A COMPARATIVE STUDY OF SYNTHETIC ESTIMATION STRATEGIES WITH APPLICATIONS TO DATA FROM THE NATIONAL HEALTH CARE EXPENDITURES STUDY

Steven B. Cohen, National Center for Health Services Research

I. Introduction

Planners, researchers, government and the business sector are but a few of the many sources in need of reliable small area statistics to formulate and implement policies at the local level. Timely local area estimates of assorted health, social, political and economic parameters are not directly obtained from sample surveys, primarily due to cost and sample size considerations. To satisfy this need, several alternative strategies have been developed to yield the required small area statistics. These techniques utilize available data at the national level, local data on population, and accessible data on auxiliary variables to derive prediction models. Unlike direct estimators which presume sample data from every local area under consideration, these synthetic estimators utilize auxiliary information (in the estimation equation) from local areas not represented in a respective national survey. Presently, a limited number of such procedures have gained respectability under certain qualifying assumptions, as a consequence of empirical tests of validity and widespread usage.

Although familiarity with a particular small area estimation strategy often dictates its use, certain guidelines should be considered in the selection process. This is a consequence of no technique being consistently superior in precision to its alternatives. The type of available predictor information, its functional relationship with the specified criterion variables and the underlying assumptions of each prediction model helps narrow the choice. Previous empirical studies on similar data sets or related criterion variables, which compare the precision of alternative techniques should be considered.

In this paper, alternative small area estimation strategies are examined as to their potential applicability using data from the National Health Care Expenditures Study. This data base is particularly attractive in its similarity in structure to many of the national survey designs encountered (i.e., HIS, CPS) when using national survey data to synthetically yield subnational estimates. The NHCES was established to provide a detailed assessment of the utilization, costs, and sources of payment associated with medical care in the United States. The survey was designed to provide data for a major research effort in the Division of Intramural Research of the National Center for Health Services Research (NCHSR).

The area sampling design for NHCES can be characterized as a stratified three-stage area probability design from two independently drawn national area samples of the Research Triangle Institute (RTI) and the National Opinion Research Center (NORC). These national samples have the capacity to independently yield unbiased national estimates of relevant health characteristics. The first stage in both designs consists of primary sampling units (PSU's) which are SMSA's, counties, part of counties, or groups of contiguous counties. The second stage consists of secondary sampling units (SSU's) which are census enumeration districts (ED's) or block groups (BG's). Smaller area segments constituted the third stage in both designs, from each of which a sub-sample of households was selected in the final stage of sampling. Combined stage specific sample sizes over the two designs were 135 PSU's (covering 108 separate localities), 1,290 SSU's and 1,290 segments. Sampling specifications required the selection of approximately 14,000 households.

II. Synthetic Estimators

Our comparisons of alternative small area estimation strategies focused upon those techniques whose use was most widespread in the statistical literature. The estimators considered in our study included:

a. the NCHS synthetic estimator,

b. a regression estimator,

c. a post-stratified estimator, and

d. a composite estimator

a. The NCHS Synthetic Estimator

Procedurally, the NCHS method requires the selection of suitable predispositional demographic variables (i.e., race, income, sex, age), and national survey data is used to determine estimates of relevant criterion variables for each of the G mutually exclusive and exhaustive domains defined by the respective demographic cross-classifications. To produce the synthetic estimate of a criterion variable Y for local area ℓ, the NCHS model takes the form of a weighted average

The author wishes to thank Mr. Alan H. Dorfman for his helpful discussion of the paper, Polly Estrin and Deanna Alton for their conscientious typing of this manuscript.

The views expressed in this paper are those of the author and no official endorsement by the National Center for Health Services Research is intended or should be inferred.

$$\bar{Y}^*_{\ell(s)} = \sum_{j=1}^{G} P_{\ell j} \hat{\bar{Y}}_{\cdot j} \qquad (1.1)$$

where $P_{\ell j}$ is the proportion of local area ℓ's population represented by domain j, so that $\sum_{j=1}^{G} P_{\ell j} = 1$ and $\hat{\bar{Y}}_{\cdot j}$ is a probability estimate of the criterion variable for domain j obtained from a national sample. The more detailed estimating equation includes a regional adjustment.

