AMERICAN COMMUNITY SURVEY DATA FOR ECONOMIC ANALYSIS

Charles H. Alexander
Demographic Statistical Methods Division
Bureau of the Census

For presentation at the October 18-19, 2001 Meeting
Census Advisory Committee of the American Economic Association

AMERICAN COMMUNITY SURVEY DATA FOR ECONOMIC ANALYSIS

Charles H. Alexander
Demographic Statistical Methods Division

The American Community Survey is the Census Bureau's proposed new way to collect census "long form" information. It will provide estimates of a variety of demographic and economic characteristics, for geographic areas of all sizes regularly throughout the decade.

This paper provides an overview of the anticipated data products from this new survey, and discusses some issues about how to use them in econometric modeling.

INTRODUCTION

The decennial census has two parts. The basic census contacts all dwelling units, counting the number of residents and obtaining their age, sex, race and Hispanic origin, and a few other variables. A sample of about one-sixth of all dwelling units receives a "long form", collecting a much longer list of demographic, economic, and housing variables. This long form sample is the important source of information about the general characteristics of the population below the national level.

The American Community Survey (ACS) is the Census Bureau's proposed new way to collect "long form" information, measuring these characteristics continuously throughout the decade starting in 2003. The ACS will regularly update the "snapshot" of communities that the census gives, and will produce a time series that will measure changes over time. This will provide more information than ever before to understand economic and social changes for the nation's communities, both geographic domains and demographic groups. The ACS, like the census long form, covers a variety of topics that are mandated or required by federal law.

Replacing the census long form with the ACS is part of a plan to re-engineer the 2010 census. The other components of the plan are a program to modernize the Census Bureau's Master Address File and TIGER geographic database, and early planning and testing to take advantage of the opportunities provided by these other changes to simplify and improve the census. There will, of course, still be a "short form" census in 2010 to enumerate the population.

PLANS TO INTRODUCE THE ACS

Testing of the ACS methods began with 4 *demonstration sites* in 1996, expanding to 8 by 1998. Most of these started with a 15 percent sample the first year to get a quick look at small-area estimates, and then dropped to a 3 percent sample. For the years 1999-2001, there are a diverse set of 31 *comparison sites*, counties or small groups of counties, chosen to represent a range of situations where ACS data collection methods might conceivably give different results than census long form methods. Most of these sites have a 5 percent annual sample, so that the 1999-2001 averages for small areas such as census tracts can be compared to the census. The larger comparison sites have a 3 percent sample, with one (Houston, TX) having a 1 percent sample.

In addition, the ACS questionnaire was used as part of an operational feasibility test of collecting "long form" information separately from the census. This test, called the Census 2000 Supplementary Survey (C2SS), had an annual sample of about 700,000 addresses in a sample of 1203 counties nationwide throughout the year 2000. The C2SS is designed to produce state and national estimates,

and also can make reasonably precise estimates for counties over 250,000 population, and many places or metropolitan areas over that size. A similar census Supplementary Survey is being conducted in 2001 and is planned for 2002, as part of a transition to the ACS. Data from the ACS comparison sites are included in the C2SS national and state estimates.
The full ACS is planned to have an annual sample size of 3 million addresses, which is an average sampling rate of about 2.5 percent, in all parts of all counties and American Indian Reservations. Each month, there will be a separate panel of 250,000 addresses, with no address repeating in sample for at least 5 years. As with the census long form sample, there will be a higher sampling rate in small governmental units, with a somewhat lower rate in large census tracts.

THE ACS DATA COLLECTION METHODS

Each year the ACS sample addresses will be selected from the Master Address File, and mailed out in 12 monthly panels. Data collection for each monthly panel extends over a three-month period, with telephone follow-up in the second month for addresses where a telephone number can be obtained, and personal-visit followup for a one-third subsample of the remaining nonrespondents. For units with no usable mailing address, for example those with only a physical description, a two-thirds subsample is sent straight to personal visit.

The telephone and personal-visit interviews use computer-assisted interviewing techniques (“CATI” or “CAPI”). The interviewing staff are permanent employees; many also work on the Current Population Survey (CPS) or other surveys. There is a telephone ”failed edit follow-up” to obtain missing information from mail returns which fail a “content edit” because of too much missing data.

The questionnaire asks for the characteristics of the residents of the unit as of the time of the interview. Anyone who is “currently living or staying” at the unit is included as a resident; unlike the census, people who live somewhere else most of the time are included if they are staying at the unit for more than two months.

The C2SS, and the 2001 and 2002 supplementary surveys, do not include group quarters, such as prisons, hospitals, college dormitories, or homeless shelters. The ACS comparison sites, in 1999 and 2001, include group quarters. However, in 2000, they did not include group quarters, to avoid burdening the facilities with both the ACS and the census long form; data for group quarters from the long form will eventually be included in the 2000 ACS estimate in the test sites. In months when a particular group quarters facility is selected for sample, a sample of beds or rooms is designated for interview.

