EVALUATION OF EFFICIENCY OF SYNTHETIC ESTIMATES

P.D. Ghangurde and M.P. Singh
Statistics Canada

1. Introduction

In many large scale household surveys, data are obtained by sample designs involving geographic stratification and area sampling. The sample size of these surveys is large enough to yield estimates of reasonable accuracy at national and provincial levels. Also, smaller areas composed of complete strata do not pose a problem in estimation, though the reliability of estimates for these areas may be low. If small areas cut across design strata and are thus composed of areal domains within strata, the reliability of design-based estimates for these areas is severely reduced due to the effect of clustered sample design used in these surveys. In cluster sampling,ratio estimates for total or mean per unit for areal domains can be biased due to significant probability of having no sampled clusters in the domains. For the same reason,the unbiased estimates of totals have large variances (see Kish and Frankel [9]). In particular cases where a domain has no sampled clusters, both estimates become impractical.

Previous work on evaluation of efficiency of synthetic estimates in large scale household surveys has been done on the Health Interview Survey conducted by the U.S. National Center of Health Statistics (see e.g. Levy and French [10], Schaible, Brock and Schnack [14]), Current Population Survey (see Gonzales [5], Gonzales and Hoza [14]) and Australian Labour Force Survey (Purcell and Linacare [12]). The method of synthetic estimation uses knowledge of population structure for improving efficiency of design-based estimates. The knowledge of population structure has been used in sample surveys in design (as in probability proportional to size (pps) sampling) and in estimation (as in ratio and regression estimation) where some relationship between an estimation variable and an auxiliary variable is assumed. In household surveys, the characteristics of interest are usually counts or proportions of various attributes, the basic assumption in synthetic estimation being that of homogeneity of these counts in socio-economic or demographic subgroups in the population.

This paper considers the problem of evaluation of efficiency of synthetic estimates for cluster sampling with probability proportional to size, which is the basic design used in many household surveys. Previous papers on evaluation of efficiency have assumed simple random sampling. The evaluation is carried out by simulating the sample design of the Canadian Labour Force Survey (LFS) and by using the 1971 Census data. The parameters in the expressions for bias and efficiency under super-population models are estimated by weighted least squares using data of the characteristics and population obtained in the census. By assuming the finite population to be a sample from a super-population, the effect of population growth and heterogeneity of the population on the bias and efficiency is examined and the assumptions underlying synthetic estimation are evaluated. A measure of relative accuracy of the synthetic estimate as compared to the design-based estimate, for a set of strata which include the domain, is suggested and its values are obtained for the LFS design using census data. The bias and variance of the synthetic estimate with ratio adjustment based on projected population in a large area, are also derived for any sample design within strata.

2. A Short Review of Literature and Approaches to Evaluation

The problem of small area estimation can be posed as a problem of estimation of counts of a population in the cells of a contingency table of three dimensions - areal domains J, categories of interest X (e.g. unemployed, employed, not in labour force) and subgroups A (e.g. age-sex groups). The estimates for domains can be obtained from survey estimates in X x A cells by applying the Deming-Stephan Iterative Proportional Fitting (IPF) algorithm to adjust these estimates to conform to known marginal totals (based on the last census) for A and J (see Chambers and Feeney [1] and Freeman and Koch [3]). Chambers and Feeney [1] investigate the optimality properties of estimates obtained by IPF imposing the two constraints of additivity to known marginal totals for categories of A and J and show that these estimates preserve the structure of the population as given by interactions of J, X and A.

The problem of lack of accurate intercensal population projections for small areas is well-known (see e.g. Eriksen [2]). The estimates obtained by imposing the constraint of additivity to A totals only have been called synthetic estimates. These estimates can be defined as weighted means of survey estimates in X x A cells, the weights being the proportion of population as of the last census, in the categories of A.

Royall [13] and Holt, Smith and Tomberlin [8], model the structure of the population under various assumptions and obtain the best linear unbiased estimates of various parameters in the framework of the linear model theory. The synthetic estimates for domains are based on the estimates of parameters obtained under the assumption of equal probability sampling from infinite population. The expressions for bias and variance of these synthetic estimates have been derived in these two papers.

A problem in evaluation of synthetic estimates lies in the difficulty in estimation of their bias and efficiency. The approach used in estimation of bias in a number of studies is to obtain the average difference of synthetic estimates and the census-based true values for a number of domains. In the case of conceptual differences between the characteristic of interest in a survey and in the census, these estimates may not be realistic. The efficiency of synthetic estimates for simple random sampling has been evaluated in a number of studies on the basis of comparison of estimates of mean square error of

the synthetic estimate with that of variance of an unbiased estimate. A problem with this approach is that the variance of the unbiased estimate is very large due to small sample in the domain. Gonzales and Waksberg [7], therefore, introduced the concept of evaluation of average mean square error for a set of synthetic estimates and developed an estimate of the average mean square error.

3. Synthetic Estimate

In household surveys, many characteristics are homogeneous within demongraphic or socio-economic subgroups, but due to the lack of availability of frames for these subgroups, the homogeneity cannot be exploited in the design through stratification. However, these subgroups are used as post-strata at the estimation stage. Also, ratio estimation with population within subgroups as an auxiliary variable, is used to improve efficiency of design-based unbiased estimates within post-strata.

We introduce synthetic estimation for stratified sampling with an unspecified design within strata. Let the population in a large administrative area p be divided into a number of geographic strata.

