INTEGRATING THE AMERICAN COMMUNITY SURVEY AND THE INTERCENSAL DEMOGRAPHIC ESTIMATES PROGRAM

Charles H. Alexander, Jr. and Signe Wetrogan U. S. Census Bureau
Charles H. Alexander, Jr. Demographic Statistical Methods Division, U.S. Census Bureau, Washington, D.C. 20233

I. INTRODUCTION

This paper is a review of issues and possible methods for using information from the American Community Survey (ACS) to improve the Census Bureau's intercensal population estimates program, and for using the estimates from that program as weighting controls for the ACS. We have previously referred to this integration of survey data with model based estimates as part of the "Program of Integrated Estimates".

II. BACKGROUND

The ACS Design: The basic American Community Survey design has a "rolling sample" of approximately 250,000 addresses spread across the entire Master Address File each month. A different sample of addresses is included in each month's mail-out, so that the sample cumulates to about a 2.5 percent sample over the course of a year, and a 12.5 percent sample over five years. The basic survey estimates are annual averages of the number of people or households with specific characteristics. People are included at their current residence at the time their data are collected. As with recent census long form samples, the ACS will have a higher sampling rate in small government units, and a lower rate in large census tracts. No address will be in sample more than once in a five-year period.

The ACS will start nationwide in 2003. In 1999 there are 36 comparison counties using the ACS design with a 5 percent annual sample. Some of these counties were included in previous years of the ACS "Demonstration Period", with 3 counties having been in sample since 1996.

The ACS is a mail survey with follow-up of nonrespondents by telephone and, for a subsample, in person. The data collection for a given monthly sample of addresses takes place over a three-month period. In the first month there are repeated mailings. In the second month there is an attempt at a telephone interview for addresses where no mail form has been returned, and where a telephone number can be obtained. In the third month, a random sample of one-third of the remaining nonrespondents are contacted in person.

The sampling frame for the ACS is the Master Address File (MAF) maintained and regularly updated by the Census Bureau's Geography Division. In some areas, there will be additional corrections and updates to the MAF by the ACS field representatives.

The ACS data are weighted based on the month of collection. For example, the July data come from mail interviews from the July panel, telephone interviews (and some late mail returns) from the June panel, and, personal-visit interviews from the May panel. (Dahl, 1998)

Before applying any weighting controls, initial "pre-control" survey weights are calculated for each sample housing unit or group quarters person. Each sample unit is weighted by the inverse of its selection probability. Units followed up in person (in the third month) have an extra factor of three applied to their weight to compensate for the one-third nonresponse subsampling. Some additional factors are applied to correct for the variation in monthly sample size because of the extended follow-up period, and to correct for units lost because they cannot be interviewed, even by the personal visit follow-up. The "noninterview factors" give higher weights to interviewed units in the same census tract and month of interview as these final nonrespondents. The "pre-control" weights are the starting place for the final stages of ACS weighting, when information from the intercensal population estimates program is incorporated by applying "population controls", as decided in Section III.

Intercensal "Model-Based" Population Estimates: The intercensal demographic estimates of population are derived from complex models combining administrative records with data from the previous census. For higher geographic levels, the approach used for "official" estimates is the *cohort-component method.* The census data are updated by adjusting for recorded births and deaths, aging the survivors by one year as each year passes. This is combined with an estimate of international immigration and emigration.

Internal migration is estimated from tax records and other sources. Population estimates are given each year down to the county level by age, race, sex, and Hispanic origin. Official estimates of total population are given each year for all places with a functioning governmental unit.

For smaller geographic areas, an alternative *housing unit method* is thought to give better results. In this method, an estimate of the number of housing units in each area is made by updating the housing unit count, using information from building permits and other sources. The estimated number of persons is derived from the housing unit estimate, by applying estimates of the vacancy rate and persons per household. The latter two parameters are derived from the previous census values, with some adjustments based on observed changes at higher geographic levels. The housing unit method is not presently used for official estimates, but has proved promising on an "experimental" basis.

The intercensal estimates are generally regarded as more accurate than the initial estimates of the population size that come out of household surveys; indeed most surveys control their estimates to be consistent with the official estimates. One of the reasons the estimates are so accurate is that there is a Federal/State Cooperative Population Estimates (FSCPE) program, which involves review and discussion of the estimates with state and local experts, including comparisons of results with their models.

However, the accuracy and detail of these estimates generally declines as smaller and smaller areas are considered. One of our research questions is how far down the intercensal estimates should be used as population controls.