DATA PRODUCTS AND STANDARD ERRORS FROM THE ACS

The ACS will produce a variety of 1-year and multi-year data products, for different purposes. This includes both summary files and tables, and public use microdata samples.

The ACS data product that will most directly replace the long form summary data will be a series of 5-year moving averages for all sizes of geographic areas. These will start in 2008 with the 2003-2007 average, and will be updated each year thereafter. The standard errors of any one of these 5-year estimates will be slightly larger than those of a comparable long form estimate, because the 5-year initial mailout sample is smaller than the decennial long form and because of the sub-sampling for nonresponse follow up; the standard errors will typically be about 1.33 times as large as comparable long form standard errors. We expect this to be offset to some degree by a lower rate of missing data, because of the use of experienced interviewers. A 10-year average from the ACS would have a smaller standard error than the census long form, but we anticipate that most data users would look at the series of 5-year moving averages instead.

Single-year ACS estimates will have about 3 times the long form standard error. However, these estimates will still be useful for larger domains. We have adopted a criterion that single-year estimates will be published for geographic areas or other domains of more than 65,000 population. This corresponds to a "12 percent coefficient of variation for a 10 percent estimate", which implies a 90 percent confidence interval of 10.0 $\pm$ 2.0. This is roughly comparable to the precision of CPS estimates for states. Under the same criterion, we will publish 3-year averages for areas or domains of more than 20,000 population. These published estimates consist of "profiles" giving selected characteristics, and "summary tables", giving extensive univariate distributions and cross tabulations similar to traditional census summary tables.

In addition to these "published" products for general-purpose use, single-year estimates even for areas below the 65,000 or 20,000 limits, will be made available in a form suitable for statistical analysis, such as SAS files. These time series will be useful for fitting statistical or econometric models. They also give data users the flexibility to depart from the standard 5-year averages for specific applications where something other than 5-year averages may be preferable. For example, simulation results suggest that 3-year averages may be preferable to 5-year averages, when past years' ACS data are used to "forecast" the need for funds in the current year as part of funding formulas.

Public-use microdata samples (PUMS) will be released annually from the ACS. As with the census long form, confidentiality concerns limit the geographic detail that can be released, force some "top-coding" and swapping of the data, and prevent the entire sample from being released. We expect the ACS PUMS files to use the decennial census Public-Use Microdata Areas and that these will be of

about 100,000 population. However, the C2SS PUMS files are only at the state level.

THE DECISION NOT TO PRODUCE MONTHLY ESTIMATES

The ACS is being designed to make annual-average and multi-year estimates, and not monthly estimates. Surveys such as the CPS, that are designed to make monthly estimates, complete each month's random sample in the designated month. For the ACS, by contrast, the cases collected in each month are not a strict probability sample; for example the June data consist of early returns from June mailouts, late returns and telephone followup cases from the May mailouts, and personal-visit followup cases from the April mailouts. This is a suitable design for annual data, but not for monthly estimates. The ACS weighting, designed for annual small-area estimates, is done on an annual basis, although month is used as a category in calculating some of the weighting factors.

There is a need for some information about seasonal patterns within the annual average, for areas where there is substantial variation in the resident population across the year, for example areas with many seasonal workers or college students. There is a question on the ACS form that identifies households containing part-year residents. We plan to develop descriptive tables to give some information about seasonal patterns within areas. Such tables will be produced for the comparison sites with highly seasonal populations, and we will get advice from data users about what information is most useful.

There have been suggestions that we include month of interview on the PUMS files. This has not been absolutely ruled out, but because of the disclosure-avoidance rules this would probably come at the expense of details for geography or other variables, since the interview month could be used to help identify an individual household which was known to be in the survey.

Although the ACS cannot replace monthly surveys, data from the ACS can be of use to other federal statistical programs. The ACS data will be useful in survey design and weighting for the Current Population Survey, in the ways that census data have been used in the past, but with greater timeliness. Also ACS data can be useful as a variable in small-area models such as the Local Area Unemployment Statistics (LAUS) program at the Bureau of Labor Statistics, and the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program.

WEIGHTING AND POPULATION CONTROLS

ACS data are weighted to adjust for differential probabilities of selection, and to adjust for other known differences between the interviewed sample and the population. The approaches used are similar to those used by other surveys, including the census long form. The weighting factors with the greatest impact

are the ones that adjust for differences in selection probabilities, in particular for the oversampling of small governmental units for the one-in-three subsampling of nonrespondents, and for the two-in-three subsampling of unmailable addresses.

The other important factors are those that control the survey estimates to agree with intercensal demographic estimates by age, sex, race, and Hispanic origin. These intercensal estimates are produced by updating the previous census results using vital records and other administrative records, as part of a well established federal-state cooperative program. The estimates are available at the county level and our plans are to control the ACS at that level, and possibly for some large places within counties.

