS) adds many helpful variables to the ACS PUMS files, including a metro status variable. For ACS PUMS from 2012 to the present, our classification mirrors theirs for fully-metro and fully-nonmetro PUMAs. However, their metro status variable is coded as "Not available" for all mixed PUMAs, whereas we assign these a metro status based on their metro population shares. The difference between our approach and theirs is illustrated below, using the ACS PUMS for 2014:

ACS 2014 population percentages by IPUMS metro variable versus ERS metropuma13

```
           |  ERS: PUMA Majority Metro
    IPUMS: |  by 2013 Definition
Metro 2013 |         0          1 |     Total
-----------+----------------------+----------
        NA |      6.06       7.40 |     13.46
  Nonmetro |      8.24       0.00 |      8.24
     Metro |      0.00      78.30 |     78.30
-----------+----------------------+----------
     Total |     14.29      85.71 |    100.00
```

The metro status variable in the IPUMS ACS files from 2005-2011 follows the 1993 OMB metro classification, not the 2003 definition, and so will not completely agree with the ERS variable *metropuma03* even for those PUMAs that are not mixed.

For more information on IPUMS geography variables, see:
https://usa.ipums.org/usa-action/variables/group?id=h-geog
and
https://usa.ipums.org/usa-action/variables/METRO#comparability_section

How closely do PUMA-based approximations replicate actual metro/nonmetro differences?

This section evaluates the usefulness of the PUMA-based approximation of metro status, using the 2013 OMB definition, in ACS 2014. We compare metro/nonmetro differences using the ERS PUMA-based approximation to the true differences based on actual metro status in the full ACS survey, as reported by American Fact Finder, for several important economic outcomes; we also report the results for the IPUMS approach. Note that the public use data are a sub-sample of the full ACS, and do not exactly replicate its results at the national level, quite apart from the issue of identifying metro and nonmetro areas.

Metro-Nonmetro results, selected variables in ACS 2014: True vs. PUMA approximations

Variable	Source	US	Metro	Nonmetro	Nonmetro Share if Classified	Share not classified
Total Population	Table S0101	318,857,056	272,650,933	46,206,123	14.5%	0%
	ERS: metropuma13	318,857,056	273,289,365	45,567,691	14.3%	0%
	IPUMS: metro	318,857,056	249,680,736	26,258,424	9.5%	13.5%

Variable	Source	US	Metro	Nonmetro	Difference
All Ages Poverty Rate	Table S1701	0.155	0.151	0.181	0.030
	ERS: metropuma13	0.151	0.147	0.172	0.025
	IPUMS: metro	0.151	0.147	0.174	0.027
Child Poverty Rate	Table S1701	0.217	0.211	0.252	0.041
	ERS: metropuma13	0.212	0.207	0.243	0.036
	IPUMS: metro	0.212	0.207	0.246	0.039
Mean HH Income	Table S1901	75,591	78,931	56,813	-28.0%
	ERS: metropuma13	75,660	78,839	57,473	-27.1%
	IPUMS: metro	75,660	80,121	57,152	-28.7%
Share <HS (18-24 years)	Table S1501	0.139	0.134	0.168	0.034
	ERS: metropuma13	0.138	0.134	0.166	0.032
	IPUMS: metro	0.138	0.133	0.163	0.030
Share BA+ (25+ years)	Table S1501	0.301	0.320	0.188	-0.132
	ERS: metropuma13	0.301	0.320	0.188	-0.131
	IPUMS: metro	0.301	0.329	0.193	-0.135

The above table reveals that using the ERS variable *metropuma13*, which assigns metro status to all PUMAs that are majority metro, leads to a slight but systematic underestimation of the metro-nonmetro gap for each of the variables listed, for reasons explained below. For example, in the full ACS the nonmetro all-ages poverty rate is 3.0 percentage points higher than the metro poverty rate, whereas using *metropuma13* the gap is just 2.5 percentage points. By contrast, the IPUMS approach, which is equivalent to limiting the sample to those PUMAs that are either 100% metro or 100% nonmetro, sometimes overstates and sometimes understates the true metro-nonmetro gap. In three of the cases in the table (poverty, child poverty and mean household income), the IPUMS approach generates an estimate of the metro-nonmetro gap, and of the nonmetro average itself, which is closer to the truth, while in the other two (the education share variables), the variable *metropuma13* comes closer to the truth. Thus the principal advantage of the ERS variable *metropuma13* is not that it necessarily yields more accurate estimates of metro-nonmetro differences, or of average nonmetro results, but rather that it (imperfectly) classifies *all* observations in the dataset, whereas the IPUMS approach is silent on the metro status for 13.5 percent of the population (and those omitted are disproportionately nonmetro residents).

As noted above, the imperfect classification of people by metropolitan status creates a small bias towards zero – a systematic understatement – when estimating the difference between the metro and nonmetro averages of many important economic variables. To see why, suppose some variable's mean is 25% lower in (true) nonmetro counties than in (true) metro counties, but is otherwise random. Given the error rates reported for the vintage-2010 PUMAs, under the 2013 metro definition, using the PUMA-based approximation of metro status would reduce the observed gap to 22%. This is the combined result of a very slight (0.5%) understatement of the metro mean, and a somewhat larger (3.6%) overstatement of the nonmetro mean.

In some cases this bias can be reduced, if not entirely eliminated, by estimating metro-nonmetro differences using a regression approach: We simply regress the outcome variable against *mpopshare2010* rather than comparing means for *metropuma13=0* and *metropuma13=1*. The coefficient for *mpopshare2010* is then an estimate of the metro-nonmetro difference in the outcome, since it is an estimate of the difference we would see between a fully metro PUMA and a fully nonmetro PUMA. If the outcome variable has an expected value that is linear in the share of the PUMA that is metro, this approach should be unbiased. If that expectation is not linear, then we can include a quadratic term, *mpopshare2010* squared, and estimate the metro-nonmetro gap by summing the linear and quadratic terms. This approach can also be used in regressions with other covariates appearing among the x-variables.