This procedure is particularly appealing as a consequence of its straightforward application to survey data. However, synthetic local area estimates generated in this manner will generally cluster near the mean for a specific geographical region by the nature of their derivation. Consequently, the method is not particularly sensitive to many of the internal forces operating at the local level. By assuming the small areas share the same characteristics as a standard national distribution, they can only be distinguished by their respective demographic configurations. It is also assumed that the estimate for the proportion of each local area's population represented by a specific domain is unbiased and readily available. This condition is satisfied when the target year for which local area estimates are required is a census year. Contrarily, the availability of direct estimates for the $P_{\ell j}$'s diminishes as the target year deviates markedly from the censal date. With respect to the NHCES survey, whose target year was 1977, direct estimates of local area (i.e., state) population totals for specific domains defined by demographic variable cross-classifications were unavailable. Though some marginal distributions of the domain defining variables were available for local areas (i.e., states), we did not wish to resort to raking or synthetic estimation procedures to yield the domain population estimates. These restrictions motivated our consideration of alternative techniques for NHCES applications.

b. A Regression Estimator

Ericksen (1974) developed an alternative technique for computing local area estimates which, unlike the NCHS estimator, solely combines symptomatic information and sample data into a multiple regression format (assuming an underlying linear model). Referred to as the regression-sample data method of local area estimation, the procedure can be outlined as follows:

1. Initially, a sample of n local areas, referred to as primary sampling units, (PSU's), is selected from the N local areas in the population. Estimates of the criterion variable are then computed for the respective PSU's in the sample.

2. Symptomatic information is collected for both sample and nonsample PSU's. Linear least squares regression estimates are computed using data for the sample PSU's only. Estimates for all subareas are then determined by substituting values of the symptomatic indicators, whether included in the respective sample or not.

The model assumes the availability of criterion variable estimates for each of n sample PSU's and the values of p symptomatic indicators for the universe of N local areas. It takes the matrix representation:

$$\hat{Y}^*_{(R)} = XB + u \quad . \quad (1.2)$$

where Y, an nx1 vector, is the criterion variable vector consisting of a set of observed values; X, an nx(p+1) matrix denoting the set of predictor variables; B, the (p+1)x1 vector of regression coefficients; and u, an nx1 vector, a stochastic error term.

c. A Post Stratified Estimator

The method advanced by Ericksen is most feasible when the linearity assumption is satisfied and the observed multiple correlation is high. For those situations when the multiple correlation level is moderate (.4-.7) and a nonlinear model is more appropriate, Kalsbeek (1973) and Cohen (1977) have developed a procedure whose most limiting assumption is the availability of good symptomatic information. It has usually been common practice to treat the local area units as the smallest level for which the estimates are made. Here, contrarily, the local unit is broken up into constituent geographical sectors called "base units," such as counties, enumeration districts, or other geographical subunits of a county. The local area for which a variable of interest is to be estimated is referred to as the "target area" and further subdivided into "target area base units." Unlike other methods which use symptomatic information directly for the purposes of estimation, this procedure uses the information to group base units (sample base units) from the total population.

Initially, a sample of n base units is selected from the total population of N base units. The sample base units are required to possess both symptomatic and criterion information. These units are divided into K groups (strata) using either or both types of the information available. The object is to form groups which are most homogeneous within while dissimilar between themselves. Grouping can be handled by any one of several interactive procedures in cluster analysis (i.e., Automatic Interaction Detector (A.I.D.), Multivariate Interactive K-Means Cluster Analysis (MIKCA)), or minimum variance stratification schemes (cum $\sqrt{f}$ rule). It is noteworthy that the respective groups may be defined by either rectilinear or non rectilinear boundaries.

All "target area base units" belonging to the local area in question are then assigned (classified) to one of the K groups with respect to symptomatic information. Consequently, each "target area base unit" is associated with a group of base units both similar to itself and internally homogeneous. An estimate for each of the "target area base units" with respect to the criterion variable is obtained from the sample base units in the group to which it has been assigned. In essence, each target area base unit estimate can be perceived as a synthetic estimate. These synthetic estimates are then pooled to arrive at a final estimate for the respective target area. More specifically, the respective group (strata) estimate of the criterion variable of interest takes the form:

$$\hat{Y}_g = \sum_{i \epsilon g} W_i \hat{Y}_i \qquad (1.3)$$

where W_i estimates the proportion of the total population of sample base units classified in group g that are represented by base unit i,

and $\hat{Y}_i$ is an estimate of the criterion variable of interest for the ith sample base unit.

