IMPACT OF MULTIYEAR AVERAGING OF DATA FROM THE AMERICAN COMMUNITY SURVEY

Charles H. Alexander, Bureau of the Census, Room 3705-3, Washington, D.C. 20233

Key Words: Rolling Samples, Small-area Data, Moving Average

I. INTRODUCTION

A. Background

The Census Bureau's new American Community Survey (ACS) will use monthly rolling samples to collect basic population and housing data throughout the decade. This will update the kinds of basic socioeconomic and housing information traditionally available every ten years from the census "long form" sample.

Each month's ACS sample addresses will be spread across all parts of the U. S. There will be an initial mailout of 4.8 million addresses per year for 1999-2001 and 3.0 million addresses for all subsequent years. After additional follow up by telephone there will be a personal visit follow up of about one-third of nonrespondents. The ACS design is described in Alexander (1996).

The census long form is presently the main source of data to profile the characteristics of areas smaller than States. The ACS will update these profiles throughout the decade, going all the way down to the smallest areas, such as block groups. Annual estimates from the ACS for large areas will start in 1999. Small-area ACS estimates comparable to the 2000 census will be available in 2002. There will be a partial update in 2003. The first full update, comparable to a 2001 census, will be available in 2004; these updates will be provided annually thereafter. In the year 2000, there will still be a census long form sample to "benchmark" the new ACS profiles; the 2010 census would no longer collect these detailed characteristics, concentrating on the basic count.

B. Multi-year averages from the ACS

For large areas (250,000 population or more) fairly detailed profiles will be produced each year by cumulating the twelve months of ACS data for that year. For medium-sized areas (50,000-250,000) similarly detailed profiles would require cumulating three years of data to have an adequate sample. For smaller areas the most widely used estimates would probably cumulate 5 years of data. This is called "asymmetrical" cumulation of data for areas of different sizes (Kish,1990).

The ACS multi-year cumulations for small and medium-sized areas will have many uses, ranging from basic description of the area, to allocation of government assistance according to measures of need, to providing input for mathematical models used in local planning. The way in which the data are "cumulated" would depend on the intended use. For the basic descriptive statistics for small areas, the cumulation may simply consist of taking an average or weighted average of five consecutive annual estimates for each small area.

Many data users, with decades of experience with the decennial census point-in-time "snapshots", are concerned about how these multi-year averages are to be interpreted, and how such averaging would affect their analysis of the data. This paper is an initial attempt to address these concerns, by looking at State poverty estimates from the Current Population Survey (CPS) March Supplement. A longer version of the paper, available from the author, also proposes additional research, including the use of simulated time series under a variety of models to reflect ways in which data for small areas, such as counties or census tracts, might differ from the State data.

II. DISCUSSION OF ISSUES AND RESULTS

A. Purposes of the ACS 5-year Averages

The ACS 5-year averages have two main purposes:

a. each year, provide an analogue to fresh census estimates for describing and comparing small areas;
b. describe long-term trends.

The use as a "census analogue" is the primary purpose of these estimates. The intent is that, for example, the 2000-2004 small-area estimates can be used as the practical equivalent of data from a 2002 census. These estimates would be available in mid-2005. In general, a multi-year average spanning an odd number of years would be regarded as analogous to an estimate for the middle year that is not available until after the last year of the average.

The intended use of a "census analogue" is to give a recent description of each small area and make comparisons with other areas for the same time period. For this purpose, measurement of changes over time is not an issue; the point to updating the 5-year averages is to keep the "recent description" from getting older. The choice of a 5-year "window" for the small areas was dictated by the objective of keeping the sampling error close to that of the long form, without having too long a lag between the middle year of the interval and the year the data are available.

The time series of "census analogues" can serve a second purpose, description of long-term trends. In the statistical analysis of time series, moving averages are commonly used to "smooth away short-term fluctuations". There is a long-standing literature on the optimum length of the average for different purposes, and the meaning of the trend line. (See Kendall, 1976, Chapter 3). The interest is not just in reducing the sampling error, but in smoothing away short-term ups-and-downs in the actual population values.

The 5-year averages are not a useful tool for estimating the change between the characteristics of an area in two consecutive years or otherwise describing short-term changes in an area. Other tools must be used for this, by analyzing the time series of annual ACS estimates as discussed in Section IV.

