Methodology for the Production of American Community Survey Multiyear Estimates

Anthony G. Tersine, Jr. and Mark E. Asiala, US Census Bureau

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, and operational issues are those of the authors and not necessarily those of the US Census Bureau.

Keywords: American Community Survey, Estimation.

Abstract

The American Community Survey (ACS) will be producing three- and five-year period estimates starting in 2008 and 2010, respectively. Before these are released the ACS has created a series of period estimates for research using the data in thirty-four counties from data collected in 1999-2005. These estimates were produced using the methods planned for 2008 and 2010. We will share our research results on the methods that the Census Bureau has developed to produce the multiyear estimates. There are four dimensions of the methodology that we will emphasize: the conceptualization of the estimates, the estimation process, changes in geographic definitions, and use of inflation factors for dollar-valued estimates.

1. Introduction

The ACS collects the sample data that were collected historically from the Census long form. The main difference is that the ACS data is collected continuously throughout the decade instead of the long form's once every ten years data collection. The trade-off of the ACS providing much more timely data each year is the much smaller annual sample size (about 3,000,000 addresses) relative to the long form (about 19,000,000 addresses).

Combining sample from multiple years can reduce the degree of the differences in terms of both timeliness and reliability. Based on consultations with data users, a compromise position was reached by aggregating sample from 5 years of ACS data. The annually updated ACS 5-year estimates will be less timely than ACS 1-year estimates, but for most of the decade they will be more current than once-a-decade long form estimates. The sample size of ACS 5-year estimates (about 15,000,000), while still less than that of the long form, will result in substantially more reliable estimates than ACS 1-year estimates.

The Census Bureau will produce ACS 5-year estimates for the same set of legal, administrative and statistical geographic entities as the long form all the way down to the tract and block group level. So, in this sense, these ACS 5-year estimates will replace the estimates historically produced from the long form.

However, the ACS design provides the flexibility for the production of additional estimates in legal, administrative and statistical geographic entities with larger populations based on aggregating sample from less than 5 years. Specifically, the Census Bureau will produce tabulations for entities with population of 20,000 or more based on 3 years of sample and for entities with population of 65,000 or more based on 1 year of sample. In many ways, these 3-year and 1-year estimates are highly useful and important auxiliary data products from the ACS Program. Plans call for the ACS to produce 1-year, 3-year and 5-year estimates using sets of sample cases as illustrated in Table 1.

Table 1. Sets of Sample Cases Used in Producing ACS Estimates

Data Product	Population Threshold	Year of Data Release							
		2006	2007	2008	2009	2010	2011	2012	2013
1-year Estimates	65,000+	2005	2006	2007	2008	2009	2010	2011	2012
3-year Estimates	20,000+			2005-2007	2006-2008	2007-2009	2008-2010	2009-2011	2010-2012
5-year Estimates	All Areas*					2005-2009	2006-2010	2007-2011	2008-2012

*All legal, administrative and statistical geographic areas down to the tract and block group level.

From this table, one can see the "rolling" nature of the annually updated multiyear estimates. The 5-year estimates released in 2010 are based on sample from 2005 through 2009. In 2011, the updated 5-year estimates are based on sample from 2006-2010. In addition, the table shows that once steady state is reached, the Census Bureau will release all three sets of estimates each year for entities with populations of 65,000 or more.

We will discuss the methodology for producing the multiyear of estimates in Section 2 and research data produced using this methodology in Section 3.

2. Formulation of Multiyear Estimates

While data users have discussed and understood the basic plan for releasing ACS estimates described in the introduction section, the exact formulation of multiyear estimates has been the subject of greater discussion.

These discussions about what the multiyear estimates are supposed to represent includes several dimensions:

- How should the estimates be conceptualized?
- How should the estimation process be defined, particularly as it relates to the use of independent housing unit and population estimates (i.e., controls) and variance reduction methods?
- How should changes in geographic definitions be reflected in the tabulations?
- How should the impact of inflation be reflected in dollar-valued estimates?