We consider a post-stratified ratio estimate $\hat{X}_{\underline{h}}$, for characteristic total $X_{\underline{h}}$, given by

$$\hat{X}_{\underline{h}} = \sum_a \Big(\sum_{h \in \underline{h}} \hat{X}_{ha} \Big) \cdot \frac{P_{pa}}{\hat{P}_{pa}}, \qquad (3.1)$$

where

$\hat{P}_{pa}$ = design-based estimate of population in p and subgroup a,
P_{pa} = projected population in p and subgroup a,
$\hat{X}_{ha}$ = design-based estimate of characteristic total in stratum h and subgroup a, and
$\underline{h}$ = a group of strata in p.

When $\underline{h} = p$, (3.1) is the usual post-stratified ratio estimate combined over strata in p. The population projections, P_{pa}, based on the last census are available for large administrative areas like provinces. However, for small areas like groups of counties, population changes are significantly affected by many factors and hence projections of high enough reliability are not easily available for such areas. When the small area consists of a group of strata, no special problem in estimation arises. If the small area cuts across stratum boundaries, the usual domain estimate corresponding to $\hat{X}_{\underline{h}}$ is given by

$$^{j}\hat{X}_{\underline{h}} = \sum_a \Big(\sum_{h \in \underline{h}} {}^{j}\hat{X}_{ha} \Big) \frac{P_{pa}}{\hat{P}_{pa}}, \qquad (3.2)$$

where $^{j}\hat{X}_{ha}$ is design-based estimate based on sampled units in the domain in stratum h and subgroup a. This estimate for a small area total has been found to be very unstable, particularly in clustered sample designs. For example, if no sampled units belong to any of the domains in $\underline{h}$, the estimate is zero.

We now define a synthetic estimate $^{j}\tilde{X}_{\underline{h}}$ for estimating characteristic total $^{j}X_{\underline{h}}$ in domain j in $\underline{h}$. The estimate is given by

$$^{j}\tilde{X}_{\underline{h}} = \sum_a \Big(\sum_{h \in \underline{h}} \hat{X}_{ha}\, {}^{j}W_{ha} \Big) \frac{P_{pa}}{\hat{P}_{pa}}, \qquad (3.3)$$

where $^{j}X_{ha}$ = proportion of population of subgroup a in the domain j from stratum h as of the last census. With $^{j}W_{ha}$ as defined above, it is possible to evaluate the effect of population growth on the bias and efficiency in the framework of super-population models. The estimate (3.3) has the same form as that in Gonzales and Hoza [6] and Levy and French [10] except for the ratio adjustment. The ratio adjustment factor, $P_{pa}/\hat{P}_{pa}$, based on the projected population in large area p makes the small area estimates for any characteristic additive over the large area p. This holds for all the three estimates defined above.

The effect of the ratio adjustment on the bias and variance of synthetic estimate is discussed in the following section, where expressions for bias and variance of the synthetic estimate (3.3) are derived for any sample design within strata. For simplicity, the superscript j in $^{j}W_{ha}$ has been dropped in later sections.

4. Bias and Variance

We shall use the well-known truncated Taylor series approximation (Woodruff [15], Tepping [16]) in deriving bias and variance of the synthetic estimate. We consider

$$^{j}\tilde{X}_{\underline{h}} - {}^{j}X_{\underline{h}} = \sum_a \Big(\sum_{h \in \underline{h}} \frac{\hat{X}_{ha}\, W_{ha}}{\hat{P}_{pa}} - \frac{{}^{j}X_{\underline{ha}}}{P_{pa}} \Big) P_{pa}. \qquad (4.1)$$

By first order Taylor series approximation to $\hat{X}_{ha}/\hat{P}_{pa}$ (assuming large sample size in p and subgroup a), we get

$$^{j}\tilde{X}_{\underline{h}} - {}^{j}X_{\underline{h}} \doteq \sum_a \Big(\sum_{h \in \underline{h}} \hat{X}_{ha}\, W_{ha} - \frac{{}^{j}X_{\underline{ha}}\, \hat{P}_{pa}}{P_{pa}} \Big) \frac{P_{pa}}{E(\hat{P}_{pa})}.$$

The bias of $^{j}\tilde{X}_{\underline{h}}$, to this order of approximation, is thus given by

$$B(^{j}\tilde{X}_{\underline{h}}) = E(^{j}\tilde{X}_{\underline{h}} - {}^{j}X_{\underline{h}})$$

$$= \sum_a \Big(\sum_{h \in \underline{h}} X_{ha}\, W_{ha} - {}^{j}X_{\underline{ha}} \Big). \qquad (4.2)$$

It may be noted that P_{pa} is projected population based on the last census. Because of possible undercoverage due to missed dwellings or persons in the sample, the projected population P_{pa} may not be equal to $E(\hat{P}_{pa})$. In deriving the bias, it is assumed that the multiplier $P_{pa}/E(\hat{P}_{pa})$ corrects the estimate $\hat{X}_{ha}$ for coverage bias and thus $E(\hat{X}_{ha}\, P_{pa}/E(\hat{P}_{pa})) = X_{ha}$. It may be noted that

the undercoverage for subgroup a is assumed to be the same in all strata of p.

The variance of ${}^{j}\tilde{X}_{\underline{h}}$ can be obtained as (see Appendix I),

$$V[{}^{j}\tilde{X}_{\underline{h}}] = \sum_{h\in\underline{h}} V[\Sigma_a(\hat{X}_{ha}W_{ha} - R_{\underline{h}a}\hat{P}_{ha}) \frac{P_{pa}}{E(\hat{P}_{pa})}]$$

$$+ \sum_{h\in(p-\underline{h})} V[\Sigma_a(R_{\underline{h}a}\hat{P}_{ha}) \frac{P_{pa}}{E(\hat{P}_{pa})}], \qquad (4.3)$$

where

$$R_{\underline{h}a} = (\sum_{h\in\underline{h}} X_{ha}W_{ha})/P_{pa}.$$

The second term in (4.3) shows the contribution from strata $(p-\underline{h})$, due to ratio adjustment based on the large area p. The ratio adjustment corrects the small area estimates for coverage bias and ensures the additivity of small area estimates for a characteristic total over large area p. The expression for $V(\hat{X}_h)$ can be obtained by substituting $W_{ha} = 1$ in (4.3) and that for $V({}^{j}\hat{X}_{\underline{h}})$ by substituting $W_{ha} = 1$ and replacing $\hat{X}_{ha}$ by ${}^{j}\hat{X}_{ha}$.