B. Uses of the "census analogue estimates"

To evaluate an estimate, we must first pin down what it is supposed to be estimating. If we had a 2002 census or a 2002 census analogue estimate, then in 2005 it could be used in two basic ways, either as:

i) "a current value": use it without updating as though it describes 2005; or
ii) "an historical value": use it as an estimate for 2002.

Many users of census data do not carefully distinguish between the two uses, but sometimes the distinction is clear. If the 2002 values are directly used in 2005 to distribute funds, this is a "current value" use; the difference between the 2002 estimate and the actual 2005 value is the "estimation error". If the 2002 value is compared to other data about 2002, or if the 2002 value is going to be updated using other information about 2002-2005 changes, then this is an "historical" use: in this case the difference between the 2002 estimate and the actual 2002 value is the "estimation error".

In Section II.D, there will be a distinction between "specific year" and "generic year" which applies to decennial or quinquennial estimates.

C. Summary of results: How does the ACS 5-year "census analogue" compare to the alternatives?

Three alternatives can be considered:

i) a decennial census long form;
ii) a quinquennial (every five years) census long form;
iii) an annual census long form.

To isolate the effect of averaging, assume that each hypothetical census, including the hypothetical annual census, would have the same long-form sample size as the total 5-year ACS sample. For purposes of comparison with the ACS the main interest will be the "annual census with three-year lag", in which the results come out three years late. For example, the hypothetical 2002 annual census results would be released in 2005.

The following conclusions are suggested by our analysis of the CPS poverty data and some general considerations about time series. Our research goal concerning multi-year averages is to see whether further analysis confirms these conclusions or refutes them in some circumstances.

1. Conclusions for current year uses:

1.a. For "current year" uses the ACS 5-year census analogue performs similarly but slightly better than an annual census with 3-year lag, as measured by agreement with the current year.

1.b. Using a 2002 census or census analogue for an example, neither one tends to give a very good value to use in 2005 if there are large changes in 2003 and 2004. These changes would be reflected better in the 2003 or 2004 census analogue estimates.

1.c. The ACS 5-year census analogue, with its 3-year lag, is not much worse for current year uses than an annual census which had a 2 year lag, and under some time series models is better.

2. Conclusions for "historical" uses:

2.a. For "historical" uses the ACS 5-year census analogue gives a good approximation to the midpoint "census year" value if the annual values are constant or changing linearly i.e., (at a constant rate) during the five-year period. For other patterns of change, the average may differ substantially from the middle value. The annual census would do much better in the latter case, and is always somewhat better.

2.b. For "historical" uses, the comparison of the ACS with the decennial or quinquennial census depends on whether the use is "specific" or "generic". The census is better if the interest is only in the specific census years; the ACS may be a better generic value.

These conclusions may be paraphrased as follows:

1) **for current-year uses the ACS "census analogue" is about as good as having just gotten new census data: typically slightly better than looking at 2000 census data in 2003, but slightly worse than looking 2000 census data in 2002;**

2) **for looking at historical values the ACS census analogues are not as good as an annual census with five times the ACS sample size: not a lot worse on average, but missing some important short-term fluctuations;**

3) **for historical purposes, a quinquennial census is preferable if the interest is only the specific census years, the ACS census analogues are better as a "generic" mid-decade value.**

D. "Specific" and "generic" historical estimates.

If the goal of the historical estimate is to give an accurate value only for the *specific* year 2000 or 2005, then the quinquennial census is clearly superior to the corresponding 5-year average "census analogue" for describing those two years. The quinquennial census does not give specific estimate for other years.

Alternatively, the mid-decade value may be interpreted as a *generic* value for the years in the middle of the decade. A good test of whether generic or specific-year historical estimates are of interest for a particular purpose is to think about an example like Figure 1. For a specific-year historical use, it would be completely satisfactory to report that in 1985 the South Carolina and Nevada poverty rates were fairly close (15.2 and 14.4 respectively). If instead there is regret that the "unrepresentative" 1985 values are the only ones observed, so that the generally lower 1983-87 rates in Nevada are missed, then the "generic" 1983-87 averages (17.3 and 10.6) may be preferred. (See Figure 2).

