AN EMPIRICAL COMPARISON OF THE SIMPLE INFLATION, SYNTHETIC AND COMPOSITE ESTIMATORS FOR SMALL AREA STATISTICS

Wesley L. Schaible, Dwight B. Brock, and George A. Schnack
National Center for Health Statistics

I. Introduction

Large samples such as those of the Current Population Survey (CPS) and Health Interview Survey (HIS) have been designed to provide national and regional estimates. As useful as such statistics are, there is considerable demand for additional estimates for smaller geographic areas, for example, States and counties. One way to meet this demand is to redesign the survey, but this can be both expensive and time consuming. Depending upon resources and objectives, other approaches, although they may produce biased estimates, deserve consideration. Several biased estimators were considered in 1968 in the publication, Synthetic State Estimates of Disability. The authors stated that the sample size (and design) of HIS was inadequate to make State estimates by conventional procedures and suggested that a synthetic estimator be used.

This estimator has since received considerable attention. Levy (1971) used mortality data to compute average relative errors of synthetic estimates for States. Gonzalez and Waksberg (1973) calculated mean square errors averaged over all small areas and compared synthetic and direct estimates for selected Standard Metropolitan Statistical Areas. Gonzalez and Hoza (1975) investigated errors of synthetic estimates using unemployment data for counties from the CPS and the 1970 Census. Namekata, Levy and O'Rourke (1975) investigated synthetic State estimates of work loss disability in a similar manner. Schaible, Brock and Schnack (1977) compared the average squared errors of synthetic and direct estimates of unemployment rates for county groups in Texas.

It is the purpose of this paper to compare the synthetic estimator, a simple direct estimator and a composite estimator, which is a weighted function of the other two. To provide information regarding the performance of the three estimators each was used with 1970 HIS data to produce estimates of unemployment rates for 25 HIS primary sampling units (PSU's) in Texas. Comparable parameter values were obtained from the U.S. Bureau of the Census (1972), General Social and Economic Characteristics. A similar procedure was followed in estimating the percent of the population completing college for each of the fifty States and the District of Columbia. Three years of HIS data were combined (1969-1971), and comparable population values were obtained from the 1970 Census Public Use Sample Tapes. The State values obtained from the one-in-one hundred sample on these tapes were treated as population values for comparison with estimates.

Traditionally the estimator used to produce estimates reflects the design of the sample from which the data were collected. Even though this is not entirely true of the estimators considered in this paper a few remarks about the HIS design will be useful as background for the comparisons presented. For more complete details on this design, see NCHS (1958). The HIS uses a multistage probability design which permits continuous sampling of households from the civilian, noninstitutionalized population of the United States. The first stage of the 1970 design consisted of a sample of 357 primary sampling units (PSU's) chosen from approximately 1,900 geographically defined units covering the 50 States and the District of Columbia. A PSU is defined as a county, a group of contiguous counties or a Standard Metropolitan Statistical Area. Within each PSU, Census enumeration districts are ordered geographically and divided into small clusters of households. A systematic sample of clusters is then selected. The 1970 HIS sample was composed of some 37,000 households, or a total of about 116,000 individuals.

II. Estimators

The synthetic estimator generally has a negligible variance but often a nonnegligible bias that can only be estimated under special conditions (Schaible, 1975). This non-quantifiable bias is a serious problem and is one reason the synthetic estimator has been used only in special situations. The justification for using this estimator is based on the assumption that the characteristic being estimated is correlated with certain demographic characteristics of the population. The first step in constructing a synthetic estimate is to create a cross-classification of demographic cells in such a way that the local area population in each cell is known. The synthetic estimate for a local area is then formed by weighting a larger area estimate of the health characteristic for each demographic cell by the proportion of the local area population in that cell and then summing over all cells.

For a more precise definition of the synthetic estimator let $y_{d\alpha i}$ denote the observation of interest on the ith individual in the αth demographic cell in the dth local area. Here $i=1,2,\ldots n_{d\alpha}$, the number of sample units in the dth local area and αth cell, $d=1,\ldots,D$, the total number of local areas, and $\alpha=1,2,\ldots k$, the number of α-cells. Also, let $N_{d\alpha}$ represent the number of people in the population in area d and cell α. The sample mean of the αth demographic cell for the larger area is then

$$\bar{y}_{.\alpha} = \sum_{d=1}^{D} \sum_{i=1}^{n_{d\alpha}} y_{d\alpha i} / n_{.\alpha} ,$$

and the simple synthetic estimator for local area d is

$$\bar{y}'_{d.} = \sum_{\alpha=1}^{k} N_{d\alpha} \bar{y}_{.\alpha} / N_{d.} . \qquad (1)$$