5. Evaluation of Bias and Efficiency

The empirical study referred to in this paper was confined to the province of Ontario using the 1971 Census data. The evaluation was done for a stratified clustered sample design with strata and cluster delineations identical to rural strata and primaries of the LFS design. The average number of dwellings per cluster is approximately 2,000. The LFS sample is a multistage area sample with three or four stages in rural strata and two stages in large cities. Details of the LFS design are given in Platek and Singh [9].

In the 1971 Census, labour force data were collected for a systematic sample of dwellings with a sampling ratio of 1/3. By assuming the systematic sample to be a random sample of all persons, the counts of 'unemployed' and 'employed' were weighted within each cluster (i.e. primary of the LFS design) with weights equal to inverse sampling ratio within these clusters.

The province of Ontario has ten economic regions divided into 20 rural strata in the LFS design. Synthetic estimates were evaluated in 17 strata of these regions. Domains composed of complete clusters were formed within these strata with a proportion of the population of the stratum in the domain, W_h, approximately given by 0.25, 0.50 and 0.75.

For simplicity, we consider a special case of the synthetic estimate for a domain defined within a single stratum. Also, the estimate does not use subgroups and the ratio adjustment factor. The estimates and expressions for bias, variance and relative gain in efficiency for the estimates are given below.

Let stratum h consist of N_h clusters of which n_h are drawn with pps with replacement. The unbiased estimate ${}^{j}\hat{X}_h$, of characteristic total ${}^{j}X_h$, in domain j, is given by

$${}^{j}\hat{X}_h = P_h \sum_{i=1}^{n_h} \frac{{}^{j}X_{hi}}{n_h P_{hi}} \qquad (5.1)$$

where P_{hi} is the population in the ith cluster, X_{hi} is the characteristic total in the ith cluster, ${}^{j}X_{hi}$ is X_{hi} if ith cluster is in the domain j and is zero otherwise and $P_h = \sum_{i=1}^{N_h} P_{hi}$. The synthetic estimate is defined as

$${}^{j}\tilde{X}_h = W_h P_h \sum_{i=1}^{n_h} \frac{X_{hi}}{n_h P_{hi}} \qquad (5.2)$$

and the unbiased estimate of characteristic total X_h in stratum h is defined as

$$\hat{X}_h = P_h \sum_{i=1}^{n_h} \frac{X_{hi}}{n_h P_{hi}}$$

The bias and variance of ${}^{j}\tilde{X}_h$ and variance of ${}^{j}\hat{X}_h$ are given by

$$B({}^{j}\tilde{X}_h) = W_hX_h - {}^{j}X_h,$$

$$V({}^{j}\tilde{X}_h) = \frac{W_h^2}{n_h}[\sum_{i=1}^{N_h} \frac{P_hX_{hi}^2}{P_{hi}} - X_h^2], \qquad (5.3)$$

$$V({}^{j}\hat{X}_h) = \frac{1}{n_h}[\sum_{i=1}^{N_h} \frac{P_h{}^{j}X_{hi}^2}{P_{hi}} - {}^{j}X_h^2],$$

where $X_h = \sum_{i=1}^{N_h} X_{hi}$. The relative gain in efficiency of synthetic estimate ${}^{j}\tilde{X}_h$ as compared to ${}^{j}\hat{X}_h$ can be defined as

$$G[{}^{j}\tilde{X}_h] = \frac{V({}^{j}\hat{X}_h) - V({}^{j}\tilde{X}_h) - [B({}^{j}\tilde{X}_h)]^2}{V({}^{j}\tilde{X}_h) + [B({}^{j}\tilde{X}_h)]^2}.$$

Table 1 gives the number of clusters in stratum N_h, the number of clusters in the domain N_{Dh}, and the proportion of population of the stratum in the domain W_h. Table 2 gives % relative bias of ${}^{j}\tilde{X}_h$ and % relative efficiency of ${}^{j}\tilde{X}_h$ for domains within strata for 'unemployed'. The % relative bias of synthetic estimate ${}^{j}\tilde{X}_h$ lies between $\pm$ 8% for 23 of 29 domains considered and it increases as W_h is decreased. The efficiency decreases as domain size W_h is increased and the number of sampled clusters n_h is increased from 2 to 4. The efficiency gains for small W_h are quite high due to high values of $V({}^{j}\tilde{X}_h)$ in the clustered sample design.