While all these dimensions were considered simultaneously in deciding how to formulate multiyear estimates, the sections below present each of the issues individually.

2.1 Conceptualization of Multiyear Estimates

The National Academy of Sciences, data users, and Census Bureau staff have discussed the issue of how multiyear estimates should be conceptualized. The ACS has decided to conceptualize the multiyear estimates as period estimates. We also considered two other options to define the multiyear estimates. Using the first 5-year estimates to illustrate (sample from 2005-2009), we can define the multiyear estimates as:

- Estimates that represent the period of 2005 through 2009,
- An estimate of 2007 (the middle year), and
- An estimate of 2009 (the most recent year).

For the second two options above (i.e., single year representations), some suggested the use of differential weights with relatively more weight assigned to the sample from the year of interest and nearby years.

With the first two options, data users will have three different estimates for large areas (population of at least 65,000) coming out in the same year. The third option will have the same total population (for areas that are controlled to population estimates), but the characteristic estimates will differ.

In deciding among these alternative concepts of multiyear estimates, it is useful to consider the underlying construct of 1-year estimates, which are generally accepted and subject to little debate.

The 1-year estimates are based on the set of interviews conducted from January through December of a given calendar year. The 1-year estimates reflect the characteristics of the interviews conducted in each of the 12 months equally. No month is given preference over any other month nor is more weight given to any particular month (middle nor last). In essence, 1-year estimates represent the 12-month *period* of January through December of a given year.

This conceptualization can be extended to the multiyear estimates. For example, the first ACS 5-year estimate is based on interviews conducted from January 2005 through December 2009. That is, it's based on a 60-month period. So, for a small town with a population of 10,000, the 5-year estimate will reflect the characteristics of the population as collected at the month of interview during the 60-month period.

So, similarly to 1-year estimates, multiyear estimates will be conceptualized as period estimates that are meant to reflect the characteristics of the entity over the entire data collection period: 60 months for 5-year estimates and 36 months for 3-year estimates. This is the first option above. With this conceptualization, the ACS estimates will be labeled through the use of the period of sample years that comprise the estimates. Table 1 illustrates this.

2.2 Estimation Process

Each year the entire set of the ACS data are weighted to produce 1-year estimates. The weighting process includes several factors, such as, the probability of selection, CAPI subsampling, the monthly weighting adjustment, and noninterview adjustment. Finally, the 1-year estimates are controlled to the housing unit and population estimates released for that year. Using these 1-year weights, the Census Bureau releases tabulated data products for all entities with a population of 65,000 or more.

For multiyear estimates, Census Bureau staff has discussed two estimation methods.

- Use each of the previously generated single-year weighted estimates and combine them (e.g., as a simple or weighted average) to produce the multiyear estimates.
- Combine or pool the partially weighted samples from the multiple years and apply final weighting factors to the combined set of sample cases. These estimates are controlled to the simple average of the housing unit and population estimates from the corresponding multiple years.

Again using the first 5-year estimates to illustrate, the first method involves using the separate weighting results from each of the first five years (2005, 2006, …, 2009) and then combining the results to form the 5-year estimates. The second method requires that the partially weighted sample cases from the first five years (2005-2009) be pooled into one data set first and then to conduct the final stages of the weighting process. These estimates are controlled to a simple average of the housing unit and population estimates from the 2005 to 2009 time period released in 2010.

We identified several advantages with the second method of weighting the pooled set of sample cases relative to combining the each year's weighting results, including:

- Improved accuracy of multiyear estimates achieved with pooling by taking advantage of the increase in the number of sample cases in adjustment cells and of less collapsing of adjustment cells during the weighting processes.
- More up-to-date housing unit and population estimates would be available in producing the multiyear estimates. The historic time series going back to the previous census is updated with the release of each year's estimates. This update includes the use of "final" source files and any improvements made to the methods for producing the intercensal housing unit and population estimates.
- Flexibility of developing weighting procedures that are more tailored to the specific geographic needs of the estimates being generated, i.e., 1-year, 3-year, and 5-year. This is of particular interest in the production and release of 5-year estimates at the tract and block group level.
- Production of multiyear data products (both estimates and especially variances) would more closely mirror systems used to produce 1-year data products.

