Missouri Census Data Center

Ten Things to Know about the American Community Survey (2005 Edition)

The American Community Survey is a very important new source of data of the type usually associated with a decennial census (but based on survey data that are less than two years old instead of over 6 years). However, it has a number of things about it that a user needs to be aware of before attempting to use these data. This list of items is our attempt to make sure users are aware of some of the more important potential "gotcha"s that go with these data. Keep in mind that much of what we say here is specific to the 2005 edition of the data, and may not apply to data released in previous or future years.

We have tried to avoid as much as possible getting into statistical technicalities, but it is not always possible to avoid such topics, since many of the items cited here are the result of statisticians doing things to reshape the data. In such cases even if you cannot follow the details of why something might be the way that it is, at least know that the problem exists.

1. **First, the good news.** The 2005 ACS data provides us with more information about our population and housing stock than we have ever had in our history outside of a decennial census year. The results of the survey for 2005 have been tabulated in a very detailed fashion, with over 1000 detailed ("base") tables, to go along with a series of custom profile, ranking and subject tables, geographic comparison tables and even (coming soon) narrative profiles summarizing the data for a geographic area. These data are rather easily accessed by most users via the Census Bureau's American FactFinder web application, which has undergone a very significant upgrade that not coincidentally was made available with the first release of the 2005 ACS data in August (2006).

   In addition to making the data itself readily available, the Bureau has also provided the usual access to excellent metadata and background information. If anything, most users may find the latter material to be much more than they really want to deal with. There are even a series of on-line tutorials/powerpoint presentations that provide excellent introductions to the data geared to new users. See, for example, the **2005 User's Guide**, which is just one of many key links accessible from the **ACS Home Page**.

2. **"Total Population" is really "Total Population *in Households*."** The 2005 survey **did not include persons in group quarters** (i.e. living in dormitories, nursing homes, prisons, military barracks, etc.) They *will* be included next year for the 2006 data year (and, hopefully, all future years). This limitation makes it rather difficult to cite any trend based on comparing 2000 decennial census data and the 2005 ACS data. The different survey universes used must always be taken into consideration. This limitation is noted in footnotes on the American FactFinder web site, but is not mentioned in table labels. The label says "Total Population" with the *implicit* qualification is "included in this year's sample universe".

3. **Data from the 2005 ACS is only available for geographic areas with a total population of at least 65,000.** This somewhat arbitrary magic number is designed to avoid creating tables where the sample size would result in large standard errors (aka "sampling error"). Of course, many tables that *are* published in the ACS have universes well below the 65,000 threshold. Tables come with confidence interval sizes (MOE for "margin of error" items) to alert users to the reliability of the numbers. (These MOE values are often quite large, especially when dealing with detailed subpopulations, and just because an item gets published does not mean it can or should be used without noting the significant uncertainty involved.)

   We shall not see any numbers for smaller geographic areas until 2008. In that year, the Bureau plans to publish tables for geographic areas of at least 20,000 population. These tables will be based on combining the survey data for 3 consecutive calendar years - 2005 through 2007. In 2010 the Bureau will publish tables for geographic areas of essentially any size (down to census tracts and even block groups) based on data collected over the previous 5 calendar years (2005 through 2009). These tables based on multiple years of survey results are commonly referred to as ***moving average*** tables. (The Census Bureau has decided to discourage use of the "Moving Average" terminology, preferring to have them referred to as "Period Estimates". This reflects the fact that they are not the averages of multiple one-year estimates, but instead the single estimate based on multiple years of data. It's rather a fine point, and we suspect that many people (including us) are still going to call them *moving averages*, at least part of the time.) Note that for larger areas,

you will be able to choose from different sets of tables starting in 2008. For example, in 2010 there will be 3 sets of data for Boone county, MO (over 65,000 population):

- **single-year** tables based on just the 2009 surveys.
- **3-year period estimate** tables based on survey data for 2007 through 2009
- **5-year period estimate** tables based on survey data for 2005 through 2009.

The tradeoff will be between larger sample size vs more current data. Which, if any, of these will come to be considered the "official" tables for the area remains to be determined.

