WHEN ROBUST ESTIMATION IS NOT AN OBVIOUS ANSWER: THE CASE OF THE SYNTHETIC ESTIMATOR VERSUS ALTERNATIVES FOR SMALL AREAS

Carl Erik Sarndal, University of Montreal

Summary. The synthetic estimator and alternative small area estimators are examined. A robust, approximately unbiased alternative is developed that borrows strength in the same way as the biased synthetic estimator. However, the special problems of small estimation are such that it is not a foregone conclusion that a robust method should take precedent over non-robust possibilities.

The paper emphasizes that the choice of estimation method depends on a complex interplay of factors, including sample size, sampling fraction, area smallness and departure from a basic model assuming that small areas behave like large areas.

Key words: Synthetic estimator, small areas, generalized regression approach, regression model, robustness.

1. INTRODUCTION

Most standard dictionnaires give at least two meanings of the word "synthetic": (a) something pertaining to or consisting in synthesis, in contrast to analysis, that is, the composition or combination of parts, elements, etc. so as to form a whole; (b) something pertaining to or formed by artificial synthesis, hence, not genuine, artificial.

Gonzạlez (1973) describes the method of synthetic estimation as follows: "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates." One can view this statement as making reference to the combination-of-parts aspect as well as to the artificality aspect of "synthetic": Class mean estimates Y_{s_h} (h=1,...,H) for larger areas are combined to build the synthetic estimator for small area a, $T_{SY} = \sum_{h=1}^{H} W_{ah} Y_{s_h}$ (National Center for Health Statistics, 1968). When the assumption that small areas resemble large areas fails, T_{SY} becomes design biased, therefore artificial for the convinced probability sampler. Despite the bias, the probability sampler often gambles on a synthetic estimate, because strength will be "borrowed" if the assumption holds. However, subtle departures from the assumption puts the whole method in question. We shall show in detail how the quality of T_{SY} deteriorates under departures.

Our results stem from research aiming at answering the following questions: 1) What does the concept of "borrowing strength" mean more precisely? 2) Can robust, approximately design unbiased methods be found that borrow strength in precisely the same way as the biased synthetic estimator? 3) For what small area sizes, sample sizes and other factors will the robust method, if it exists, be better than the synthetic estimator? (It is safe to assume that a robust method be conveniently estimated?

The paper also thoroughly examines the synthetic estimator itself. Its properties are now known in depth, partly due to difficulties of evaluation referred to below.

The answer to question 2 turns out to be "yes". A method is proposed which builds a new, robust small area estimate from exactly the same building blocks used by T_{SY}, but a different assembly of the blocks makes the design bias negligible in large samples. Thereby the artificiality (the design bias) of synthetic estimator is essentially removed. The new, robust method is compared with T_{SY} and other possibilities. Variance estimation of the new method is discussed.

Probability samplers are sometimes criticized for their strong preference for consistent, (approximately) design unbiased methods. If small MSE can be achieved by design biased methods, why not use them? The probability sampler's traditional answer (Hansen, Madow and Tepping, 1978) lies in robustness. But faced with the special difficulties of small area estimation, some probability samplers seem tempted to abandon their usual principles. The small area problem becomes an interesting testing ground for the opposition between "small MSE at any price" and "negligible design bias at the price of somewhat larger MSE".

2. NOTATION.

Suppose that the finite population U, containing N units labelled K = 1,...,N is divided into A mutually exclusive and exhaustive small areas (domains) labelled a = 1,...,A. For each small area, units are further classified into L mutually exclusive and exhaustive classes, which could be based on age, sex, race, etc., they are labelled h = 1,...,H. The population, and the sample drawn from it, will thereby be completely cross-classified into AH cells.

Let U_{ah} (s_{ah}), of size N_{ah} (n_{ah}), denote the set of units in the population (sample) that fall in cell ah. The N_{ah} are assumed known from a previous census or other accurate source. For aggregations across areas, let

$$U_h = \mathop{}_{a=1}^{A} U_{ah} \quad (s_h = \mathop{}_{a=1}^{A} s_{ah})$$

denote the set of units in the population (in the sample) that fall in class h of size

$$N_{.h} = \sum_{a=1}^{\bar{A}} N_{ah} \quad (n_{.h} = \sum_{a=1}^{A} n_{ah}).$$

For aggregations across groups, let

$$U_a = \mathop{}_{h=1}^{H} U_{ah} \quad (s_a = \mathop{}_{h=1}^{H} s_{ah})$$

and

$$N_{a\cdot} = \sum_{h=1}^{H} N_{ah} \quad (n_{a\cdot} = \sum_{h=1}^{H} n_{ah})$$

denote, respectively, the set of population (sample) units in area a and its size. For aggregations across groups and cells, let

$U = \overset{H}{\underset{h=1}{\cup}} U_h$ ($s = \overset{H}{\underset{h=1}{\cup}} s_h$); $N = \overset{H}{\underset{h=1}{\Sigma}} N_{.h}$ ($n = \overset{H}{\underset{h=1}{\Sigma}} n_{.h}$) be the entire population (sample) and its size.