III. POPULATION CONTROLS IN THE ACS WEIGHTING

As for many demographic surveys, ACS annual estimates are controlled to agree with intercensal population estimates. For the 36 comparison counties, ACS weighted estimates of the number of persons are controlled to the intercensal estimates by age, race, sex, and Hispanic origin at the county level. The control is implemented by "ratio estimation", in other words by applying a "post-stratification" factor to the weight of each person depending on the person's age/race/sex/origin "cell", so that the final weights sum to the population control for that cell. (Dahl, 1998)

Currently, the ACS controls use intercensal estimates that update the 1990 census, but we will soon switch to controls based on the 2000 census.

The estimated number of households is calculated from the final housing unit weight, which is not controlled as part of the person post-stratification. The results from the household weights are therefore not exactly consistent with the results from the person weights. For example, the weighted number of people in households given by the household weights will typically be less than the total person weight for the households, because the person weights include the additional post stratification factors. The remainder of this paper will focus on ACS estimates for the number of persons and will not cover the separate issues of survey weighting for household estimates.

IV. USE OF ACS TO IMPROVE THE INTERCENSAL ESTIMATES

The ACS doesn't replace the census. The census, with a "short form," updated by the intercensal population estimates program, will continue to be the source of the *number* of people and housing units, nationally and for different places. The ACS estimates the *characteristics* of the population, just as the long form sample has done in the past six censuses.

In particular, the ACS can measure changes in race or Hispanic origin distributions, as well as changes in vacancy rates and household size, that can be used to improve the estimates of these parameters used in the demographic models. It can also detect changes in the number of addresses on the Master Address File in a particular area, which can give an early indication of growth in the area.

In our initial test counties, we have seen several instances where the ACS estimates seem to have reflected changes in race distribution better than the demographic estimates, probably because the administrative records that track internal migration for the demographic models do not have much information about race or ethnicity. There are also indications of undercoverage in the ACS that call for using population controls, as discussed in Section III. At first glance this seems less severe than what has been reported for some of our major household surveys but it is hard to be sure based on data from only a few counties.

V. REASONS FOR USING POPULATION CONTROLS

For the 1999 ACS estimates that were recently released, we do not yet have controls for the new "mark all that apply" race questions used on the survey. So the survey basically "speaks for itself" in giving the weighted estimates of the number of people in the detailed race categories, although controls for less detailed categories are used. There will be controls for the new categories after the Census 2000 data have been processed.

Adjustment of the survey estimates to agree with population controls is important for several reasons:

1. *Correction for systematic undercoverage.* Household surveys typically fail to locate every person in the population, so that the weighted number of people using the pre-control person weights tends to fall noticeably short of the census counts or intercensal estimate. The causes have not been firmly established, but are thought to be a combination of missing some living quarters and leaving some people off the household roster. (Hainer, et al, 1998) This *undercoverage* is differential by age, race, and sex. Controlling to the intercensal population estimates can reduce this differential. Of course, controlling the weights cannot correct for any differences in characteristics between the missed people and the people interviewed in the survey except for age, race, Hispanic origin and sex.

2. *Reduction in sampling variance for survey estimates.* The estimated number of people using the pre-control survey weights varies from sample to sample, depending on whether the sample addresses have higher-or-lower-than-average vacancy rates and persons per household. The estimated number by age, race, sex and Hispanic origin also depends on which addresses are chosen by the random sample selection. This *sampling variance* in the number of people contributes to the total survey error in other estimates, such as the number of people in poverty. Controlling to a fixed set of population estimates, regardless of which sample is selected, eliminates the sampling error[1] for the controlled estimates, and tends to reduce the sampling error in other estimates.

3. *Smoothing of age distributions and trends.* Actual age distributions, even for single years of age, tend to be fairly smooth when plotted as a histogram, except for areas with very small population. Similarly, with a few exceptions, the populations of areas tend to grow or decline fairly smoothly over the years. However, even with fairly large samples, survey estimates of number of people, using the pre-controlled weights, very often show large jumps from one age to the next in the age distribution, or large jumps up and down over time, due to sampling error. These random jumps are typically not "statistically significant" when compared to the measured survey standard error, but the patterns displayed when graphing the age

distribution or the trend over time appear unrealistic, and very different in shape, than the actual pattern in the population.

By contrast, the population estimates tend to have a much more realistically shaped age distribution, since they start with the actual distribution from the previous census. For describing trends over time, the population estimates have a less clear theoretical advantage, in that, compared to the survey, they may tend to make the opposite error of showing an unrealistically smooth trend line. Pragmatically, this is less disturbing to data users than an artificially "jumpy" trend line, even if the jumps are declared to be not statistically significant.[2]

Comp

[1] The population controls may have *other* errors (biases), especially errors in the assumptions about migration below the national level, but they do not have *sampling* error, since they are based on a complete census, updated by all available data on births, deaths, migration, and so forth.