4. **The ACS is about characteristics of the population, not the size.** If you are looking forward to getting new and improved data regarding the *number* of hispanics (or African Americans, or foreign born persons, or poor persons, etc.) living in your state or metropolitan area, or city or county don't get your hopes up. **The ACS does not provide any new data regarding the *counts* of persons or households.** This is because the Census Bureau does not weight ACS survey returns the way they do with decennial census surveys. In the decennial census, the Bureau assigns weights based upon their master address list which is assumed to be definitive and complete. This is not the case with the MAF (Master Address File) used for the ACS. While an initial weight may be assigned based on the number of households found in an area on the MAF, the person record weights are adjusted so that total population counts at the county level by certain age, race, gender and hispanic cohorts will match numbers published in the Bureau's detailed county-level demographic estimates. The result is that the number of cases (persons) in a table is really just a reflection of those estimates, and the data collected in the ACS simply controls the apportioning of those cases (total persons in households, households with hispanic head, total males living in households, etc.) based on characteristics. So the ACS may tell us **what portion** of African Americans are classified as living in poverty in a county, but the actual number of such persons is the result of applying that *portion* to the *number* of African Americans that are estimated (some would say "guesstimated") in the Bureau's estimates program.

   (To make matters worse, the Bureau also adjusts the weights at the household level separate from the weights for persons. But that's a separate issue.)

5. **Income as reported on the ACS is not compatible with income reported in the decennial census.** This is a surprising, rather frustrating and unintended result that the Bureau does not yet fully understand. It has to do with how the questions regarding income are asked on the two survey instruments. The decennial survey asks a person about their income in the previous calendar year, while the ACS survey asks about income in the previous 12 months. Everything gets adjusted for inflation, but when the Bureau looks at test results they have strong evidence indicating that income reported with the ACS version of the question is consistenly lower (by about 4.4% nationwide and 5.2% in Missouri). See the Bureau's rather readable **13-page paper** (pdf document) about this issue by Nelson, Welniak and Posey. The bottom line is that the official Bureau stance is that users should "exercise caution" when trying to do trend anlaysis regarding income or poverty measures using the decennial census vs. the ACS data. It should be noted that when the Bureau did a press conference on the day that the income data were released from the 2005 ACS, all of the economic data trends that they cited were based on the data from the Current Population Survey (which coincidentally were released on the very same day) and **not from the ACS**.

   This has caused considerable grief among journalists and other data analysts who were chomping at the bit to publish articles regarding trends related to these hot topics, only to be told (somewhat belatedly) that they should probably not do it (that is what "exercise caution" really means in this context) since the data were not truly comparable.

   The income comparability problem is just one rather dramatic instance of an item being collected in the ACS that has issues of comparability with the same subject area as measured in the decennial census. For a good summary of such issues we recommend **Census 2000 And ACS 2005 Comparison Issues** prepared by the New York State Data Center (August, 2006).

6. **The 2005 ACS tables contain suppressed data.** Users of decennial census data who have been around long enough to remember the problems we had with the 1970 and 1980 summary data sets because of data suppression will be disappointed to find out data suppression is back for the ACS. It happens at the base table level for the 1 and 3-year data products (but will *not* be done for the 5-year data, to be released for all geographic areas starting in 2010). The Bureau applies what they refer to as their "Data Release Rules" to the base tables in order to protect us from tables "whose reliability is unacceptable". These rules are described in one of the Bureau's methodology papers (see http://www.census.gov/acs/www/Downloads/tp67.pdf if you are not put off by discussions that include references to coefficients of variation and medians.) Some of us are not impressed with

these rules, which seem somewhat arbitrary and which suppress entire tables rather than just the unreliable cells within the tables (and, conversely, allow the publishing of very unreliable cells within tables whose overall reliabilty is deemed acceptable). We'll not go into the gorey details here. But we do want to warn users about some of the unfortunate consequences of this approach by citing an example. Base Table **B17010** deals with the poverty status of families. It breaks the data down by type of family and presence of related children. The table has 41 cells in it. Many of these cells pertain to rather uncommonly-occurring categories such as "Non family, male-headed family households with no related children < 18". Because of this detail, and because the Bureau's algorithm for suppressing tables is designed to protect us from tables with small cell counts, this table winds up being suppressed for 4 of the 16 Missouri counties for which we have ACS data for 2005. The way this is *supposed* to work is that when a table is too detailed like this, then there will be a comparable "C" table with less detail. But there is no table C17010. So, you might think that at least we can go to the Economic Profile table (D03), which has an item telling us what percentage of all families in an area are below the poverty line. But it turns out that the Bureau does not go back to the original data to generate the profiles, but instead just derives/copies them from the base tables. This results in a missing value for the "Percent Poor Families" item on the economic profile for Cape Girardeau county, MO. This, in spite of the fact that there are almost 19,000 family households in that county. And in the very same profile a poverty estimate appears for "Related Children < 5 Years", even though the number of children under 5 in the county is less than 4,000.