The population and characteristic counts in the domains and strata used in the evaluation are as of census time. In intercensal years, the bias of synthetic estimates can increase due to uneven population growth. The effect of this uneven growth is examined below. For simplicity, we consider again the case of a single stratum and no subgroups. The relative bias is given by

$$\frac{B({}^{j}\tilde{X}_h)}{{}^{j}X_h} = \frac{X_h W_h}{{}^{j}X_h} - 1, \qquad (5.4)$$

where $W_h = {}^{j}P_h'/P_h'$, ${}^{j}P_h'$ and P_h' being population counts in the domain and stratum, as of the last census. Let

$$Q_h = W_h \frac{P_h}{{}^{j}P_h} - 1$$

$$= \frac{{}^{j}P_h'}{{}^{j}P_h} \cdot \frac{P_h}{P_h'} - 1, \qquad (5.5)$$

where ${}^{j}P_h$ and P_h are the current population counts in the domain and stratum respectively. Let $\frac{P_h}{P_h'} = 1+g_h$ and $\frac{{}^{j}P_h}{{}^{j}P_h'} = 1+{}^{j}g_h$, where ${}^{j}g_h$ and g_h denote growth rates in the domain and stratum respectively. Thus, $Q_h \doteq g_h - {}^{j}g_h$ if $|{}^{j}g_h| < 1$, i.e. non-zero value of Q_h represents uneven population growth within stratum h. Also,

$$W_h = \frac{{}^{j}P_h}{P_h}(1 + Q_h) \text{ and}$$

$$B({}^{j}\tilde{X}_h) = \left(W_h \frac{{}^{j}P_h}{P_h} - {}^{j}X_h\right) + Q_h \frac{{}^{j}P_h}{P_h} X_h .$$

The relative bias is given by

$$\frac{B({}^{j}\tilde{X}_h)}{{}^{j}X_h} = \frac{\left(X_h \frac{{}^{j}P_h}{P_h} - {}^{j}X_h\right)}{{}^{j}X_h} + Q_h \frac{X_h}{P_h} \cdot \frac{{}^{j}P_h}{{}^{j}X_h} \qquad (5.6)$$

The first term of (5.6) represents relative bias as of census time and the second term the contribution of uneven population growth.

In order to evaluate average bias and efficiency for domains of a particular size W_h, we consider a super-population model expressing the relationship between characteristic and population counts in clusters. Such models have been extensively used in the sample survey literature for evaluation of efficiencies of alternative sample designs and estimates.

We assume that a stratum consists of two domains j and its complement with the following model. For the ith cluster of stratum h let

$$X_{hi} = \beta_h P_{hi} + e_{hi} \qquad i \in \text{domain } j$$
$$X_{hi} = \delta_h \beta_h P_{hi} + e_{hi} \qquad i \notin \text{domain } j, \delta_h \neq 1, \qquad (5.7)$$

where β_h and $\delta_h\beta_h$ are regression coefficients and

$$\varepsilon(e_{hi}|P_{hi}) = 0, \varepsilon(e_{hi}^2|P_{hi}) = \sigma_h^2 . P_{hi}^t, \; t>0, \sigma_h^2 \geq 0,$$

$$\varepsilon(e_{hi} . e_{hi}'|P_{hi}, P_{hi}') = 0, i \neq i', i, i' = 1,2,\ldots,N_h,$$

where ε denotes expectation over domains. The model is appropriate for characteristics which are defined as counts of persons with certain attribute. The regression coefficient can be considered as a proportion of the population with the attribute in a domain in stratum h. The heterogeneity between clusters is shown by a non-zero value of σ_h^2, with dependence of variance on cluster size being appropriate for categorical data. It is known that for socio-economic characteristics t lies between 1 and 2. The parameter δ_h represents heterogeneity due to different regression coefficients, i.e. proportions between domains within a stratum, the form of regression coefficient in the complement of the domain being taken for simplicity of later expressions in this section. The implicit assumption in using separate ratio estimates over subgroups (combined over strata) as in (3.1), (3.2) and (3.3) is that heterogeneity between subgroups is a more important source of variation of characteristic counts than heterogeneity between strata within subgroups. The above model is, therefore, also appropriate for individual subgroups in a group of strata h by changing h to $\underline{h}$.

Under model (5.7), it can be proved by first order Taylor series approximation that

$$\varepsilon\left[\frac{B({}^{j}\tilde{X}_h)}{{}^{j}X_h}\right] \doteq (1-\delta_h)(W_h - 1) + Q_h\delta_h . \qquad (5.8)$$

Thus, if $\delta_h = 1$ and $Q_h = 0$, i.e. growth in intercensal period is even, the estimate is model-unbiased. The expression (5.8) shows the effect of heterogeneity, population growth and relative size of the domain on the relative bias of the synthetic estimate. The super-population variance of relative bias, assuming $\delta_h = 1$ and $t = 1$, is given by

$$\upsilon\left[\frac{B({}^{j}\tilde{X}_h)}{{}^{j}X_h}\right] \doteq \frac{\sigma_h^2}{{}^{j}P_h^2 \beta_h^2}\left[W_h . {}^{j}P_h . Q_h + (1-W_h){}^{j}P_h\right] \qquad (5.9)$$

When $Q_h = 0$, the variance is given by

$$\upsilon\left[\frac{B({}^{j}\tilde{X}_h)}{{}^{j}X_h}\right] \doteq \frac{\sigma_h^2}{{}^{j}P_h . \beta_h^2}\left[1 - W_h\right]. \qquad (5.10)$$

Since $\sigma_h^2/{}^{j}P_h . \beta_h^2$ is approximately equal to the square of the coefficient of variation of the characteristic total in the domains of size W_h, the relative bias is expected to be more stable for characteristics with greater homogeneity within post-strata. Also, the relative bias is expected to be stable for domains of large sizes. These results can be used in obtaining rough confidence intervals on the relative bias for intercensal years by assuming that errors are normally distributed.