[2] This entire issue can be regarded as a special case of the previous topic about "sampling variance", but the tradeoffs for age distributions or trends among different types of errors are very different than for individual point estimates.

Controlling estimates from different surveys to a single set of intercensal population estimates makes it easier to compare the survey results, since they all give a consistent value for the population for any particular time period. It is also an advantage to be consistent with either the official census counts or the intercensal estimates that have been through the FSCPE process.

In practice, the consistency produced by using controls is not required to be perfect. Many household surveys exclude part of the population from their target population; for example, people living in institutions or the military may be excluded. The place of residence for people in institutions may be deliberately defined differently; for example, the Current Population Survey includes college students at their parents' home even if they are away from home during the school year, while the census counts them at the college if they live at the college. There may also be differences in reference period; for example many surveys define the "annual" population as of July rather than April as in the census. Further, there are differences in the "vintage" of the estimates, with the surveys using the best population estimate at the time the survey data are published, while the official population estimates may be subject to subsequent revisions.

VI. DIFFERENCES IN RESIDENCE RULES

In principle, the decennial census counts people at their "usual residence" as of census day. The intercensal estimates start with the census, and so in effect, use the same rule. The American Community Survey uses a "current residence" concept, with the reference date as of the time the form is filled out, or the interview is conducted. So there is a difference between the survey reference population and the available controls.

The Census 2000 instructions are to include "people staying here on April 1, 2000 who have no other permanent place to stay" but not to include "people who live or stay at another place most of the time." . The ACS instructions say to:

- list everyone who in living or staying here for more than 2 months
- list everyone else staying here who does not have another us
- do not list anyone who is living somewhere else for more than 2 months, such as a college student living away

It is not yet known whether respondents read the instructions so carefully that the differences will have much effect, but the intended concepts are quite different in some common situations:

1. In the census, college students living on campus are to be counted in their dormitory. For the ACS, they are to be included on campus during the school year, but if they are away from campus for more than two months, they are to be included wherever they are staying during those months.

2. "Snowbirds" who have both a winter and summer residence are to be included in the census wherever they "spend most of the time". For the ACS, they would typically be included in each residence for part of the year.[3]

---

[3] More precisely, any given sample individual

3. Seasonal workers with no usual residence are to be counted in the census wherever they are living in April.

In the ACS, they are to be included for part of the year at the various locations where they stay.

4. "Loosely attached" household members who move from residence to residence for short periods are to be counted in the census wherever they stay most of the time. In the ACS, they are to be included wherever they are living or staying at the time of the survey.

5. For "commuter workers", who regularly spend week nights near where they work, but have family homes where they return on weekends, the census would count them wherever they stay most of the time, namely at the worksite. The ACS would count them wherever they "are living or staying"; which many of the respondents may consider to be the "family home".

The ACS uses the different rules because it collects data continuously throughout the year. It would be difficult to collect residence as of April 1 for interviews early or late in the year. Application of a "usual" residence rule as of the time of interview would also be difficult for a sample survey, since it would require either 1) moving data from the sample address where the interview took place to the usual residence which is not in sample, or 2) trying to collect information about people who are absent from the sample address for the entire three-month period of data collection for that address. In theory, this problem is avoided for the census, since both addresses are included in the census, so information collected at the "non-usual" address can be moved to the usual address. The ACS' "current residence" rule increases the chances that the people defined as residents will be available to have their data collected. The two-month limit on temporary absences was selected so that the people defined as residents would not be absent for the entire three-month period of data collection for a given address.

An additional consideration in adopting the "current residence" rule is that, in principle, it allows the questionnaire to probe more aggressively for loosely attached individuals who are currently living or staying at the address, but may not be "usual" residents. The theory is that some people do not have a "usual" residence, or are not sure what the term means. The idea of the current residence rule is that if there is any ambiguity about the usual residence, the person should be included "here" if he or she is currently living or staying here. Research will continue on how well the present ACS instructions implement this idea; in particular, we would like to find a way to eliminate the word "usual" from the instructions altogether.
Differences between the ACS pre-control estimates and the population controls, due to the residence concept, have not been noticeable for the ACS demonstration counties, but such differences could be important for smaller areas with a substantial seasonal population. More information will be available in 2000 from the additional ACS comparison sites introduced in 1999, several of which were selected because they contained areas with highly seasonal population.

VII. AN ILLUSTRATIVE EXAMPLE

The following hypothetical example illustrates our problem; it has to be hypothetical because we have not yet encountered a clear problem of this type in our demonstration sites.

Example: A seasonal resort area has:
1,000 year-round usual residents,
all employed;
10,000 half-year "non-usual" residents,
all retired.
The annual average population is 6,000.