Two synthetic estimators are used here, one when the small areas investigated are States and a slightly different one when the small areas are county groups. The estimator used to produce estimates for States is described above, except for the addition of appropriate sampling weights and a ratio-adjustment. This ratio-adjustment forces the weighted sum of the individual State synthetic estimates in a geographic region to be consistent with the usual HIS probability estimate for that region. There is evidence to suggest that when estimating for States, the synthetic estimator with this adjustment has smaller average squared error than without the adjustment (Schaible et al, 1976). The α-cells for State synthetic estimates in this paper were defined to be the 64 cells created by cross-classifying the following variables:

1. Color: white; other

2. Sex: male; female

3. Age: under 17 years; 17-44 years; 45-64 years; 65 years and over

4. Family size: fewer than 5 members; 5 members or more

5. Industry of head of family: Standard Industrial Classifications: (1) forestry and fisheries, agriculture, construction, mining, and manufacturing; (2) all other industries.

For these cells 1970 State populations are available from the Census Public Use Tapes, and reliable national estimates are available from HIS. However, for county estimates, where the larger area was defined as the Southern Region, the HIS sample sizes in some cells are small. In this case, 8 cells were defined by the age and sex groups above. County populations were available for these cells in the Bureau of the Census (1971), General Population Characteristics. The synthetic estimator used for county group estimates did not contain a ratio adjustment.

If data from a sample designed to make estimates for a large area are to be used to make estimates for a small area and there are no sample units in the small area of interest, then obviously conventional estimation methods cannot be used and a synthetic approach must be considered. However, when estimating for a small area that contains sample units the possibility of using conventional estimators should not be ignored. It is evident that at some point, as the number of sample units in an area increases, a conventional estimator becomes more desirable than a synthetic one. This is true whether or not the sample was designed to produce estimates for small areas. Thus a second, more direct approach, is to use conventional estimators with the sample units that fall in the small area of interest. This approach, while not new, has received little attention in the literature, but it would seem to have potential for areas where sample sizes are reasonably large. For example, in California 96 percent of the population reside in the primary sampling units surveyed by HIS and the total HIS sample size exceeds 10,000 persons each year. In situations such as this, one suspects that a conventional direct estimator might be more appropriate than a synthetic one.

The simplest of the conventional estimators is the unweighted sample mean or simple inflation estimator, which for local area d may be written as

$$\bar{y}_{d.} = \sum_{\alpha=1}^{k} \sum_{i=1}^{n_{d\alpha}} y_{d\alpha i} / n_{d.} \quad . \qquad (2)$$

The simple inflation estimator is by far the most widely used of the three considered here. Its simplicity is appealing and with appropriate sample design it is unbiased and its variance can be estimated. However, when used to estimate for small areas from samples designed for large areas (as are the HIS and CPS) the conventional sampling theory model yields little information about the properties of this estimator. For this reason alternative estimators have been proposed.

The idea of a composite estimator is not new; it was discussed in the 1968 publication cited above. It was also mentioned there that a desirable feature of such an estimator would be that the synthetic component receive more weight when the State sample size was small and the direct component receive more weight when the sample size was large. Royall (1973) in a discussion of papers by Gonzalez (1973) and Ericksen (1973), also suggested that a choice between direct and synthetic approaches need not be made but that "... a combination of the two is better than either taken alone." Also, as related by Gonzalez and Hoza (1975), "In a seminar given at the Bureau of the Census in March 1975, William G. Madow suggested a combination of synthetic estimates and observed values for the primary sampling units included in the CPS." Investigations into the basis for and properties of the composite and other related estimators are presently taking place (Royall, 1977, Schaible, 1977).

One rather obvious approach to arrive at a specific composite estimator is to weight each component by the inverse of its squared error and then normalize so the sum of the resulting weights is unity. Empirical comparisons of the errors of various direct and synthetic estimators for States and county groups led to a specific formulation of such a composite estimator. Given a design assume the expected squared error of the simple inflation estimator is of the form $b/n_{d.}$, and that of the synthetic estimator is b', where b and b' are constants. Then if each component estimator is weighted by the inverse of its expected error, the following composite estimator results:

$$\bar{y}''_{d.} = (c_d)\,\bar{y}_{d.} + (1-c_d)\,\bar{y}'_{d.} \quad , \qquad (3)$$

where $c_d = n_{d.}/(n_{d.} + b/b')$. The quantity b/b' is the small area sample size at which the expected errors of the two component estimators are equal.