E. Measuring long-term trends.

The above discussion has concerned only the quality of the estimates as a description of some particular year. The ACS 5-year averages also describe "long-term" trends. Figures 2 and 3 show the "trend lines" for the two worst-fitting States, Louisiana and the District of Columbia. Even though some important short-term movements are not well reflected, the trend lines clearly give information about changes across the decade. For example, contrast the experience of the District of Columbia with that of Louisiana.

III. EMPIRICAL BASIS FOR THE CONCLUSIONS

Table 1 summarizes the estimation errors for various 5-year "census analogue" estimates and single year "annual census" values, based on the CPS State poverty rates.

The "unadjusted RMSE" is

$$SQRT\left(\sum_{s}\sum_{t=1982}^{1991}(Estimate\ (s,t)\ -\ CPS\ (s,t))^2\right)$$

where CPS(s,t) is the CPS poverty rate for state s in year t, and Estimate (s,t) is the corresponding estimate based on either a 5-year average or the CPS value for a different year. The "lag" describes the difference between the center year of the estimate and year t. For example, using the "5-year Avg. Lag 0", Estimate(s, 1982) would be the 1980-84 average; using the "5-year Avg. Lag 3", it would be the 1977-81 average; using "CPS lag 0", it would be the 1982 CPS value; using "CPS lag 3" it would be the 1979 CPS value. The years 1982 to 1991 are used because those are the years for which all the indicated estimates can be computed. By definition, "CPS lag 0" gives a perfect estimate of the CPS value.

The unadjusted RMSE is not fair to the estimators using one-year CPS because those estimators have substantially higher sampling error than the 5-year averages. To remove this effect, the second column subtracts the estimated sampling error from the quantity (Estimate(s,t) - CPS(s,t))^2 so that what is left is an estimate of the bias due to using values from other years to estimate year t. However, an actual annual census or ACS census analogue would not be without sampling error, so in Column 2 the one-year CPS estimates now look better than what we would see if we compared an annual census to "the truth". The third column adds back to each Estimate(s,t) value the amount of sampling error that would be present in a CPS estimate with five times the actual CPS effective sample size, to reflect what would happen if the CPS annual estimates had the same variance as the 5-year averages they were being compared to. This sampling error is on the order of what the ACS would have for an area of population 20,000.

Either column 2, which gives the estimated bias, or column 3, which puts the bias in the context of the sampling errors for a typical medium-small area using either the ACS or an equivalent-sized annual census, is a reasonable way to make the comparison.

Conclusion 1.a (similarity of 5-year census analogues and corresponding annual census.) In column 3, the RMSEs for comparable "3 year old" 5-year averages and "annual census lagged 3" are respectively 1.73 and 2.04, similar but with a slight advantage to the 5-year average. Column 2 gives the same basic result.

The general similarity of the errors is illustrated in the graphs in Figures 5 and 6, which compare CPS(s,t) to the "5-year average lag 3" and "CPS lag 3". The problem is clearly related to the lag 3 not to whether the average is used. West Virginia is the worst-fitting State for the lagged 5-year average.

Conclusion 1.b (lack of fit if there are large changes between "census year" and "current year".) This is implied by the basic arithmetic of averages. For illustration, consider West Virginia for t = 1989.

Conclusion 1.c (5-year average lagged 3 almost as good as one-year value lagged 2, and sometimes better.) Here the RMSE comparison in column 3 is 1.73 for the 5-year average and 1.576 for the one-year value lagged 2, not a large difference. As before, column 2 gives a similar conclusions (1.12 vs .92). A simplistic time series model under which the lagged 3 average would do better is

$$Y_t = u + a_t + e_t$$

where $\{a_t\}$ and $\{e_t\}$ are series of uncorrelated errors with mean zero, due to "true noise" and "sampling error". Of course, there are many other models under which the one-year value lagged 2 would be better.

Conclusion 2.a. This is also based on the arithmetic of averages. The graphs in Attachment A illustrate the point. Observe that the 5-year average fits fairly well for D.C. during t = 1988 and 1989 in the middle of a steadily increasing trend, but does very poorly in 1986 in the middle of a "V"-shaped pattern.

