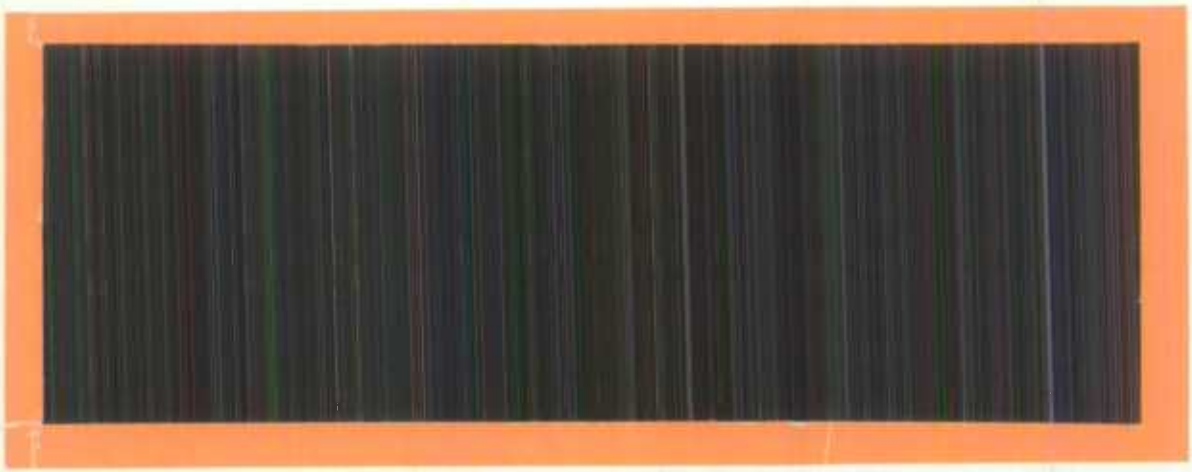
Statistics Canada Statistique Canada



Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

Canada

SMALL AREA ESTIMATES FROM SAMPLE SURVEYS

G.H. Choudhry, and Y. Bélanger

Census and Household Survey Methods Division

Methodology Branch

Number: CHSM 85-075E

SMALL AREA ESTIMATES FROM SAMPLE SURVEYS

G.H. Choudhry, and Y. Bélanger

A Monte Carlo study was carried out to evaluate the performance of some alternate Small Area Estimation techniques to obtain Labour Force estimates from large scale household surveys. The Canadian Labour Force survey which utilizes a clustered multi-stage sample design has been taken as an example of a large scale household survey and the survey design was simulated using data from 1976 and 1981 censuses. The census divisions (CD's) which cut across the boundaries of design strata are taken as small areas (domains) and 1,000 Monte Carlo samples were selected to obtain LF estimates. The alternate small area estimators which include post-stratified domain, synthetic, composite and SPREE were evaluated for their Biases and Mean Square Errors. Keyfitz's (1957) variance estimator for estimating the variance of a ratio has been considered for small area estimators using population adjustment at the small area level.

Keywords: Monte Carlo Study; Small Area Estimation; Multi-Stage Survey Design; Mean Square Error

1. INTRODUCTION

The demand for more and better quality data for small areas (or domains) has been constantly increasing during the recent years. The choice of an estimation method, among other factors, depends on the availability of data sources and the particular type of applications.

In our study, the small areas are the Census Divisions (CD's) which are unplanned areal domains and were not distinguished at the time of sample design and thus cut-across the boundaries of the design strata. The sizes of these domains are such that the reliability of regular estimates would have been satisfactory had these domains been designed with separate fixed sample sizes from individual domains. The study was undertaken to evaluate some of the alternate small area estimation