In the past, there have been limits on the accuracy of the race/origin detail for counties, but in the next decade there will be improved race/origin information available from the administrative records used in making the intercensal estimates. In addition, information from the ACS on changes in race/origin distribution and household size will be fed into the demographic models to improve their accuracy. Circularity will be avoided, in using ACS data to improve the ACS weighting controls, because the intercensal estimates will still be based on the demographic models, in which the ACS estimates of changes in the population are only one piece of information.

An important decision in controlling the ACS to the intercensal estimates is how to handle differences in residence rules. The ACS includes everyone who is currently living or staying at the sample address, except for people who usually live at some other address and are staying at the sample address for two months or less. The census, and by extension the intercensal estimates, include people at the place they “live or stay most of the time”, as of April 1. The difference will be apparent for some counties or places where there are large numbers of seasonal vacationers or seasonal workers, especially if the April population is much different than the annual average. Also, in college towns the intercensal estimates in theory include all students who stay on campus for most of the year, while the ACS annual averages would not include students during the summer if they are living somewhere else for more than two months.

For the C2SS, the ACS estimates were controlled to agree with the census for states and counties of over 250,000 population. In spite of the conceptual differences in residence rules, there did not seem to be any dramatic differences in the population, before and after controlling the estimates, attributable to the differences in rules for those large areas. This may suggest that there is less difference between the “usual” and “current” populations than might be expected, or it may suggest that respondents do not read the residence instructions very carefully and report similar results regardless of how the instructions are phrased.

However, for some smaller counties known to have highly seasonal populations, there were noticeable differences between the ACS estimate before controls and the census count. In these cases, it does not make sense to force the ACS data collected for the "current residents" to agree with the intercensal estimates based on "usual" residents. Consider the hypothetical extreme example of a summer resort with 1,000 full year residents, all employed, and 6,000 retirees who live there for four months of the year. The ACS annual average would be 3,000 people, two-thirds of whom are retired. It would probably not make sense to control this to equal 1,000 "usual residents, two-thirds of whom are retired".

Based on those considerations, our plan is to start by controlling ACS estimates to agree with the intercensal population estimates (by age, sex, race/origin) for large counties or groups of counties where there is relatively little difference between the "usual" and "current" populations. This will generate ratio adjustment factors that will be applied to each sample person; these factors help to adjust for different coverage. However for smaller, more seasonal, counties and places, the population estimated by the ACS would not be controlled and would therefore show the current residence population. There may be some time series smoothing of the population estimates for these smaller areas.

Additionally, research has started on models to generate intercensal estimates of the current residence population. The idea is to adjust the usual residence controls based on the ACS questions about seasonal residence, and to combine these adjusted controls with the direct ACS estimates of the current residence population, as described in the previous paragraph, but smoothed over time. The combination of the two would use a "shrinkage" or "empirical Bayes" estimator, giving greater weight to whichever component is estimated to have the lower mean squared error in a particular area. If this research is successful, these adjusted population estimates could be used to more exactly reflect the current residence definition in the survey controls.

SOME ISSUES ABOUT THE BASIC ACS APPROACH TO MULTIPLE YEAR DATA

The basic approach of spreading the long form sample over the decade derives from the "rolling sample" design long advocated by the late Leslie Kish of the University of Michigan. Kish recommended using cumulations over different periods of time for different purposes.

Data users familiar with the census long form have wondered whether the familiar decennial "snapshot" could be adequately replaced by a series of somewhat noisier annual estimates, which have to be cumulated into multi-year averages to achieve a precision comparable to the decennial estimate.

Our basic argument in favor of the rolling sample has been as follows. For characteristics that are stable, or changing slowly, in a particular area, using the average over the previous 5 years will be similar to using a larger single-year sample in the third year of the 5-year period. Since long form data takes between 2 and 3 years to be released, the 5-year average is roughly comparable in timing to newly released long form data. When there is a dramatic change in an area, than having an annual time series is especially valuable, as compared to having data only one year in ten

In the latter situation, a satisfactory analysis may require supplementing the 5-year averages with an analysis of single year data; this will be possible with the ACS, since single-year data will be available even for areas below the normal publication thresholds. One potential area of development for the ACS is how to alert users of the 5-year averages to unusual variation in the annual numbers and to display this variation in a way that is helpful to interpreting the 5-year average.

In discussions of this issue, we have yet to encounter an application for which a decennial snapshot of population characteristics is clearly preferable, and there are many where the ACS is clearly preferable. There are certainly potential applications for which an annual sample the size of the census would be helpful: the ACS cannot measure year-to-year change very precisely for small areas. However, realistically the choice is between a decennial long form and annual estimates based on a smaller sample.

SOME QUESTIONS FOR THE COMMITTEE

1. Can you suggest improvements in our plans for data products? In particular, should we be considering alternatives to the standard release of 1-year, 3-year, and 5-year averages for areas of different sizes? What is the best way to release data for use in econometric modeling?

2. Do you have recommendations on the handling of the differences between "usual" and "current" determination of residence, due to collecting data throughout the year for ACS?

3. More generally, what concerns and opportunities do you see for using ACS data in economic analysis and modeling?