The relative gain in efficiency under the super-population model is given by

$$\frac{\varepsilon[V[{}^{j}\hat{X}_h] - V[{}^{j}\tilde{X}_h] - (B[{}^{j}\tilde{X}_h])^2]}{\varepsilon[V[{}^{j}\tilde{X}_h] + (B({}^{j}\tilde{X}_h))^2]} = A/B \qquad (5.11)$$

where

$$A = \frac{{}^{j}P_h[P_h - {}^{j}P_h]}{n_h} - {}^{j}P_h^2[(1-\delta_h+\delta_h\frac{P_h}{{}^{j}P_h})W_h - 1]^2$$

$$- \frac{W_h^2 P_h}{n_h}[\sum_{i=1}^{N_h} P_{hi}[{}^{j}A_{hi}+\delta_h(1-{}^{j}A_{hi})]^2 - P_h$$

$$[\frac{{}^{j}P_h}{P_h} + \delta_h(1 - \frac{{}^{j}P_h}{P_h})]^2]$$

$$+ \frac{\sigma_h^2}{n_h\beta_h^2}[\sum_{i=1}^{N_h} P_{hi}^{(t-1)}[(P_h-P_{hi})({}^{j}A_{hi}-W_h^2) - n_h P_{hi}$$

$$({}^{j}A_{hi} - W_h^2)],$$

$$B = {}^{j}P_h^2[(1-\delta_h+\delta_h\frac{P_h}{{}^{j}P_h})W_h - 1]^2 + \frac{W_h^2 P_h}{n_h}[\sum_{i=1}^{N_h} P_{hi}$$

$$[{}^{j}A_{hi}+\delta_h(1-{}^{j}A_{hi})]^2 - P_h[\frac{{}^{j}P_h}{P_h}+\delta_h(1-\frac{{}^{j}P_h}{P_h})]^2]$$

$$+ \frac{\sigma_h^2}{n_h\beta_h^2}[\sum_{i=1}^{N_h} P_{hi}^{(t-1)}[(P_h-P_{hi})W_h^2 + n_h P_{hi}({}^{j}A_{hi} -$$

$$W_h)^2]] \qquad (5.12)$$

${}^{j}A_{hi}$ = 1 if the ith sampled cluster is in the domain and

${}^{j}A_{hi}$ = 0 otherwise. When $\delta_h = 1$,

$$A = \frac{1}{n_h}[{}^{j}P_h(P_h - {}^{j}P_h)] - {}^{j}P_h^2[\frac{P_h}{{}^{j}P_h}W_h - 1]^2$$

$$+ \frac{\sigma_h^2}{n_h\beta_h^2}[\sum_{i=1}^{N_h} P_{hi}^{t-1}[(P_h-P_{hi}).$$

$$({}^{j}A_{hi}-W_h^2) - n_h P_{hi}({}^{j}A_{hi}-W_h)^2]]$$

$$B = {}^{j}P_h^2[\frac{P_h}{{}^{j}P_h}W_h - 1]^2 + \frac{\sigma_h^2}{n_h\beta_h^2}[\sum_{i=1}^{N_h} P_{hi}^{t-1}[(P_h-P_{hi})$$

$$W_h^2 + n_h P_{hi}({}^{j}A_{hi} - W_h^2)]]. \qquad (5.13)$$

If $\sigma_h^2 = 0$, the gain depends on the number of clusters in the sample, relative size of the domain, inaccuracy of weights shown by $(\frac{P_h}{{}^{j}P_h}W_h - 1)$ and heterogeneity in the stratum $(\delta_h \neq 1)$. When $\delta_h = 1$ and the domain is equal to the stratum (i.e. $W_h = 1$) the efficiency gain is zero. The additional terms in (5.12), as compared to (5.13), represent loss in efficiency due to heterogeneity in the stratum as represented by $\delta_h \neq 1$.

The parameters in (5.12) and (5.13) introduced by the model are β_h, σ_h^2 and δ_h. The best linear unbiased estimates by generalized least squares under model (5.7), (assuming $\delta_h = 1$) are given by

$$\hat{\beta}_h = \frac{\sum_{i=1}^{N_h} X_{hi} P_{hi}^{(1-t)}}{\sum_{i=1}^{N_h} P_{hi}^{(2-t)}}$$

$$\hat{\sigma}_h^2 = \frac{1}{(N_h-1)}\sum_{i=1}^{N_h} P_{hi}^{-t}(\hat{e}_{hi} - \bar{e}_h)^2,$$

where

$$\hat{e}_{hi} = X_{hi} - \hat{\beta}_h P_{hi}, \quad i=1,2,\ldots,N_h,$$

$$\bar{e}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} \hat{e}_{hi}.$$

The parameter δ_h can be estimated as a ratio of the regression coefficient under model (5.7) in the complement of the domain and that in the domain. However, in the empirical evaluation selected values of δ_h are used. The expressions for efficiency gains based on an extension of (5.7) to subgroups in a group of strata are given in Ghangurde and Singh [4].

The expressions (5.12) and (5.13) were computed for characteristic 'unemployed' using census data on X_{hi} and P_{hi}, $i=1,2,\ldots,N_h$. Table 3 gives % efficiency gains under the model with $Q_h=0.0$, $\delta_h=1.0$, 1.1, 1.2 and $t=1.0$ for the same domains in economic regions 53 to 57 as in Table 2. It may be noted that $Q_h=0.0$ (i.e no population growth) at census time. The efficiency gains shown in Table 3 for domains in these regions can be compared with efficiency gains based on census data as shown in Table 2. The efficiency gains under the model decrease as δ_h is increased. The relative bias value in stratum 1 of economic region 53 is high and that in stratum 1 of economic 57 is low in comparison to those in other strata. This could explain high discrepancies in efficiency gains between Tables 2 and 3 for these two strata. For other regions the two efficiency gains compare well, thus providing a rough validation of the model used in the evaluation. More extensive empirical results on efficiency gains, based on the above model and its extension for subgroups and assuming

other values of Q_h and t are given in Ghangurde and Singh [4].