One of the disadvantages of the first method is that it would require the production of the one-year numbers for all geographic areas. This additional production would require both time and resources to complete and be more than that required for the second method.

We will discuss the use of controls and an additional weighting step for multiyear estimates in more detail below.

Since the multiyear estimates represent estimates for the period, the controls used are not a single year's housing or population estimates from the Population Estimates Program but are an average of these estimates over the period. For the housing unit controls, a simple average of the one-year housing unit estimates over the period is calculated for each county. The version or vintage of estimates used is always the last year of the period since these are considered to be the most up-to-date and are created using a consistent methodology. For example, the housing unit control used for a given county in the 2005–2009 weighting would be equal to the simple average of the 2005, 2006, 2007, 2008, and 2009 estimates that were produced using the 2009 methodology (the 2009 vintage). Likewise, the population controls by race, ethnicity, age, and sex are obtained by taking a simple average of the one-year population estimates at the county by race, ethnicity, age, and sex. For example, the 2005–2009 control total used for Hispanic males age 20–24 in a given county would be obtained by averaging the one-year estimates for that demographic group for 2005, 2006, 2007, 2008, and 2009.

Using the pooled weighting also allows us to add a model-assisted (specifically generalized regression

estimation or GREG application) weighting step to the estimation process (Fay 2005, 2006, 2007). The objective of this additional step is to reduce the variances of base demographics. While reducing the variances, the estimates themselves are relatively unchanged. This process involves linking administrative record data with ACS data. As first noted by Paul Voss and his colleagues (Van Auken et al. 2004) and detailed by Starsinic (2005), tract level sampling variances for ACS estimates are considerably larger than initially projected, whereas county-level variances generally meet design predictions. The 3-year GREG application is to help reduce the variances of base demographic estimates at the place and Minor Civil Division (MCD) level, and the 5-year GREG application for census tract level estimates of base demographics.

2.3 Geographic Definitions

Each year the Census Bureau's Geography Division updates the geographic definitions of tabulation entities. For 1-year estimates, we use the definitions that exist as of January of that year (and submitted by April) to tabulate the data. Any changes that occur during the January to December data collection period are not available in time for processing and tabulation. So, only one set of geographic definitions is available.

For multiyear estimates, where the data collection occurs over 3 to 5 years, changes in geographic definitions will occur and be provided prior to producing the multiyear estimates tabulations. However, discussions on how to deal with the different sets of definitions fairly quickly converged on the decision to use the most recent set of definitions. In essence, the same set of definitions used for tabulating 1-year estimates.

Using the 2005-2009 multiyear estimate to illustrate, if a town annexed a set of blocks in 2007 to be part of the incorporated town, then we would tabulate as part of the town the entire set of sample cases from 2005 through 2009 in areas that define the town as of January 2009. This would include 2005 and 2006 sample cases in the eventually annexed blocks that were not part of the town at the time of their interview.

This decision was driven by a handful of factors, including:

- ACS estimates would reflect the most current geographic definitions available.
- Maintains greater consistency with the intercensal housing unit and population time series, which also uses the most current geographic definitions.
- There is no meaningful construct for an "average" geographic definition.

2.4 Inflation Adjustment

The responses to ACS questions that require a dollar-value response are referenced to the month of interview. It may only be for the previous month or the past 12 months, but either way the reference period shifts across the interview months throughout the year. The income questions ask about the past 12 months. So an interview conducted in January 2004 would ask about income from January 2003 to December 2003, but an interview conducted in December 2004 would ask about income from December 2003 to November 2004.

Several assistance programs determine eligibility thresholds using calendar year-based income values. Others require the income data to be used in combination with other data sources that are calendar year-based, such as, tax data. So, to ensure the utility of reported ACS dollar-valued estimates, separate monthly adjustment factors are computed and applied to the corresponding monthly dollar values, to "anchor" the dollar-valued estimates to the calendar year of the interview. This procedure is used for 1-year estimates using the national level Consumer Price Index (CPI) from the Bureau of Labor Statistics since regional CPI are not available for the entire country.