Associated with the k^{th} unit is the quantity Y_k. We write $\Sigma_{U_{ah}} Y_k$, $\Sigma_{U_h} Y_k$, $\Sigma_{s_{ah}} Y_k$ etc. to denote sums over $k \in U_{ah}$, $k \in U_h$, $k \in s_{ah}$, etc. The plain symbol Σ is reserved for the frequently needed sum $\overset{H}{\underset{h=1}{\Sigma}}$. Means are denoted as follows: At the population level, Y_{ah} for $\Sigma U_{ah} Y_k/N_{ah}$, $Y_{.h}$ for $\Sigma U_h Y_k/N_{.h}$, $Y_{.a}$ for $\Sigma_{U_a} Y_k/N_{.a}$; at the sample level, $Y_{s_{ah}}$ for $\Sigma_{s_{ah}} Y_k/n_{ah}$, Y_{s_h} for $\Sigma_{s_h} Y_k/n_{.h}$, etc. Sampling fractions are denoted $f_{ah} = n_{ah}/N_{ah}$ at the cell level, $f_h = n_{.h}/N_{.h}$ at the class level; also, set $f_h' = (n_{.h}-1)/(N_{.h}-1)$. The size of cell ah relative to its area is $W_{ah} = N_{ah}/N_{a.}$; the relative size of class h is $W_h = N_{.h}/N$. The size of cell ah relative to its class is $G_{ah} = N_{ah}/N_{.h}$, and $G_{ah} > 0$ for all a,h is assumed.

If the population is stratified and stratified sampling is used, a third classification will further complicate the notation. We limit our discussion to stratified random sampling with strata identical to the classes. This design is denoted by strs and implies the inclusion probabilities $\pi_k = n_{.h}/N_{.h}$ for every unit $k \in U_h$ ($h = 1,\ldots,H$). Briefly discussed is the design of simple random sampling (srs) with $\pi_k = n/N$ for all $k \in U$.

Part of the dilemma is that the survey has not been conducted with the goal of efficient estimation for the small areas particularly in mind. A common consequence is a shortage of observations in a small area. The relative size of the small area is evidently a key element in the problem.

Purcell and Kish (1980), warnings against the mistake of considering small area estimation as one homogeneious problem, suggested that size of the small area will influence the choice of method. They classified small domains into (1) Major domains, composing 1/10 of the population or more; (2) Minor domains, comprising between 1/10 and 1/100 of the population; (3) Mini domains, comprising between 1/100 and 1/100000 of the population; (4) Rare domains, comprising less than 1/100000 of the population. An example of very small domains involves the 38,000 U.S. federal revenue sharing districts. We show how area size influences the character of an estimator. A goal of future research is to give more precise rules for the estimation technique appropriate for rare domains, for mini domains, etc.

3. DESIGN BIASED ESTIMATION

The synthetic estimator

$$T_{SY} = \Sigma W_{ah} \overline{Y}_{s_h} \qquad (3.1)$$

estimates the small area mean $\overline{Y}_a$ with a design bias, under either srs or strs, of $\Sigma W_{ah} (Y_{.h}-Y_{ah})$. In the back of the probability sampler's mind is the hope that $\overline{Y}_{ah} = \overline{Y}_{.h}$ for each h; then vanishing bias and small variance makes T_{SY} efficient. Now T_{SY} is recommended particularly when areas are so small that stable estimation of the cell means $\overline{Y}_{ah}$ is difficult; this is the situation we primarily have in mind in this paper, however, class sample sizes $n_{.h}$ are assumed to be rather large. If the $\overline{Y}_{ah}$ could be easily estimated, the (post-) stratified estimate $\Sigma W_{ah} \overline{Y}_{s_{ah}}$, design unbiased under srs and strs, would appeal to the probability sampler. The "simple direct" estimate $\Sigma n_{ah} \overline{Y}_{s_{ah}}/n_{a.}$, design unbiased under srs , can always be calculated, but its variance is usually large. Some comments on T_{SY} are:

1. Design bias. The probability sampler relunctantly embraces T_{SY}. He is willing to tolerate its design bias only because meaningful estimation seems otherwise impossible; T_{SY} is preceived as "...a dangerous tool, but with careful further development, it has attractive potential". (Simmons, 1979).