The values of relative gains in efficiency obtained are very high due to the clustered sample design. For practical use of synthetic estimation in such designs, we need a criterion for relating the accuracy of synthetic estimate for a set of domains to that of design-based estimate for a group of strata. In section 6, we propose a criterion for this purpose.

6. A Measure of Relative Accuracy

We consider a ratio of relative mean square error of the synthetic estimate to relative variance of the unbiased estimate for the group of strata, which include the set of domains. The ratio is given by

$$R[{}^{j}\tilde{X}_{\underline{h}}] = \frac{V({}^{j}\tilde{X}_{\underline{h}}) + [B({}^{j}\tilde{X}_{\underline{h}})]^2}{V(\hat{X}_{\underline{h}})} \cdot \frac{X_{\underline{h}}^2}{{}_{j}X_{\underline{h}}^2} \qquad (6.1)$$

and can be considered as a measure of relative accuracy of the synthetic estimate for a set of domains as compared to the unbiased estimate for $\underline{h}$. For interpretation of this ratio, we assume that domains within strata are of the same relative size for all $h \epsilon \underline{h}$, i.e. $W_{ha} = W_{\underline{h}a} \doteq W_{\underline{h}}$. It follows that

$$X_{\underline{h}}.W_{\underline{h}} = B({}^{j}\tilde{X}_{\underline{h}}) + {}_{j}X_{\underline{h}}, \text{ and } V({}^{j}\tilde{X}_{\underline{h}}) = W_{\underline{h}}^2\, V(\hat{X}_{\underline{h}}).$$

Hence

$$R[{}^{j}\tilde{X}_{\underline{h}}] \doteq [1 + \frac{[B({}^{j}\tilde{X}_{\underline{h}})]^2}{V({}^{j}\tilde{X}_{\underline{h}})}][1 + \frac{B({}^{j}\tilde{X}_{\underline{h}})}{{}_{j}X_{\underline{h}}}]^2. \qquad (6.2)$$

The ratio is a simple function of bias ratio and relative bias. If bias ratio and relative bias are small, synthetic estimate (for a set of domains) will have approximately the same accuracy as that of the unbiased estimate $\hat{X}_{\underline{h}}$ for $\underline{h}$. The above ratio may be called synthetic (or small area) estimation effect, since it is a factor by which accuracy of small area data is decreased (i.e. relative mean square error is increased) due to non-conformity of small areas to design strata.

The ratio was evaluated for the same domains within strata of economic regions 53 to 57 as those considered for evaluation of bias and efficiency. The ratio $R({}^{j}\tilde{X}_{\underline{h}})$ is obtained by substituting h for $\underline{h}$ in (6.1). Tables 4 and 5 give values of $R({}^{j}\tilde{X}_{\underline{h}})$ for 'unemployed' and 'employed' for domains of various sizes W_h and sampled clusters n_h. The value of the ratio increases as n_h is increased and W_h is decreased. Due to high values of relative bias and bias ratio for 'unemployed' as comapred to those for 'employed', for the same values of W_h and n_h, the ratios are, in general, greater for 'unemployed' than for 'employed'. It may be noted that the ratio can be expressed in terms of coefficient of variation (CV) and relative bias.

In practice, the relative bias can be estimated using census data and $CV(\hat{X}_{\underline{h}})$ and $CV({}^{j}\tilde{X}_{\underline{h}})$ using survey data. For census years, the ratio can thus be evaluated for various characteristics and domain sizes. In case census data is not available for some characteristics, the results on relative bias and its variance under models, given in section 5, can be used for evaluation of bias. The values of the ratio can be used to appraise accuracy of synthetic estimates for domains in comparison to that of estimates for complete strata. For characteristics and domain sizes with the value of the ratio close to one, synthetic estimates can be used with similar interpretation as that for the design-based estimates for complete strata.

7. Concluding Remarks

The main feature of this paper has been the derivation of bias and variance of synthetic estimate with ratio adjustment based on projected population in a large area and evaluation of bias and efficiency using census data. The framework of a super-population model (considered in this paper) provides analytical results on the bias and efficiency and shows the effect of population growth and heterogeneity of the population. The mean square error under the super-population model, as a measure of accuracy of the synthetic estimate, corresponds to the concept of average mean square introduced by Gonzales and Waksberg [7]. The basic sample design assumed in the evaluation of efficiency gains is cluster sampling with pps with replacement. However, the results on bias and variance given in section 4 are valid for any sample design. The values of efficiency gains obtained are rather high for small domains due to the clustered sample design.

Acknowledgements

The authors wish to express their appreciation to Mr. R. Platek, Director, Household Surveys Development Division for helpful discussions.

Appendix I

We have from (3.3)

$${}^{j}\tilde{X}_{\underline{h}} = \sum_{a} \Big(\sum_{h \epsilon \underline{h}} \hat{X}_{ha} W_{ha} \Big) . \frac{P_{pa}}{\hat{P}_{pa}}$$

Let $R_{\underline{h}a} = \big(\sum_{h \epsilon \underline{h}} X_{ha} W_{ha} \big) / P_{pa}$. Then by first order Taylor series approximation to $\hat{X}_{ha}/\hat{P}_{pa}$

$${}^{j}\tilde{X}_{\underline{h}} - \sum_{a} \sum_{h \epsilon \underline{h}} X_{ha} W_{ha} = \Big[\frac{\sum_{h \epsilon \underline{h}} \hat{X}_{ha} W_{ha}}{\hat{P}_{pa}} - \frac{\sum_{h \epsilon \underline{h}} X_{ha} W_{ha}}{P_{pa}} \Big] P_{pa}$$

$$= \sum_{a} \Big[\sum_{h \epsilon \underline{h}} \hat{X}_{ha} W_{ha} - \frac{\big(\sum_{h \epsilon \underline{h}} X_{ha} W_{ha} \big) \hat{P}_{pa}}{P_{pa}} \Big] \frac{P_{pa}}{E(\hat{P}_{pa})}$$