7. **The ACS collects data for all 12 months of the year, not for just a single point in time.** The decennial census takes a snapshot of the population and housing stock based on a single day - April 1 of the decennial year. But ACS surveys are distributed year-round, so we have January data and December data. This can be a key factor in interpreting differences in data between the Census and the ACS. Especially so in areas that have seasonal populations, such as resort areas or college towns. In the decennial census you are counted where you are residing on April 1 (with a very few exceptions, such as a person who was on a trip that day and fills out the form when they get back home.) With ACS it is more complicated; where you get counted is based on where you reside when you get the survey (unless you are only staying there "temporarily", defined as less than 2 months). This *should* mean increased populations for places like Lake of the Ozarks (resort area with a large summer-only population) and lower populations for places like Lawrence, KS (college town, where most students are there on April 1 but not in the summer when many will be away for more than 2 months.) However, since the population counts are then "adjusted" so that they sum to the numbers from the estimates program (which are all anchored by the census counts, which are single-point-in-time based), maybe not. It may wind up affecting the characterstics (educational attainment goes down in Lawrence) without affecting the actual head counts.

8. **Data products** for the ACS are (as mentioned in item 1, above) are numerous. They include:

   - "Base tables" (aka "Detailed Tables") that are the usual kinds of summary tables that users of decennial census SF's (Summary Files) are used to. All the data in the other table types can be derived from the information in these tables. These tables have a naming scheme that consists of up to 4 parts. The first lettef of the table ID is either a B or a C; a C table is a collapsed version of a B table with the corresponding table-code. So tables B13008 and C13008 contain similar data (women who gave birth in the last year by marital status and foreign born status). The more detailed B table provides a breakdown of the foreign born category by U.S. citizenship status. There will be instances when data will be suppressed in the B tables but *will* be available for the less detailed C table. Each base table is identified by a 5-digit number with leading zeroes, with the first 2 digits of the code corresponding to a topic. E.g. tables in the 02 series are about race, while tables numbered 15xxx have to do with educational attainment. Some tables have alpha suffixes as part if their IDs. Alpha suffix codes A through I used to indicate that the table has a special race/hispanic subpopulation. Thus table B05003 and B05003A contain similar data but the former is a summary of the total population, while the latter is for the "White Alone" subpopulation. A few tables have a "PR" suffix; these are tables that are specific to Puerto Rico. These will have the same root Table-ID as the corresponding table for U.S. geography. For example, tables B05001 and B05001PR have similar data, with the former available for all U.S. geographies and the latter just for Puerto Rico geographic entities.

   - Profile tables are special extracts based upon the base tables. These tables are much smaller and are used to provide good overviews of an area. There are 4 of them, referenced by codes DP01 through DP04 which correspond to the descriptive titles:

     DP01: General demographic summary {age, sex, race, hispanic, etc.)

DP02: Social profile (education, merital status, fertility, etc.)
DP03: Economic profile (income, poverty status, employment status, etc.)
DP04: Housing profile (Housing values, tenure, units in structure, etc.)

- Ranking and Subject tables - accessible from American FactFinder.

- PUMS files. These for users who want to "roll their own" data tables/analyses. See the next item.

The data products are being (or were) released in *waves* over the late summer and early fall of 2006. As of mid-September Waves 1 and 2 have been released, covering subjects in the first 3 DP categories. The housing data are to be released the first week of October; the Narrative Profile data products along with some more etailed data regarding population subgroups are due in November.

9. **PUMS data included in ACS products.** The Public Use Microdata Sample data allows users to have access to a 1% sample of ACS surveys. This represents about 40% of the available data, since the overall ACS sample is about 1-in-40 or 2.5% of all households within a given year. Researchers who are comfortable with the statistical aspects of analyzing such data, typically with a commercial statistical software package such as SAS or SPSS, can create their own custom tables. The smallest unit of geogrpaphy on these files is the PUMA, or Public Use MicroSample Area -- the same units identified on the 2000 Census PUMS files. (See much more about PUMAs in the next item.) Care must be taken when using PUMS files because of the small sample size.

   The MCDC has a complete collection of the ACS PUMS data, which is kept in its own separate data directory ("filetype") called *acspums*. This directory contains such files for multiple years. These datasets are basically just copies of the datasets as released by the Census Bureau; we did not have to convert them.