2. Estimation of the bias of T_{SY}. Levy (1979) points out that "the bias of T_{SY} can not be estimated from the data used to construct it". This may be true; however, the robust method T_{RB} below shows that the bias can be reduced to one that vanishes with increased sample size.

3. Problems of evaluation. Ericksen (1979) justifiedly claims that "the accuracy of synthetic estimates has not usually been assessed and we do not have a systematic method which could tell us how inaccurate or biased the estimates might be". Some reported evaluations are empirical; Levy (1971), Gonzalez (1973), Gonzalez and Hoza (1978), Schaible (1979). Here we evaluate by means of model-expected design MSE, which reveals that in cloosing an estimator, the statistician confronts a complex interaction of factors: sample size, sampling fraction, area smallness, degree of departure from the supposed ideal model.

Holt, Smith and Tomberlin (1979) maintain that T_{SY} is an intuitive and ad hoc estimator. They, as well as Laake (1979), carried the idea of homogeneity of cell means to its full logical consequence by first posing the super population model ξ_0 such that

$$E_{\xi_0}(Y_k) = \beta_h \; ; \; V_{\xi_0}(Y_k) = \sigma_h^2 \qquad (3.2)$$

for all $k \varepsilon U_h$ $(h = 1,\dots,H)$. (The Y_k's are assumed independent throughout.) They then derived the prediction approach estimator, which for every fixed s minimizes $E_{\xi_0}(T-\bar{Y}_{a.})^2$ given that $E_{\xi_0}(T-\bar{Y}_{a.}) = 0$; their result is

$$T_{MOD} = \Sigma W_{ah} \bar{Y}_{s_h} + \Sigma W_{ah} f_{ah} (\bar{Y}_{s_{ah}} - \bar{Y}_{s_h}). \tag{3.3}$$

It is composed of T_{SY} plus a second term working slightly in the direction of reducing the design bias, which is given under strs by (6.2). Thus T_{MOD} hardly alleviates the bias problem; it shows, however, that a combination of class means $\bar{Y}_{s_h}$ and cell means $\bar{Y}_{s_{ah}}$ may have some merit. A zero cell frequency n_{ah} nullifies the contribution $W_{ah} f_{ah} (\bar{Y}_{s_{ah}} - \bar{Y}_{s_h})$.

REFERENCES

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 63, pp. 615-620.

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd edition. New York: Wiley.

Ericksen, E.P. (1979). Discussion of paper by R.E. Fay. In: Synthetic Estimates for Small Areas, ed. J. Steinberg, NIDA Research Monograph 24. Rockville, Maryland: National Institute on Drug Abuse, pp. 185-190.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. Proceedings American Statistical Association, Social Statistics Section, pp. 33-36.

Gonzalez, M.E. and Waksberg, J. (1973). Estimation of the error of synthetic estimates. 1st meeting, International Association of Survey Statisticians, Vienna.

Gonzalez, M.E. and Hoza, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, pp. 7-15.

Hansen, M.H. and Tepping, B.J. (1978). Foundations of inference in survey sampling. Proceedings American Statistical Association, Social Statistics Section.

Holt, D., Smith, T.M.F. and Tomberlin, T.J. (1979). A model based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, pp. 405-410.

Laake, P. (1979). A prediction approach to subdomain estimation in finite populations. *Journal of the American Statistical Association,* 74, pp. 355-358.

Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. Proceedings of the American Statistical Association, Social Statistics Section, pp. 328-331.

Levy, P.S. (1979). Small area estimation - synthetic and other procedures, 1968-1978. In: Synthetic Estimates for Small Areas, ed. J. Steinberg, NIDA Research Monograph 24. Rockville, Maryland: National Institute on Drug Abuse, pp. 4-19.

National Center for Health Statistics (1968). Synthetic State Estimates of Disability PHS Publication No. 1759. Public Health Service, Washington: U.S. Government Printing Office.

Purcell, N.J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, pp. 3-18.

Sarndal, C.E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. In press, *Biometrika*.

Schaible, W.L. (1979). A composite estimator for small area statistics. In: Synthetic Estimates for Small Areas, ed. J. Steinberg, NIDA Research Monograph 24. Rockville, Maryland: National Institute on Drug Abuse, pp. 36-53.

Schaible, W.L., Brock, D.B. and Schnack, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. Proceedings of the American Statistical Association, Social Statistics Section, pp. 1017-1021.

Simmons, W.R. (1979). Discussion of paper by P.S. Levy. In: Synthetic Estimates for Small Areas, ed. J. Steinberg, NIDA Research Monograph 24. Rockville, Maryland: National Institute on Drug Abuse, pp. 20-23.