The estimate of the criterion variable for each local area (ℓ) takes the form:

$$\bar{Y}^*_s = \sum_g P_{\ell_g} \hat{Y}_g \qquad (1.4)$$

where p_{ℓ_g} estimates the proportion of local area ℓ's population classified in group g.

The estimator bears a striking resemblance to the NCHS synthetic estimator previously discussed. The primary difference concerns the method of group or domain formation. For the NCHS method, domain estimates are generated once individuals are classified according to their demographic characteristics. Alternatively, this procedure links all individual observations within a sample base unit to a particular group, based on the unit's symptomatic information. Population estimates for the p_{ℓ_g} defined in this manner are more readily obtainable than the small area demographic distributions required for the NCHS estimator.

d. A Composite Estimator

A final method which has gained recent attention is the composite estimator which takes the form of a weighted average of two component estimators:

$$\bar{Y}^*_c = C_1 \bar{Y}^*_I + C_2 \bar{Y}^*_{II} \qquad (1.5)$$

where C_1 and C_2 are appropriately chosen weights and $\bar{Y}^*_I$ and $\bar{Y}^*_{II}$ are alternative synthetic estimators. It has the property of a mean square error that is smaller than the larger of the mean square errors for the two component estimators. Given the variety of alternative synthetic estimators available, one often does not have a priori information to determine the most appropriate and precise method for his data. The composite estimator is a welcome alternative that minimizes the selection of a less precise technique. Schaible, Brock, and Schnack (1977) have demonstrated that with a judicious selection of the composite weights, the composite estimator will have a mean square error which is smaller than either of the component estimators. They have shown that when $E(\bar{Y}^*_I - \bar{Y})(\bar{Y}^*_{II} - \bar{Y})$ is small relative to MSE $(\bar{Y}^*_{II})$, the appropriate weights that will minimize the composite estimator's mean square error can be approximated by

$$C_I = \frac{1}{1+R}, \quad C_{II} = 1 - C_I, \text{ where}$$

$$R = \frac{\text{MSE } (\bar{Y}^*_I)}{\text{MSE } (\bar{Y}^*_{II})}$$

Here, the respective weights have been normalized so that they sum to unity.

III. A Comparison of Alternative Synthetic Estimation Strategies

A major limitation one confronts when conducting reliability studies of alternative small area estimation strategies is the unavailability of independent local level estimates of relevant criterion variables. Consequently, measures of bias and precision such as mean square errors are often difficult to directly estimate. Fortunately, design features of the National Health Care Expenditures Study do not subject our study to such constraints. Here, the complex nature of the survey design, which is a stratified multistage area probability design from two independently drawn national samples, is ideally suited for the diverse methodologies in small area estimation. The replicated property of the survey design will yield independent unbiased direct estimates for several local areas and allow for a reliability study of the alternative strategies.

To implement a comparison of the alternative methods of small area estimation, the NHCES PSU's (primarily SMSA's and counties) were taken as base units for the Kalsbeek-Cohen strategy, and as observational units in the regression strategy. Two considerations affected this choice of units for the alternative estimators:

1. the availability of independent unbiased estimates of a select number of PSU's, and

2. the capacity to demonstrate the use of PSU's as elemental units in synthetic models for estimates of larger local areas (i.e., states) will result in estimated bias terms, measures of precision and deviations between synthetic estimators that are functions of PSU differences.

More specifically, data on insurance coverage for 1977 was selected from the NHCES data base to derive unbiased PSU estimates of the following dependent variables:

1. percent of population ever on Medicaid,

2. percent of population ever on Medicare,

3. percent of population ever covered with private insurance,

4. percent of population ever uninsured.