procedures to produce CD level estimates from the Canadian Labour Force Survey (LFS), using data from the population census in an auxiliary fashion. Since the post-censal population estimates of CD's are also available (Verma and Basavarajappa; 1985), we have also made use of the current CD population during estimation. This study is similar to the one reported by Drew, Singh, and Choudhry (1982), but differs from the above in the following aspects: (a) the number of Monte Carlo replicates was increased to 1,000 from 100 in the earlier study, (b) the structure preserving estimation (SPREE) method of Purcell and Kish (1979) was also evaluated in this study, and (c) Keyfitz's (1957) variance estimator for estimating the variance of a ratio was applied to estimate the variance of small area estimators using population adjustment at the CD level. In addition to SPREE, we have evaluated post-stratified domain, synthetic and composite estimators which are linear combinations of post-stratified and synthetic estimators. The composite estimators we have considered are of the particular type where the weight given to the post-stratified domain estimator depends on the sample size in the domain. Basically the weight given to the post-stratified domain component increases as the sample size in the domain increases until the weight becomes one for some pre-specified sample size in the domain. We have considered two such estimators and these are: (i) sample size dependent estimator proposed by Drew, Singh, and Choudhry (1982) where the weight increases linearly as a function of the ratio of sample size in the domain to the expected sample size until it becomes one, and (ii) modified regression (MRE) estimator of Hidiroglou and Sarndal (1984) where the weight increases in a quadratic fashion as a function of the ratio of sample size in the domain to the expected sample size until it becomes one. Both of these estimators become synthetic estimator when there is no sample in the domain and these become post-stratified domain estimator when the sample size in the domain is equal to or greater than the expected sample size.

For all the estimators considered in this study, the average percent absolute relative biases and the average root mean square errors were obtained in a Monte Carlo study in which the LFS design was simulated using data from 1976 and 1981 censuses. Keyfitz's (1957) variance estimator for the variance of a ratio was applied and evaluated for estimating the variances of those small area estimators which use population adjustment at the small area level.

2. Small Area Estimators

We consider a finite population U consisting of N units, (e.g., households or dwellings in household surveys), divided into L (areal) design strata labelled 1, 2, ..., h, ..., L. The population U is also divided into A non-overlapping small areas (or domains) 1, 2, ..., a, ..., A for which estimates are required. If we denote by ${}_aU$ the set of units belonging to the small area 'a', then the parameter to be estimated is the total of y-variable for all units in ${}_aU$ which we denote by ${}_aY$. Let ${}_aU_h$ be the set of units belonging to the intersection of small area 'a' and design stratum 'h', and y_j be the y-value (e.g., number of persons employed, unemployed, etc.) associated with jth unit, j = 1, 2, ..., N, then

$$
\begin{aligned}
{}_aY &= \sum_h \sum_{j \in {}_aU_h} y_j \\
&= \sum_{j \in {}_aU} y_j \ .
\end{aligned}
\tag{2.1}
$$

The particular design under consideration follows a multi-stage clustered area sample design which is self-weighting within each stratum. A sample s_h of size n_h given by

$$
n_h = N_h / W_h \tag{2.2}
$$

is selected from stratum h independently where N_h is number of units in the stratum and $1/W_h$ is the sampling rate in the stratum. All persons belonging to the households selected in the sample are interviewed and their Labour Force status is determined. Let (h, j, k) denote the kth person in the jth household in stratum h, then define

$$
\tau_{hjk} = 1 \quad \text{if (h, j, k) possesses the characteristic}
$$

$$
= 0 \quad \text{otherwise}
$$

$$_a\delta_{hj} = 1 \quad \text{if } (h, j) \epsilon\, _aU$$

$$= 0 \quad \text{otherwise}$$

Let W_{hjk} be the final weight for the person denoted by (h, j, k) whenever $(h, j) \epsilon s_h$. The final weight W_{hjk} is different from the design weight W_h because it incorporates efficient estimation procedures such as post-stratification by age/sex or raking ratio estimation.

The direct (or also referred to as expansion or simple domain) estimator of the total $_aY$ is given by

$$_a\hat{Y}_{EXP} = \sum_h \sum_{j \epsilon s_h} {}_a\delta_{hj} \sum_k (\tau_{hjk} W_{hjk}). \tag{2.3}$$

Similarly an estimate of the total population $_aP$ is given by

$$_a\hat{P}_{EXP} = \sum_h \sum_{j \epsilon s_h} {}_a\delta_{hj} \sum_k W_{hjk}. \tag{2.4}$$