Assuming $E[\hat{X}_{ha} \frac{P_{pa}}{E(\hat{P}_{pa})}] = X_{ha}$, i.e. that the multiplier $P_{pa}/E(\hat{P}_{pa})$ corrects the estimate $\hat{X}_{ha}$ for

undercoverage errors in subgroup a in h, we have

$$E(\hat{{}^{j}X}_{\underline{h}}) = \sum_{a} \sum_{h\epsilon\underline{h}} X_{ha}W_{ha}.$$

By changing the order of summation, which avoids derivation of variances and covariances for post-strata within strata (Woodruff [11])

$$\hat{{}^{j}X}_{\underline{h}} - \sum_{a} \sum_{h\epsilon\underline{h}} X_{ha}W_{ha} = \sum_{h\epsilon\underline{h}} \left[\sum_{a}(\hat{X}_{ha}W_{ha} - R_{\underline{h}a}\hat{P}_{pa}) \frac{P_{pa}}{E(\hat{P}_{pa})}\right]$$

$$= \sum_{h\epsilon\underline{h}} \left[\sum_{a}(\hat{X}_{ha}W_{ha} - R_{\underline{h}a}\hat{P}_{ha}) \frac{P_{pa}}{E(\hat{P}_{pa})}\right]$$

$$- \sum_{h\epsilon(p-\underline{h})} \left[(R_{\underline{h}a}\hat{P}_{ha}) \frac{P_{pa}}{E(\hat{P}_{pa})}\right]$$

Since sampling is done independently within each stratum

$$V[\hat{{}^{j}X}_{\underline{h}}] = \sum_{h\epsilon\underline{h}} V\left[\sum_{a}(\hat{X}_{ha}W_{ha} - R_{\underline{h}a}\hat{P}_{ha}) \frac{P_{pa}}{E(\hat{P}_{pa})}\right]$$

$$+ \sum_{h\epsilon(p-\underline{h})} V\left[\sum_{a}(R_{\underline{h}a}\hat{P}_{ha}) \frac{P_{pa}}{E(\hat{P}_{pa})}\right]$$

which is (4.3).

Table 1

Domain Sizes and Weights

Economic Region	Stratum	N_h	N_{Dh}	W_h
52	1	13	3	0.22
			6	0.46
			10	0.77
	2	11	3	0.25
			6	0.53
			9	0.83
	3	15	4	0.25
			7	0.48
			12	0.82
58	1	18	4	0.27
			9	0.48
			14	0.76
	2	17	4	0.21
			9	0.50
			14	0.72
	3	13	3	0.27
			6	0.46
			10	0.63
53	1	13	6	0.46
54	1	15	8	0.55
55	1	17	8	0.47
	2	14	7	0.50
56	1	14	7	0.50
	2	13	6	0.39
57	1	14	7	0.49
	2	17	8	0.69
50	1	14	6	0.44
59	1	14	6	0.45
	2	19	7	0.39

Table 2

% Relative Bias and Efficiency Gain (Unemployed)

Economic Region	Stratum	W_h	% Rel Bias	% Efficiency Gain n_h=2	n_h=3	n_h=4
52	1	0.22	10.98	5,689	4,724	4,035
		0.46	-3.68	3,540	3,670	3,513
		0.77	1.39	1,300	1,023	1,016
	2	0.25	2.00	10,761	10,602	10,449
		0.53	1.34	3,670	3,266	3,244
		0.83	.08	792	790	783
	3	0.25	5.33	4,584	4,413	4,254
		0.48	5.48	1,463	1,384	1,312
		0.82	3.94	303	294	286
58	1	0.27	2.69	5,172	5,119	5,067
		0.48	-3.70	1,808	1,768	1,730
		0.76	-2.15	502	497	493
	2	0.21	2.82	1,315	993	791
		0.50	-8.15	1,509	1,370	1,253
		0.72	-6.85	589	545	506
	3	0.27	1.24	3,680	3,016	2,550
		0.46	8.43	2,149	1,878	1,668
		0.63	2.23	1,655	1,631	1,607
53	1	0.46	27.32	316	237	182
54	1	0.55	2.43	2,824	2,767	2,712
55	1	0.47	8.02	1,352	1,251	1,162
	2	0.50	-7.06	2,128	1,892	1,702
56	1	0.50	-6.40	1,553	1,463	1,382
	2	0.39	-4.09	5,080	4,819	4,583
57	1	0.49	1.59	7,853	6,690	5,824
	2	0.69	-3.16	2,808	2,639	2,488
50	1	0.44	9.11	1,430	1,299	1,189
	1	0.45	4.25	2,669	2,569	2,475
	2	0.39	-7.68	1,211	1,161	1,115

Table 3

% Efficiency Gain Under Model (Unemployed)

Q_h=0.0, t=1.00

ER	Stratum	n_h	δ_h=1.00	δ_h=1.10	δ_h=1.20
53	1	2	1030	595	789
		3	946	864	681
		4	873	786	597
54	1	2	2759	2253	1483
		3	2116	2082	1212
		4	2486	1866	1044
55	1	2	1670	1488	1115
		3	1567	1356	956
		4	1476	1244	834
	2	2	1994	1729	1226
		3	1865	1563	1039
		4	1750	1425	899
56	1	2	1309	1190	930
		3	1219	1083	804
		4	1141	993	705
	2	2	4202	3328	2031
		3	3798	2857	1615
		4	3464	2499	1335
57	1	2	1322	1202	939
		3	1231	1093	811
		4	1150	1001	711
	2	2	2784	2200	1333
		3	2709	2052	1166
		4	2638	1922	1034