Subsequently, only the 76 NORC sample PSU's were used in the determination of synthetic estimates for the alternative methods. These PSU's served as the sample base units for the Kalsbeek-Cohen model and as observational units in the regression approach. Our composite estimator was a weighted function of these estimators. A screening technique (a'la stepwise regression) was implemented to determine those symptomatic variables with the greatest predictive capacity for the respective measures of insurance coverage among several relevant symptomatic variables selected from the 1977 County City Data Book (a U.S. Census Bureau data source). For medicaid coverage, measures of population density, income, and the number of children were chosen. For medicare coverage, measures of the aged population, income, and facility supply were selected. Similarly, measures of the aged population, physician supply, and the birth rate were predictors for private insurance. Income, physician supply, and birth rate were also predictors for the uninsured. The percent of variation explained by each of these predictor variable models (R^2) can be observed in Table I. One can note only a moderate level of multiple correlation. With the selection of appropriate symptomatic data, synthetic estimates of the different measures of insurance coverage could easily be derived for the 59 PSU's represented in the RTI sample. Since unbiased estimates of insurance coverage for the 59 RTI PSU's were available from the NHCES sample, estimates of bias and mean square error could also be derived.

The small number of NORC PSU's (76) restricted the number of groups formed by the Kalsbeek-Cohen technique to eight to insure an average sample size of 9 for each cluster estimator. A minimum variance stratification scheme which employed the cum $\sqrt{f}$ rule on the marginal distributions of the symptomatic variables was used in group formation. Previous studies by Cohen (1978) have demonstrated that further increases in the number of clusters at the expense of reduced numbers of sample base units coincide with diminishing returns with respect to further proportional reductions in the estimator's mean square error.

Our composite estimator was a weighted function of the Ericksen and Kalsbeek-Cohen estimators. Here, we considered two sets of weights with respect to the composite estimator. One choice approximated the set which would minimize the composite estimator's mean square error. The approach assumes the availability of unbiased estimates of the criterion variable under-consideration in the determination of mean square errors for the component synthetic estimators. Most applied settings for synthetic estimation will be deficient in this requirement. Application of alternative weights which are constants set at $C_I=C_{II}=1/2$ was the alternative scheme we considered (to average the component estimators).

To measure the precision of the respective small area estimators, we estimated the average mean square error for the synthetic PSU estimators. An unbiased estimator of the mean square error for each synthetic estimator takes the form:

$$MSE\ (\hat{Y}^*_{(s)}) = (\hat{Y}^*_{(s)} - \hat{Y})^2 - \sigma^2_{\hat{Y}} \qquad (1.6)$$

where $\hat{Y}$ is an unbiased estimator of a criterion variable for a specific PSU,

$\sigma^2\ \hat{Y}$ an unbiased estimate of the variance of $\hat{Y}$ and

$\hat{Y}^*_{(s)}$ the synthetic estimate of a criterion variable for the same PSU.

Consequently, the average mean square error over M=59 PSU's was estimated as:

$$AMSE(Y_{(s)}) = \frac{\sum_{i=1}^{M} (\hat{Y}^*_{i(s)} - \hat{Y}_i)^2}{M} - \frac{\sum_{i=1}^{M} \sigma^2_{\hat{Y}}}{M} \qquad (1.7)$$

Coefficients of variations for the alternative synthetic estimators were approximated to serve as an additional measure of precision. Here the estimated coefficient of variation takes the form:

$$C.V.(Y_{(s)}) = \frac{\left[AMSE(Y_{(s)})\right]^{\frac{1}{2}}}{Y_{(s)}} \qquad (1.8)$$

Comparisons of average mean square errors and coefficients of variation between alternative synthetic estimators can be observed in Table II and Table III. These measures of variation indicate a consistent gain in precision for the Ericksen technique over the Kalsbeek-Cohen procedure for all insurance indicators. Although the R^2 levels for most regression models were generally only moderate, there was no indication of model "lack-of-fit" for the assumed linear relationship. Simultaneously, the relatively small number of PSU's used in the estimation of strata means via the clustering strategy for the reliability study reduced the potential number of strata in the model. Use of the entire 135 NHCES PSU's over the 76 NORC PSU's in strata formation would have improved the discrimination and predictive capacity of the model. In this setting, however, the regression approach might still have proven more appropriate.