The direct estimator does not utilize any auxiliary information, all it requires is the identification of those sampled units which belong to the small area of interest. The sample size in the small area is a random variable and due to the clustered nature of the design, the variation in the sample size falling in the domain could be large, resulting in high variance for this estimator. The other estimators considered for evaluation purposes rely on information on an auxiliary x-variable which is often taken as the count of persons or count of persons by population sub-groups (usually defined on the basis of age/sex) from a recent census. These estimators in our study are: (i) post-stratified domain, (ii) synthetic, (iii) sample size dependant, and (iv) structure preserving estimates (SPREE). The estimators (ii), (iii), and (iv) also make use of sample and auxiliary data external to the domain to a different degree. As discussed by Drew, Singh, and Choudhry (1982), the adjustment based on auxiliary information can be made either separately for each stratum intersecting the domain, or by applying an overall adjustment for all strata intersecting the domain. Accordingly, the estimators were further classified as <u>separate</u> or <u>combined</u> depending on the level at which the adjustment was applied. In this study we have only considered the <u>combined</u> version of these estimators due

to their stability.

(i) Post-Stratified Domain Estimator

Let g = 1, 2, ..., G denote the population sub-groups by age/sex, and x-variable denote the population count from recent census, then define

$_aX_{hg}$ = total of x-variable for age/sex group g in small area 'a' intersected by stratum h,

and

$$_aR_g = \sum_{h\in\tilde{h}} {}_aX_{hg} / \sum_{h\in\tilde{h}} {}_a\hat{X}_{hg} , \qquad (2.5)$$

where $_a\hat{X}_{hg}$ is the direct estimate of $_aX_{hg}$ constructed from the selected sample of PSU's and $\tilde{h} = [h \mid {}_aU_h \neq \emptyset]$. If a stratum is completely contained in the small area, then $_a\hat{X}_{hg}$ was replaced by $_aX_{hg}$. Then the post-stratified domain estimator is given by

$$_a\hat{Y}_{PSD} = \sum_g \left(\sum_{h\in\tilde{h}} {}_a\hat{Y}_{hg}\right) {}_aR_g , \qquad (2.6)$$

where $_a\hat{Y}_{hg}$ is the direct estimate for population sub-group g in the small area intersected by stratum h. The post-stratified domain estimator is unbiased except for the effect of ratio estimation bias which is usually negligible. The estimator was defined to be zero when there was no sample in the domain. This estimator is not very reliable for small sample sizes.

(ii) Synthetic Estimator

We consider the synthetic estimator defined as

$$_a\hat{Y}_{SYN} = \sum_g \left(\sum_{h\in\tilde{h}} \hat{Y}_{hg}\right) \frac{\sum_{h\in\tilde{h}} {}_aX_{hg}}{\sum_{h\in\tilde{h}} X_{hg}} , \qquad (2.7)$$

where

X_{hg} = total of x-variable for age/sex group g in stratum h,

and

${}_aX_{hg}$ was defined before

The above estimator has been investigated in Australia by Purcell and Linacre (1976), and in Canada by Ghangurde and Singh (1977, 1978). A different form of synthetic estimator has been used by Gonzalez (1973); Gonzalez and Waksberg (1973); Schaible, Brock, and Schnack (1977); Gonzalez and Hoza (1978), and Laake (1978) for estimates of total number of persons and unemployment. The synthetic estimator is based on the assumption of homogeneity of the characteristics for the population subgroups across small areas and the estimator could be highly biased if the assumption of homogeneity does not hold.

(iii) Sample Size Dependent Estimators

The sample size dependent estimator is a linear combination of the post-stratified domain and synthetic estimators, where the weights depend on the sample size in the domain. It is constructed using the result that the performance of the post-stratified domain estimator depends on the sample size in the domain. If the sample size in the domain is more than some predetermined cut-off value then the estimator is essentially the post-stratified domain estimator, otherwise the weight on the post-stratified domain component decreases as the sample size in the domain decreases and it becomes the synthetic estimator when there is no sample in the domain. We will consider two such estimators where the weights are functions of sample size in the domain. The first one is due to Drew, Singh, and Choudhry (1982) and is defined as