If we control the survey estimates to agree with the usual residence population, the survey results would say that there are 1,000 residents, 5/6 of whom are retired. This is clearly not a valid description of this population.

Instead we would propose to give the following

---

with this residence pattern, is included at the residence where he or she is living or staying at the time of interview. It is rare for both of a person's addresses to fall into sample in the same year. However, on average, a group of individuals with this pattern would be included for part of the year in one place and part of the year in the other place.

information combining results from both the ACS and the intercensal estimates:

- The area has an annual average of 6,000 "current residents", 5/6 of whom are retired.

- The area has 1,000 "usual residents".[4]

This gives a valid description of the area, although we know that data users will want to know more about the nature of the seasonal patterns in the area: Are there a large number of seasonal vacationers, seasonal workers, or college students? What time(s) of year are they there?

VIII. OPTIONS FOR DEALING WITH DIFFERENCES IN RESIDENCE RULES IN THE ACS WEIGHTING

One option is to use *usual-residence controls* for all places, even though the data are collected for the current residents. As the example illustrates, the results may not make much sense when there is a substantial difference between the usual-resident and current-resident populations.

A second option is to come up with *current- residence controls* for all areas. For large areas, such as large counties or Indian Reservations, or groups of smaller counties or Reservations, we would estimate the difference between the current-residence and usual-residence populations, and adjust the controls accordingly. For smaller places within the large areas, where there is not enough information to estimate the differences, the uncontrolled estimates could be smoothed across time and across the age distribution; smoothed estimates would be used as controls for the annual estimates of characteristics.

To estimate the difference between the usual- residence and current residence populations, it is necessary to look at both the total number of people in different "seasonal" groups, such as college students, seasonal vacationers, and seasonal workers, as well as an estimate of monthly variation in the total population and in these seasonal groups. Research is needed on how to put this information together to estimate the difference. The ACS may need to add a question about seasonal workers, and additional questions about seasonal residence of individuals more generally.

A third option is a *hybrid approach*, which is currently used by the ACS. For larger areas, where there is little difference between the current-and-usual residence populations, usual residence controls are used. For smaller places within the larger areas, the uncontrolled estimates, which are essentially on a current residence basis, could be smoothed.[5]

Thinking back to the four objectives given in Section V, all three options address the first three objectives to some degree. The higher-level controls apply factors that adjust for systematic undercoverage (Objective 1), although these may not pick up local variations within the larger areas. Either controls, or a combination of controls and smoothing will reduce sampling error (Objective 2) and smooth age distributions and trends (Objective 3).

The usual-residence approach also meets Objective 4, since the population estimates are consistent with the official estimates. However, as the example illustrates, this creates a serious estimation problem since the population in the controls is not the population who are there to have their data collated.

For either the current-resident approach or the hybrid approach, we would be able to get valid characteristic estimates for the current-resident population, but the population base for the characteristics estimates would

[4]Note that for the ACS, the usual residents are also estimated on an annual average basis. In principle, it gives the average over the months of the year of the number of people who in that month would say they are usual residents.

[5]The current ACS estimates do not smooth the lower-level populations, but when more years of data are available, we will experiment with this.

not be the same as the traditional, official population estimates for the area.

The fact that the population base for the characteristics estimates is not *numerically* the same as the official population of the area is not necessarily a problem. Many characteristic estimates use something other than the total population as their reference population. For example, in the census and the ACS, disability status is reported for the civilian noninstitutional population, poverty rates are reported for families not all households, occupation is restricted to employed persons, and many characteristics have age restrictions, such as educational attainment being reported for persons age 25 or more. Whether it is important that the population base for the characteristics estimates would be *conceptually* different than the total population is a different question; we need input from potential analysts of ACS data about the practical effects of having a reference population for characteristics that is based on a different residence rule that the official total population.

IX. NEXT STEPS

The discussion in this paper suggests a number of areas where we need future research or development.

A. Research on improving the usual residence intercensal estimates using ACS information, as discussed in Section IV.

B. Research on converting usual- residence controls to current- residence controls for larger geographic areas, as discussed Section VII.

C. Research on smoothing uncontrolled current-residence estimates for smaller geographic areas.

D. Make a decision between the "hybrid" and "full current residence" methods, as the preferred approximate current-residence approach.

E. Work with potential data users on what information to present about the usual-residence and current-residence populations in ACS publications.

Dahl, S. A. (1998). Weighting the 1996 and 1997 American Community Surveys. Proceedings of the *American Statistical Association Survey Research Methods Section, 172-177.*

Hainer, P., Hines., Martin, E., and Shapiro, 6. (1988). Research on Improving Coverage in Household Surveys.
Proceedings of the Census Bureau Fourth Annual Research Conference, pp. 513-539.