Table 4

$R(\hat{{}^{j}X}_h)$ (Unemployed)

ER	Stratum	W_h	$R(\hat{{}^{j}X}_h)$ n_h=2	n_h=3	n_h=4
52	1	0.22	2.0529	2.4636	2.8739
		0.46	1.0201	1.0662	1.1124
		0.77	1.0413	1.0478	1.0544
	2	0.25	1.0734	1.0892	1.1050
		0.53	1.0405	1.0471	1.0539
		0.83	1.0203	1.0226	1.0249
	3	0.25	2.5433	2.6399	2.7365
		0.48	1.2687	1.3363	1.4038
		0.82	1.1304	1.1554	1.1803
58	1	0.21	1.0763	1.0872	1.0982
		0.48	0.9689	0.9895	1.0103
		0.77	0.9715	0.9786	0.9856
	2	0.21	3.9996	5.1779	6.3557
		0.50	1.0400	1.1382	1.2364
		0.72	1.0064	1.0758	1.1452
	3	0.27	2.2042	2.6752	3.1456
		0.46	1.6109	1.8284	2.0459
		0.63	1.0756	1.0909	1.1061
53	1	0.46	3.0872	3.8204	4.5534
54	1	0.55	1.0925	1.1141	1.1358
55	1	0.47	1.3734	1.4768	1.5802
	2	0.50	1.1162	1.2479	1.3797
56	1	0.50	0.9907	1.0480	1.1054
	2	0.39	1.0292	1.0837	1.1384
57	1	0.49	2.0418	2.3915	2.7412

Table 5

$R(\hat{{}^{j}X}_h)$ (Employed)

ER	Stratum	W_h	$R(\hat{{}^{j}X}_h)$ n_h=2	n_h=3	n_h=4
52	1	0.46	1.0163	1.0217	1.0270
	2	0.53	1.0357	1.0451	1.0544
	3	0.48	1.0553	1.1031	1.1509
58	1	0.21	1.2630	1.3804	1.4979
		0.48	1.0039	1.0088	1.0136
	2	0.21	1.0208	1.0408	1.0606
		0.50	1.1009	1.1384	1.1757
	3	0.27	1.2415	1.3844	1.5275
		0.46	1.0485	1.0838	1.1191
53	1	0.46	1.3505	1.5778	1.8051
54	1	0.55	1.3729	1.5711	1.7693
55	1	0.47	1.0074	1.0096	1.0117
	2	0.50	1.3581	1.5731	1.7881
56	1	0.50	1.5016	1.7331	1.9646
	2	0.39	1.0722	1.1203	1.1683
57	1	0.49	1.0899	1.1243	1.1587
	2	0.69	1.3061	1.4456	1.5850
50	1	0.44	1.3134	1.3611	1.4090
59	2	0.46	0.9845	1.0376	1.0907
	3	0.39	1.3007	1.3585	1.4164

References

[1] Chambers, R.L. and Feeney, G.A. (1977), "Log Linear Models for Small Area Estimation", Paper presented at the Conference of the Biometric Society, Australasian Region.

[2] Eriksen, E.P. (1974), "A Regression Method for Estimating Population Changes for Local Areas", Journal of the American Statistical Association, pp. 867-875.

[3] Freeman, D.H. and Koch, G.G. (1976), "An Asymptotic Covariance Structure for Testing Hypotheses on Raked Contingency Tables from Complex Sample Surveys", Presented at the Social Statistics Section of the American Statistical Association.

[4] Ghangurde, P.D. and Singh, M.P. (1977), "Evaluation of Efficiency of Synthetic Estimation in the LFS", Technical Report, Statistics Canada.

[5] Gonzales, M.E. (1973), "Use and Evaluation of Synthetic Estimates", Proceedings of the Social Statistics Section of the American Statistical Association, pp. 437-443.

[6] Gonzales, M.E. and Hoza, C. (1978),"Small Area Estimation With Application to Unemployment and Housing Estimates", Journal of the American Statistical Association, pp. 7-15.

[7] Gonzales, M.E. and Waksberg, J.L. (1975), "Evaluation of the Error of Synthetic Estimates", Unpublished paper presented at the first meeting of the International Association of Survey Statisticians, Vienna.

[8] Holt, D., Smith, T.M.F. and Tomberlin, T.J. (1977), "Synthetic Estimation for Small Subgroups of a Population", Unpublished Report, University of Southampton.

[9] Kish, L. and Frankel, M.R. (1974), "Inference from Complex Samples (with discussion)", Journal of the Royal Statistical Society, B, 36, 1-37.

[10] Levy, P.S. and French, D.K. (1977), "Synthetic Estimation of State Health Characteristics Based on the Health Interview Survey", Series 2, Report No. 75, U.S. Department of Health, Education, and Welfare.

[11] Platek, R. and Singh, M.P. (1976), "Methodology of the Canadian Labour Force Survey", Technical Report, Statistics Canada.

[12] Purcell, N.J. and Linacare, S. (1976), "Techniques for the Estimation of Small Area Characteristics, Paper presented at the third Australian Statistical Conference, Melbourne.

[13] Royall, R.M. (1977), "Statistical Theory of Small Area Estimates - Use of Prediction Models", Unpublished Report, U.S. National Center of Health Statistics.

[14] 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" Paper presented at the Annual Meeting of the American Statistical Association.

[15] Woodruff, R.S. (1976), "A Simple Method of Approximating Variance of A Complicated Estimate", Journal of the American Statistical Association, pp. 411-414.

[16] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proceedings of Social Statistics Section, American Statistical Association.