Averaged over all States and years, the 5-year ACS averages would have a RMSE of only about 1 percentage point (.96 from Table 1). However, in some years the deviations can be quite large. High or low periods of less than five years tend to be understated (see 1983-85 and 1986-88 in the D.C. graph). Single-year "spikes" can be missed almost totally (see 1982 in West Virginia).

For a specific year, an "annual census" value for that year would be more accurate than the corresponding ACS census analogue. The 5-year average census analogue has a bias which is not present in the annual census (.62 vs zero Column 2 in Table 1). The average difference in RMSE may not be large (.96 vs .72 in Column 3 of Table 1) but this average can include large deviations, such as the previously cited 1985 D.C. values, depending on the area and what happened in the years around the particular census year.

Conclusion 2.b As discussed above, there are two ways to look at the RMSE when comparing the ACS to a quinquennial (or decennial) census in noncensus years:

i) The quinquennial census give a better specific historical estimate for the census years and gives no historical value for other years;
ii) The quinquennial census estimate gives a generic estimate for the years around the census. In this case, the errors for the years before and after the census could be defined to be the differences between the census estimate and the actual value for these years. Averaging these errors for the census year and the two years before and after the census gives (see Column 3 of Table 1) $\mathrm{SQRT}((1.57^2 + 1.18^2 + .72^2 + 1.18^2 + 1.57^2)/5) = 1.28$, compared to the RMSE of .96 using the census analogue for each of the years. The corresponding comparison from Column 2 is 1.05 vs .62.

IV. QUESTIONS ANSWERED BY SOMETHING OTHER THAN THE 5-YEAR AVERAGES

Many questions that could be addressed with ACS data would call for something other than the 5-year average "census analogue estimates". These are questions that cannot be addressed adequately (or at all) with decennial census data, so we should not expect to address them with "census analogues".

A. Analyzing the relationships of variables that change over time.

Suppose we are looking at the relationship of income, educational attainment and age, using the pooled 2000-2004 data. For this purpose, the relationship of variables should be analyzed or modelled by looking at individual microdata or annual estimates, including time as a variable in the model.

B. Detecting or explaining short-term changes.

As discussed previously, multiyear averages are not a good way to describe short-term changes. For example, comparing the 2004-2008 average to the 2005-2009 average really looks at these difference between 2004 and 2009, not the difference between 2006 and 2007 which are the mid-point "reference years" of the two averages.

Some of the most important potential new uses of the ACS involve tracking short-term change, namely measuring how various statistical indicators change after State or local governments change programs or policies.

(Center for the Study of Social Policy 1995, pp 10-12). These "policy" questions need to be addressed by analyses of the time series of single-year ACS estimates. This is beyond the scope of the current paper, which deals with multi-year averages.

C. Providing information to time series models.

Data from the ACS will prove useful for time series models, whether directly as variables in the models or indirectly for "calibrating" the models by comparing their forecasts to ACS data. An important example of the latter is the current use of census long form data to calibrate metropolitan area traffic planning models by every ten years comparing the model's predicted journey-to-work patterns with the corresponding small-area long form estimates (U.S. Department of Transportation, 1996). ACS data might also be valuable inputs to small-area estimates modelling efforts such as those that make small-area unemployment estimates or poverty estimates using combinations of survey data and administrative records.

For these purposes, the time series analyst will usually use the individual annual ACS values even for small areas where the sampling variability is high. The "cumulation" takes place by using many years of data to estimate the parameters of the model or by using the average fit over many years to "calibrate" the model. Data files of annual values will be made available to researchers for these purposes, but the annual values will not be published as "stand alone" estimates for specific small areas.

D. Doing better than a "census analogue" for "current year" uses.

Conclusion 1.b suggests that the 3-year time lag of between the mid-year of the 5-year average "census analogue" and the current year may cause problems in some applications. The estimates may be "fresh" by comparison with a decennial census, but they do not immediately pick up changes in a small area. This could be important for areas that have recently undergone a fundamental change.

A frequent suggestion has been that the unweighted 5-year average should be replaced by a weighted average on giving more weight to the more recent years. The author has resisted this suggestion on the grounds that the optimal weights depend on the specific characteristic being estimated, the set of geographic areas being analyzed, and the intended use of the data. Further, the optimal analysis usually will not be to look at a (weighted) average at all, but to use the annual time series in some other way to make forecasts or inferences.