III. Results

Figures 1, 2 and 3 show the plots of esti-

mated unemployment rates versus the actual rates obtained from the 1970 Census for the three estimators considered. The vertical distance from a point to the 45 degree line shows the magnitude of the error of the estimate represented by that point. The average squared error (Table 1) is simply the average over the 25 county groups of the squares of the differences between the estimates and the corresponding Census values. The correlations (Table 2) given are Pearson's product moment correlation coefficients.

The average squared error in estimated unemployment rates produced by the simple inflation estimator is large, 6.85 percentage points. However, it should be noted that the 1970 HIS sample sizes of the civilian labor force in these county groups are generally small. In 18 of the 25 county groups the number of sample people in the civilian labor force is less than 90. As would be expected, large errors occur in county groups where the sample sizes are small. Actual unemployment rates range from 2.2 to 6.6 percent, while the simple inflation estimates range from 0.0 to 11.6 percent. The correlation coefficient between simple inflation estimates and actual values is .52.

The plot of actual and synthetic unemployment rates is shown in Figure 2. The average squared error of the synthetic estimates is 1.27, much smaller than that of the simple inflation estimator. However, the correlation coefficient of estimates and actual values is only .08. The synthetic estimates cluster around 3.5 percent and range from 3.2 to 3.8 percent. This clustering near the value of the larger area mean is a common characteristic of the synthetic estimator. This is at least partially due to the fact that the variables used to define the α-cells are often not sufficiently correlated with the item being estimated to yield a good estimate for a given small area. When this is true, the magnitude of the bias for a given small area will increase with the difference between the small area parameter and the parameter of the larger area used to produce estimates. These results suggest that the synthetic estimator may be a poor choice if one is interested in either estimating levels of those areas with extreme values or comparing levels between small areas.

The composite estimator, by weighting the simple inflation estimate less heavily when the sample size is small, tends to reduce the large errors of the simple inflation estimates in those areas with small sample sizes; and by weighting the simple inflation estimate more heavily when the sample size is large, tends to reduce the large errors of the synthetic estimator when the actual value of the small area is very different from that of the large area. The plot of the composite estimates is shown in Figure 3. The average squared error is .92, less than that of either the simple inflation or synthetic estimator. The correlation coefficient is .51, essentially the same as that of the simple inflation estimator.

Figures 4, 5, and 6 show the plots of State estimates of percent of the population completing college versus the percent obtained from the 1970 Census for each of the three estimators. Average squared errors are shown in Table 1 and correlation coefficients in Table 2.

States of course have much larger HIS sample sizes than county groups, and this is reflected in the difference between the plots of simple inflation estimates in Figures 1 and 4. Also, as in Figure 1, the large deviations in Figure 4 are generally those of estimates for States with relatively small sample sizes. The average squared error of the simple inflation estimates of the percent completing college is 1.81 and the correlation coefficient between estimate and actual value is .69.

The synthetic estimates for the percent completing college (Figure 5) are more closely clustered around the 45 degree line than the county group synthetic estimates of the unemployment rate (Figure 2). This might be partially due to differences in the predictability of characteristics of States and counties and/or the difference in the number of cells used in the synthetic estimator and the variable estimated. The average squared error of the synthetic estimates of percent completing college is 1.76, essentially the same as that of the simple inflation estimator. The correlation coefficient is .45. The majority of the difference between the correlation coefficients of the simple inflation and synthetic estimators is explained by the point representing the District of Columbia (actual percent 11.2). The observation that the synthetic estimator often does not do well in estimating for certain areas including the District of Columbia has been made before (personal communication, Levy, Gonzalez).

The average squared error of the composite estimates of the percent completing college is 1.09 and the correlation .72. As in the previous example, the composite estimator yields a smaller average squared error than either of the component estimators and also produces a correlation as good as the better of the two component estimators.

IV. Summary

In estimating both the unemployment rates for county groups in Texas and the percent of the population completing college for States the composite estimator has an average squared error approximately 30 percent less than that of the synthetic estimator. With both variables the synthetic estimator has smaller average squared error than does the simple inflation estimator, the other component of the composite estimator. Also, when estimates are correlated with actual values the composite estimator has correlation coefficients as large as those of the simple inflation estimator which are larger than those of the synthetic estimator.