10. **The PUMA (Public Use Microdata Area) level geography lets you map and analyze data for your entire state.** One of the things people are used to doing with data from the Census Bureau is creating thematic maps or summary reports that show spatial distributions of data within their state or region. This sort of thing is not generally doable with the ACS data (at least not yet) because of the limited geography available (e.g., you cannot do a state-wide map or report by county because many or most of the counties have no published data as yet). There are two levels of geography where the data at these levels is available for all areas, covering the state; they are Congressional Districts and PUMA's. The former tend to be too large for mapping purposes, while the latter are considerably smaller and hence better suited for a mapping application. Users who are not familiar with PUMA's may find it worth their while to become more familiar; they're not just for stat geeks any more. To learn more, you can start with a set of pdf-file base maps accessible from the Bureau's web site at **http://www.census.gov/geo/www/maps/puma5pct.htm**. From this index page choose your state. When you get to the pdf document be sure to note that the **first page is an *index page*** that displays entities called "Super PUMAs". These are ***not*** the PUMA's you want. The PUMAs you *do* want are sometimes referred to a "5% PUMAs" because they were the geography used on the 5% Sample PUMS files in 2000, whereas the Super-PUMAs (also known as "1% PUMAs") were the ones used on the 1% PUMS files in 2000. The key to using these maps is to understand that the (5%) PUMAs nest within the Super-PUMAs and these pdf files have, following the initial state-level Super-PUMA overview map, 1 or more inset maps showing more detail for metropolitan areas within the state, and then **one page for each super-puma** showing the boundaries of the 5% PUMAs. The maps also show relevant place and county boundaries to help you see what geographic areas correspond to the PUMAs. For example, look at the 3rd page of the pdf file for Colorado (**http://ftp2.census.gov/geo/maps/puma/puma2k/co_puma5.pdf**). You can see from this page that **PUMA 00101** (these codes are unique within state, of course) is comprised of a series of rural counties in the northwestern corner of the state, and going across the northerm border as far east as Larimer county. We see that the PUMA is made up of 5 complete counties (Moffat, Rio Blanco, Garfield, Routt and Jackson) as well as parts of Mesa and Larimer. Sometimes it can get more than a bit tricky trying to decipher the county and PUMA boundaries on these maps to make sure you know exactly what areas a PUMA covers. A more precise and (once you get over the initial shock) easy way of seeing the relationships of PUMAs to other geographic entities, such as counties, is using the MCDC's **MABLE/Geocorr** web application. For example, we invoked the application and specified that we wanted:

    - **Colorado** as the state to process (from thje first select list on the page).

- **PUMA for 5 Pct Samples (2000)** as the source geography (from the "*Select 1 or more "SOURCE" Geocode(s)*" list.)

- **County (2000)** as the target geographic (from the "*Select 1 or more "TARGET" Geocode(s)*" list.)

- To "**Generate 2nd allocation factor (AFACT2**): portion of target geocodes in source geocodes" by checking the middle checkbox in the first line of the Output Options section of the form.

Then we clicked one of the ***Run Request*** buttons on the page. Try replicating these specs yourself. It should take about 10 seconds for geocorr to translate these specifications into a pair of output files, one a csv file that can be used for importing to an Excel or other spreadsheet application, and the other a plain text report file (html and pdf options are available, but this "plain text" version is the default). Here is what you should see on the first few lines of that report:

puma5	County	cntyname	Total Pop, 2000 census	puma5 to county alloc factor	county to puma5 alloc factor
00101	08045	Garfield CO	43791	0.438	1.000
	08057	Jackson CO	1577	0.016	1.000
	08069	Larimer CO	12749	0.127	0.051
	08077	Mesa CO	3112	0.031	0.027
	08081	Moffat CO	13184	0.132	1.000
	08103	Rio Blanco CO	5986	0.060	1.000
	08107	Routt CO	19690	0.197	1.000

Each line represents the intersection of the 00101 PUMA with a Colorado county. The 4th column shows the 2000 census pop count for the intersection (the portion of the county within the PUMA), and is followed by 2 columns of allocation factors. The first allocation factor says what portion of the PUMA's total population is in the County (43.8% of persons living in PUMA 00101 also live in Garfield county), while the second indicates what portion of the county's population also reside in the PUMA (100% of Garfield county resideents live in PUMA 00101, and only about 5% of Larimer countians reside in that PUMA.)

For more information regarding PUMA geography see the **MCDC's page describing PUMAs** in considerable detail, including a link to a **custom report** that shows all the PUMA codes in the U.S. along with their Super PUMAs and what counties and major cities are contained within each.

The author acknowledges the valuable contribution of Leonard Gaines of the New York State Data Center who reviewed an early version of the page and made several valuable suggestions and corrections.

This file last modified Tuesday October 03, 2006, 10:24:08