As anticipated, the composite estimator whose weights were selected to minimize the estimator's mean square error was consistently more precise than the alternative estimators. Application of this strategy, however, requires *a priori* estimates of the mean square errors of the component estimators. As noted, most applied settings for synthetic estimation will be deficient in this requirement. The alternative composite estimator under consideration ($C_I=C_{II}=1/2$) was less precise than the regression model but more accurate than the clustering strategy. This method always allows one the advantage of selecting an estimator which is more precise than the component estimator with the largest mean square error. Appropriate assumptions regarding the respective component estimators' mean square errors, even when sample estimates are unavailable, increases the likelihood of weight selections which resemble the optimal weights. Since composite estimators have the property of being rather insensitive to large errors in weight specifications, weight selections which allow mean square error estimates to vary across synthetic estimates based on variability estimates from related studies may yield a composite estimator more precise than either of its components (Schaible, 1977). When good judgment is possible with the choice of component estimators and weights, our study indicates the composite estimator is the recommended synthetic estimation strategy for adoption among all relevant and appropriate alternative techniques.

IV. Summary

To summarize, reliable estimates of parameters at the local level are difficult, if not impossible, to obtain directly from sample surveys, primarily due to the constraints of sample size and design. Yet, the very nature of the problem has served as the motivating force in the development of several alternative procedures. The selection of the optimal strategy is a function of the type of symptomatic information available, the appropriateness of a specified prediction model, and the number of available sample observational units (i.e., PSU's). Once appropriate and practicable strategies have been determined, the use of a composite estimator with judiciously chosen weights will avoid the chance of selecting the least precise technique and will potentially yield an estimate of superior precision.

V. References

Cochran, W. G. *Sampling Techniques*. New York: John Wiley and Sons, 1963.

Cohen, S. B. and Kalsbeek, W. D. "An Alternative Strategy for Estimating the Parameters of Local Areas" *American Statistical Association; Proceedings of the Social Statistics Section* (1977), 781-786.

Cohen, S.B., (1979). A Modified Approach to Small Area Estimation. *NIDA Research Monograph 24: Synthetic Estimates for Small Areas*, 98-134.

Ericksen, E. P. "A Regression Method for Estimating Population Changes of Local Areas." *Journal of the American Statistical Association*, 69 (1974), 867-875.

Gonzalez, M. E. and Waksberg, J. "Estimation of the Error of Synthetic Estimates." Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, 1973.

Gonzalez, M. E. "Use and Evaluation of Synthetic Estimates." *American Statistical Association, Proceedings of the Social Statistics Section* (1975).

Kalsbeek, W. D. "A Method for Obtaining Local Postcensal Estimates for Several Types of Variables." Unpublished doctoral dissertation, University of Michigan (1973).

National Center for Health Statistics. *Synthetic Estimates of Disability*, PHS publication, No.1759, 1968.

Schaible, W.L., Brock, D.B., and Schnack, G.A. An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Areas. *Proceedings of the American Statistical Association Social Statistics Section*, 1017-1021.

Schaible, W.L. A Composite Estimator for Small Area Statistics. *NIDA Research Monograph 24: Snythetic Estimates for Small Areas* 36-53, 1979.

TABLE I

Selected Predictor Variables for Respective Criterion Variables

Percent of Population	Symptomatic Variable	R^2
On Medicaid	Population per sq-mi., SSI Recipients, Birth Rate/1000	.2581
On Medicare	Persons 65 & over, Hospital Beds/100,000, Per Capita Money Income	.3557
With Private Insurance	Persons 65 & over, Birth Rate/1000, Physicians/100,000	.2175
No Health Insurance	Birth Rate/1000, Physicians/100,000, Per Capita Money Income	.3337

TABLE II

Average Mean Square Errors of Alternative Synthetic Estimators for Health Insurance Coverage Variables, National Health Care Expenditures Study, 1977

		Average Mean Square Error		
Percent of Population	Kalsbeek-Cohen	Regression	Composite $(C_I=C_{II}=1/2)$	Composite $(C_I=\frac{1}{1+R})$
On Medicaid	40.17	35.21	35.80	31.98
On Medicare	17.91	13.24	14.48	10.36
With Private Insurance	124.77	106.69	111.16	96.30
No Health Insurance	64.22	47.27	51.52	39.87

TABLE III

Approximate Coefficients of Variation for the Alternative Synthetic Estimators

Percent of Population	Kalsbeek-Cohen	Regression	Composite $(C_I=C_{II}=1/2)$	Composite $(C_I=\frac{1}{1+R})$
On Medicaid	.6194	.6361	.6130	.5932
On Medicare	.3686	.3076	.3264	.2743
With Private Insurance	.1491	.1382	.1409	.1307
No Health Insurance	.3235	.2759	.2889	.2567