There are, of course, other ways to define weights for the composite estimator. In fact, preliminary results with these and other data indicate that other weighting schemes can produce further reductions in average squared error and further increases in the correlation with actual values. Preliminary results also indicate that the composite estimator is remarkably robust against poor estimates of the unknown quantity b/b'.

The above is only a small empirical study of the performance of three estimators under rather restricted circumstances. However, these results are encouraging, and investigations of the properties of composite estimators are continuing.

References

Ericksen, Eugene P. (1973): "Recent Developments in Estimation for Local Areas." Proceedings of the American Statistical Association, Social Statistics Section, pp. 37-41.

Gonzalez, Maria E. (1973): "Use and Evaluation of Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 33-36.

Gonzalez, Maria E. and Waksberg, Joseph E. (1973): "Estimation of the Error of Synthetic Estimates." Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Gonzalez, Maria E. and Hoza, Christine (1975): "Small Area Estimation of Unemployment." Proceedings of the American Statistical Association, Social Statistics Section, pp. 437-443.

Levy, Paul S. (1971): "The Use of Mortality Data in Evaluating Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 328-331.

Namekata, Tsukasa, Levy, Paul S., and O'Rourke, Thomas W. (1975): "Synthetic Estimates of Work Loss Disability for Each State and the District of Columbia." Public Health Reports, 90, pp. 532-538.

National Center for Health Statistics (1958): "Statistical Design of the Health Household Interview Survey." Health Statistics, Series A-2. Publication No. 584-A2. Public Health Service, Washington, D.C.

Royall, Richard M. (1973): "Discussion of two papers on Recent Developments in Estimation of Local Areas." Proceedings of the American Statistical Association, Social Statistics Section, pp. 43-44.

Royall, Richard M. (1977): "Statistical Theory of Small Area Estimates - Use of Prediction Models." Unpublished report prepared under contract from the National Center for Health Statistics.

Schaible, Wesley L. (1975): "A Comparison of the Mean Square Errors of the Postratified, Synthetic and Modified Synthetic Estimators." Unpublished report, Office of Statistical Research, National Center for Health Statistics.

Schaible, W.L., Casady, B., Schnack, G.A. and Brock, D.B. (1976): "An Empirical Comparison of Some Conventional and Synthetic Estimators for Small Areas." Draft report, National Center for Health Statistics.

Schaible, Wesley L., Brock, Dwight B. and Schnack, George A. (1977): "An Empirical Comparison of Two Estimators for Small Areas." Presented at the Second Annual Data Use Conference of the National Center for Health Statistics. Dallas, Texas.

Schaible, Wesley L. (1977): "Notes on Composite Estimators for Small Areas." Unpublished memoranda, Office of Statistical Research, National Center for Health Statistics.

U.S. Bureau of the Census (1971): General Population Characteristics-Texas, PC(1)-C45, U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census (1972): General Social and Economic Characteristics-Texas, PC(1)-C45, U.S. Government Printing Office, Washington, D.C.

Acknowledgements

The authors would like to thank Barry Peyton and Eugene Diggs of the Office of Statistical Research programming staff for their efforts in computing the estimates and providing the graphical presentations for this paper. Also, the authors thank Kay Barrett and Anita Powell for typing the manuscript.

TABLE 1. Average Squared Errors of the Simple Inflation, Synthetic and Composite Estimators for Two Variables, Health Interview Survey, 1970.

Estimator	Variable	
	Unemployment Rate	Percent Completing College
Simple Inflation	6.85	1.81
Synthetic	1.27	1.76
Composite	.92	1.09

TABLE 2. Correlation Coefficients between Estimate and Actual Value for Three Estimators and Two Variables, Health Interview Survey, 1970.

Estimator	Variable	
	Unemployment Rate	Percent Completing College
Simple Inflation	.52	.69
Synthetic	.08	.45
Composite	.51	.72

FIGURE 1. Unemployment Rates, Simple Inflation Estimates and Actual Values for 25 County Groups in Texas, 1970

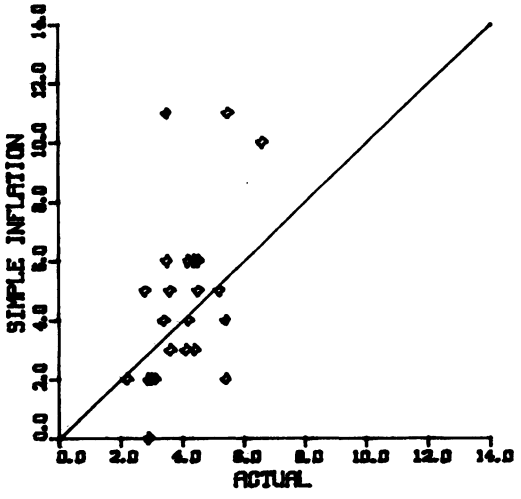

FIGURE 2. Unemployment Rates, Synthetic Estimates and Actual Values for 25 County Groups in Texas, 1970