The extension of this logic to multiyear estimates is fairly straightforward. As above, the utility of 5-year dollar-valued estimates is maintained by adjusting the reported values throughout the 5-year period to the last calendar year of the period. This is achieved by computing and applying calendar-year - to - calendar-year inflation adjustment factors as appropriate to the dollar-valued estimates from each year other than the last year in the multiyear period.

For the first 5-year estimate (2005-2009), this would entail computing and applying four different calendar year-based inflation adjustments to dollar-valued estimates from 2005, 2006, 2007, and 2008, respectively, to produce 2005-2009 5-year estimates in 2009 calendar year constant dollars.

3. Multiyear Estimates Study

In preparation for production of the first set of multiyear estimates from the American Community Survey in 2008, the Census Bureau has created a set of research data files for a sample of geographic areas. We produced these data using the methods described above to dress rehearse our production steps, as well as to evaluate the properties of multiyear estimates. These data were released publicly on the ACS website in April 2007. The data released as part of this study are considered research data. The estimates were produced to test production methods and have not undergone the subject matter and technical review required for standard ACS data products.

The study includes a series of 1-year, 3-year, and 5-year estimates for 34 of the 36 ACS Test counties (all except Fort Bend and Harris counties in TX). A total of 14 data sets were created including:

- 1-year estimates for 2000, 2001, 2002, 2003, 2004, and 2005;
- 3-year estimates for 1999-2001, 2000-2002, 2001-2003, 2002-2004, and 2003-2005; and
- 5-year estimates for 1999-2003, 2000-2004, and 2001-2005.

Data products were produced in the form of data profiles (demographic, social, economic, and housing characteristics) for a broad set of geographic areas including counties, places, Minor Civil Divisions, school districts, American Indian Areas, Public Use Microdata Areas, Zip Code Tabulation Areas, tracts, and block groups. Estimates are released as both estimated counts and estimated percentages. The profile format released with this product is not indicative of what will be released for the first multiyear estimate in 2008. The production methods for standard products, including thresholds and data release rules were used to determine the final set of products. The data use the same disclosure limitation methodology (data swapping) as the original 1-year data. The confidentiality edit was previously applied to the raw data files when they were created to produce the 1-year estimates and these same data files with the original confidentiality edit were used to produce the 3-year and 5-year estimates.

In addition, data profiles for tabulation areas that contained only a small number of households are not being released. In order to prevent the disclosure of the data for these areas through subtracting estimates from nested geographic areas, some additional tabulation areas are also not being released. We are researching alternative options to address disclosure risks for these types of areas for the production of our first 5-year data product in 2010.

The Census Bureau is conducting evaluations as well as researchers associated with these tests external to the Census Bureau. The objectives of these evaluations are to answer questions about reliability, quality, stability, and usability of multiyear estimates and to assess methodological issues involved in their production.

Tables 2-4 show data from the ACS multiyear estimates study for Franklin County, OH for the 2005, 2003-2005, and 2001-2005 periods, respectively. Notes: '*****' means that the estimate is controlled and has no sampling error. The margin of error when added to and subtracted from the estimate yields the 90 percent confidence interval for the estimate.

Table 2. 2005 American Community Survey Data for Franklin County, OH

Characteristic	Estimate	Margin of Error	Percent	Margin of Error
Total Population	1,068,080	*****	(N/A)	(N/A)
Male	524,028	+/-496	49.1%	+/-0.1
Female	544,052	+/-496	50.9%	+/-0.1
School Enrollment				
Population 3 years and over enrolled in school	289,153	+/-4,838	(N/A)	(N/A)
Nursery school, preschool	17,200	+/-1,629	5.9%	+/-0.5
Kindergarten	15,169	+/-1,765	5.2%	+/-0.6
Elementary school (grades 1-8)	116,046	+/-2,362	40.1%	+/-0.9
High School (grades 9-12)	57,681	+/-2,101	19.9%	+/-0.7
College or graduate school	83,057	+/-3,868	28.7%	+/-1.0