Accordingly the following approach is proposed by the author. Start with the unweighted average "census analogue" as the basic general-purpose official estimate provided by the ACS. If for a particular purpose, there is sufficient concern about the effect of changes between the midpoint of the 5-year average and the time the average is released or used, then a special analysis of the ACS time series should be performed relevant to the purpose at hand. This analysis would use the annual ACS estimates, so as to extract as much relevant information as possible. These analyses would not replace the census analogues as the "official numbers", except if additional steps (legislative or otherwise) were taken to declare them the official methods for a specific purpose based on some widely accepted research results.

References

Alexander, C.H. (1996). “Some Basic Technical Information about the American Community Survey.” Internal Census Bureau draft report, dated June 12, 1996.

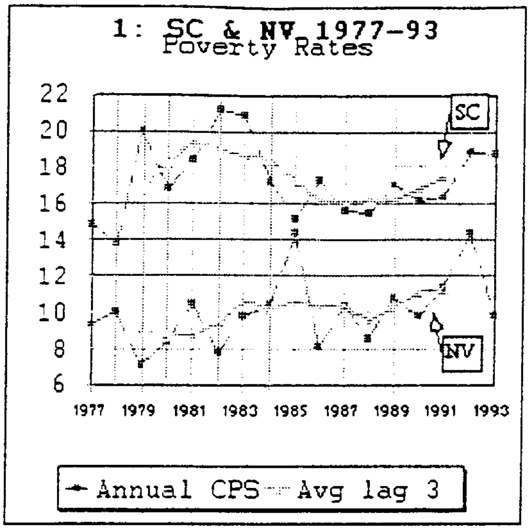
Center for the Study of Social Policy (1995). *Making Decisions Count: How the Census Bureau's Now Survey Could Transform Government.* Washington, D.C.: October 31, 1995.

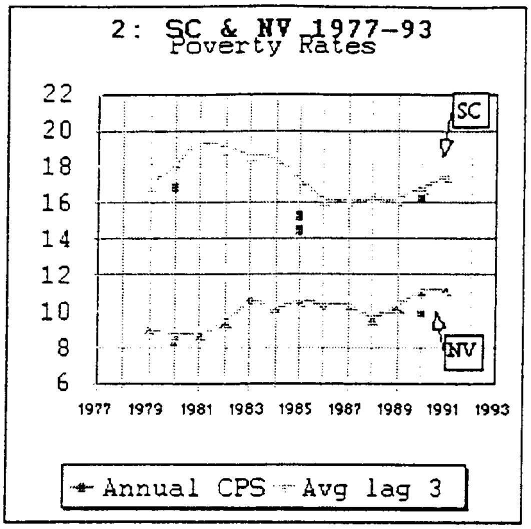
Kendall, M. (1976). *Time Series*. Hafner Press, NY
Kish, L. (1990). “Rolling Samples and Censuses.” *Survey Methodology*, 16, 1, pp. 63-79.

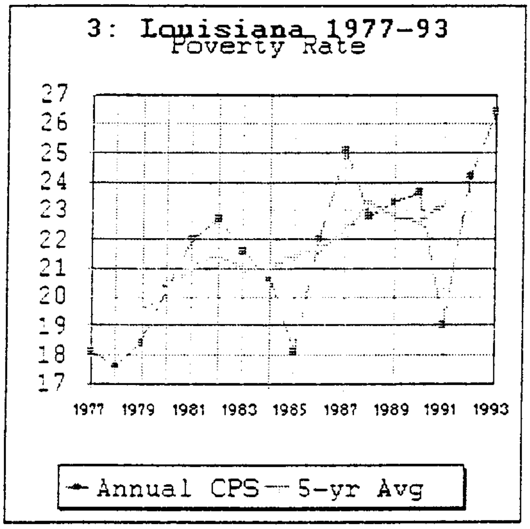
U.S. Department of Transportation, Bureau of Transportation Statistics (1996). Implications of Continuous Measurement for the Uses of Census Data in Transportation Planning. Washington, D.C.: April 1996.