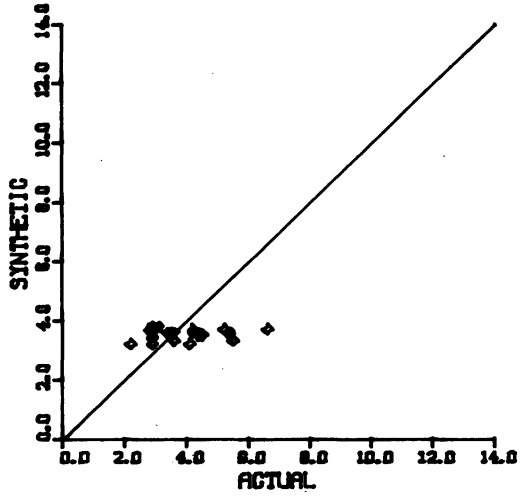

FIGURE 3. Unemployment Rates, Composite Estimates and Actual Values for 25 County Groups in Texas, 1970

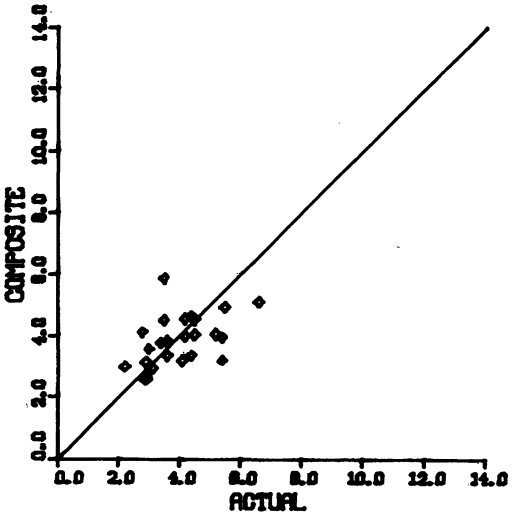

FIGURE 4. Percent of the Population Who Have Completed College, Simple Inflation Estimates and Actual Values for Fifty States and the District of Columbia, 1969-1971

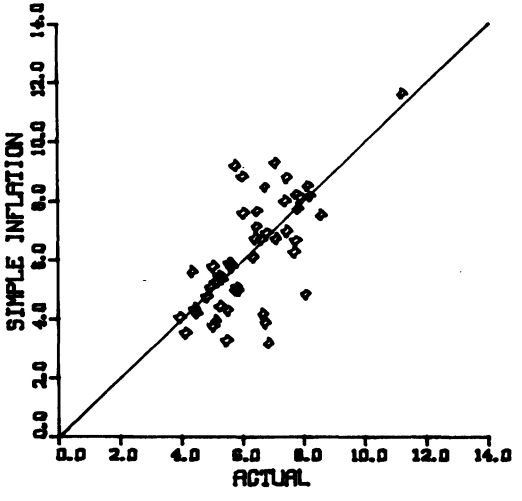

FIGURE 5. Percent of the Population Who Have Completed College, Synthetic Estimates and Actual Values for Fifty States and the District of Columbia, 1969-1971

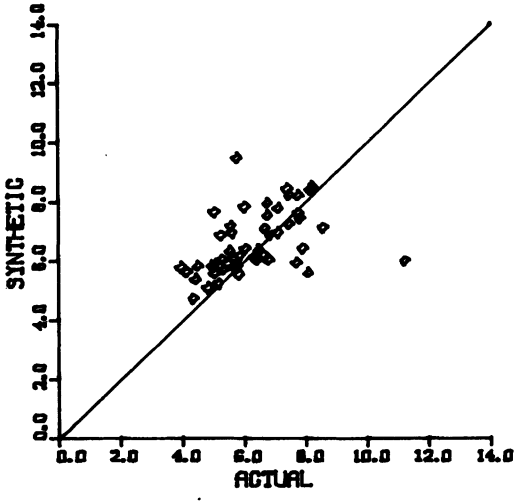

FIGURE 6. Percent of the Population Who Have Completed College, Composite Estimates and Actual Values for Fifty States and the District of Columbia, 1969-1971