Source: U.S. Census Bureau, 2005 American Community Survey

Table 3. 2003-2005 American Community Survey Data for Franklin County, OH

Characteristic	Estimate	Margin of Error	Percent	Margin of Error
Total Population	1,065,392	*****	(N/A)	(N/A)
Male	521,467	+/-135	48.9%	+/-0.1
Female	543,925	+/-135	51.1%	+/-0.1
School Enrollment				
Population 3 years and over enrolled in school	291,688	+/-2,907	(N/A)	(N/A)
Nursery school, preschool	19,217	+/-1,338	6.6%	+/-0.4
Kindergarten	15,154	+/-1,037	5.2%	+/-0.4
Elementary school (grades 1-8)	117,825	+/-1,686	40.4%	+/-0.6
High School (grades 9-12)	56,518	+/-1,077	19.4%	+/-0.4
College or graduate school	82,974	+/-2,309	28.4%	+/-0.6

Source: U.S. Census Bureau, 2003, 2004 and 2005 American Community Surveys

Table 4. 2001-2005 American Community Survey Data for Franklin County, OH

Characteristic	Estimate	Margin of Error	Percent	Margin of Error
Total Population	1,062,541	*****	(N/A)	(N/A)
Male	519,337	+/-135	48.9%	+/-0.1
Female	543,204	+/-135	51.1%	+/-0.1
School Enrollment				
Population 3 years and over enrolled in school	291,087	+/-2,161	(N/A)	(N/A)
Nursery school, preschool	19,391	+/-944	6.7%	+/-0.3
Kindergarten	14,517	+/-895	5.0%	+/-0.3
Elementary school (grades 1-8)	119,282	+/-1,187	41.0%	+/-0.5
High School (grades 9-12)	56,102	+/-934	19.3%	+/-0.3
College or graduate school	81,795	+/-1,736	28.1%	+/-0.4

Source: U.S. Census Bureau, 2001, 2002, 2003, 2004 and 2005 American Community Surveys

4. Summary

In terms of the 4 dimensions discussed, ACS multiyear estimates are intended to represent the characteristics of the average population in an entity across the entire multiple-year period. This is best accomplished by pooling the sample into a single estimation process using a single set of averaged controls with a GREG application to reduce variances for base demographics below the county level. The estimates are tabulated as geographically defined in the final year of the series and in dollar-valued terms of the last calendar year. The ACS has produced research data to evaluate these methods and allow the public the opportunity to see what ACS multiyear estimates will look like.

Acknowledgements

The authors wish to thank Scot Dahl and Alfredo Navarro for comments on the paper.

References

Fay, R.E. (2005), "Model-Assisted Estimation for the American Community Survey," *Proceedings of the 2005 Joint Statistical Meetings on CD-ROM*, American Statistical Association, pp. 3016-3023.

Fay, R.E. (2006), "Using Administrative Records with Model-Assisted Estimation for the American Community Survey," *Proceedings of the 2006 Joint Statistical Meetings on CD-ROM*, American Statistical Association, pp. 2995-3001.

Fay, R.E. (2007), "Imbedding Model-Assisted Estimation into ACS Estimation," prepared for presentation at the Joint Statistical Meetings, Salt Lake City, UT, July 29 – August 2, 2007.

Starsinic, M. (2005), "American Community Survey: Improving Reliability for Small Area Estimates," *Proceedings of the 2005 Joint Statistical Meetings on CD-ROM*, American Statistical Association, pp. 3592-3599.

Van Auken, P.M., Hammer, R.B., Voss, P.R., and Veroff, D.L. (2004), "American Community Survey and Census Comparison, Final Analytical Report, Vilas and Oneida Counties, Wisconsin; Flathead and Lake Counties, Montana," unpublished report dated March 5, 2004, available at http://www.census.gov/acs/www/AdvMeth/acs_census/lreports/vossetal.pdf.