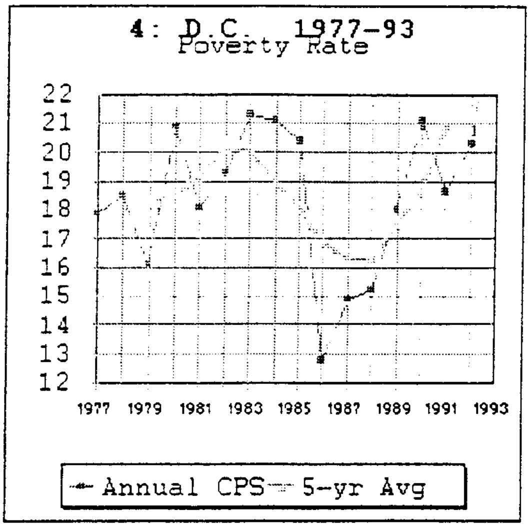
Table 1
Root Mean Squared Error (RMSE) in Poverty Rate
Averaged over all States and Years 1982-91

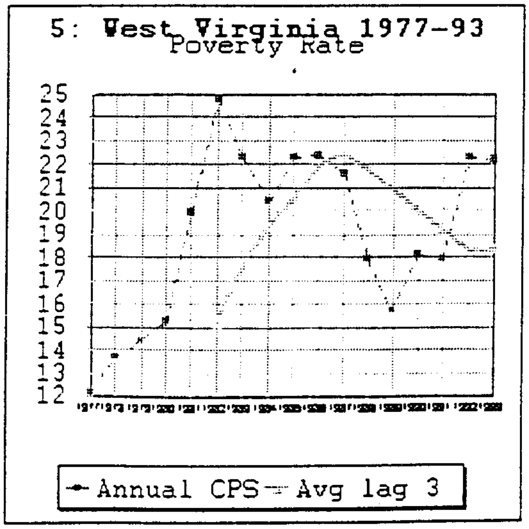
	Unadjusted RMSE	Adjusted to Eliminate Sampling Error	Adjusted to Comparable Sampling Error
5-year Avg. Lag 0	1.31	.62	.96
5-year Avg. Lag 3	2.23	1.57	1.73
CPS Lag 0	0	0	.72
CPS Lag 1	1.98	.92	1.18
CPS Lag 2	2.44	1.39	1.57
CPS Lag 3	2.80	1.90	2.04

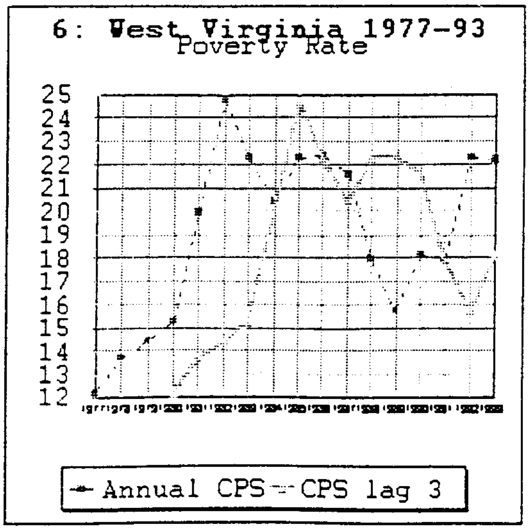

1: SC & NV 1977-93
Poverty Rates
22
20
18
16
14
12
10
8
6
1977
1979
1981
1983
1985
1987
1989
1991
1993
SC
NV
Annual CPS
Avg lag 3


2: SC & NV 1977-93
Poverty Rates
22
20
18
16
14
12
10
8
6
1977
1979
1981
1983
1985
1987
1989
1991
1993
SC
NV
Annual CPS
Avg lag 3


3: Louisiana 1977-93
Poverty Rate
27
26
25
24
23
22
21
20
19
18
17
1977
1979
1981
1983
1985
1987
1989
1991
1993
Annual CPS
5-yr Avg


4: D.C. 1977-93
Poverty Rate
22
21
20
19
18
17
16
15
14
13
12
1977
1979
1981
1983
1985
1987
1989
1991
1993
Annual CPS
5-yr Avg


5: West Virginia 1977-93
Poverty Rate
25
24
23
22
21
20
19
18
17
16
15
14
13
12
Annual CPS
Avg lag 3


6: West Virginia 1977-93
Poverty Rate
25
24
23
22
21
20
19
18
17
16
15
14
13
12
Annual CPS
CPS lag 3