$$ {}_a\hat{Y}_{SSD} = \sum_g \left[{}_a\alpha_g \left(\sum_{h \in \underset{\sim}{h}} {}_a\hat{Y}_{hg} \right) {}_aR_g + (1 - {}_a\alpha_g) \sum_{h \in \underset{\sim}{h}} \hat{Y}_{hg} \cdot \frac{\sum_{h \in \underset{\sim}{h}} {}_aX_{hg}}{\sum_{h \in \underset{\sim}{h}} X_{hg}} \right] \qquad (2.8) $$

where

$$ {}_a\alpha_g = 1/{}_aR_g \quad \text{if } {}_aR_g > 1 $$

$$ = 1 \qquad \text{otherwise} $$

When the ratio ${}_aR_g > 1$, the domain is under represented with respect to the population sub-group g and vice-versa. Another estimator where the weights depend on the sample size is the count version of modified regression estimator suggested by Hidiroglou and Sarndal (1984). When the auxiliary variable is count of persons by age/sex, the modified regression estimator can be written in the same form as ${}_a\hat{Y}_{SSD}$ by replacing ${}_a\alpha_g$ by ${}_a\alpha^2{}_g$. The modified regression estimator will be denoted by ${}_a\hat{Y}_{MRE}$. Since ${}_a\alpha_g \leq 1$, the estimator ${}_a\hat{Y}_{MRE}$ has more reliance on the synthetic component as compared to ${}_a\hat{Y}_{SSD}$ and consequently more potential for bias.

(iv) Structure Preserving Estimation (SPREE)

The structure preserving estimates (SPREE) for frequency (or count) data were suggested by Purcell and Kish (1979). The method is based on the Iterative Proportional Fitting (IPF) procedure of Deming and Stephan (1940). The association structure which are the counts at the small area level cross classified by LF status and age/sex groups, i.e., $N = \{{}_aN_{ig}\}$ where ${}_aN_{ig}$ is the count of persons in small area 'a' belonging to the labour force status 'i' and age/sex group 'g', is defined on the basis of recent surveys. Also the current information $m = \{m_{ig}, {}_am_{..}\}$, known as allocation structure, is available from the current survey and other sources, e.g., m_{ig} which is the count of persons belonging to the labour force category 'i' and age/sex group 'g' at the large area level is estimated from the labour force survey and ${}_am_{..}$, the population for small area 'a', is available from external sources. Note that i = 1 for employed, = 2 for unemployed and = 3 for not in labour force.

Each cycle of the Iterative Proportional Fitting (IPF) procedure consists of two steps and the solution at the kth cycle is given by

$$ {}_a\hat{Y}^{(k)}_{ig} = {}_a\tilde{Y}^{(k-1)}_{ig} \; \frac{m_{ig}}{\sum_a {}_a\tilde{Y}^{(k-1)}_{ig}} \;, \text{ and} $$

$$ {}_a\tilde{Y}_{ig}^{(k)} = {}_a\hat{Y}_{ig}^{(k)} \frac{{}_a m_{..}}{{}_a\tilde{Y}_{..}^{(k)}} \; ; \quad k = 1, 2, \ldots \qquad (2.9) $$

where the initial solution

$$ {}_a\tilde{Y}_{ig}^{(0)} = {}_aN_{ig}, \qquad \begin{aligned} a &= 1, 2, \ldots, A \\ i &= 1, 2, 3 \\ g &= 1, 2, \ldots, G. \end{aligned} $$

The iteration is continued until some convergence criterion is satisfied following an iteration cycle. The procedure was terminated when the absolute relative difference in each cell was less than 0.001 between two successive iterations. The final LF estimate for a particular labour force category 'i' for the small area 'a' is obtained by summing over the age/sex groups, i.e.,

$$ {}_a\tilde{Y}_{SPREE} = \sum_g {}_a\tilde{Y}_{ig}^{(k)} $$

where K is the cycle at which the iterative procedure was terminated. It was observed that the convergence was achieved in 3 iterations for majority of the cases.

3. Variance Estimation

In practice the current CD population will be available, therefore in our Monte Carlo evaluation we have made use of the total CD population for ratio adjustment. The direct estimator was excluded from this ratio adjustment because this particular estimator does not make use of any auxiliary information, and the SPREE, as described in the previous section, uses the current population of the small area in a different fashion. For other estimators in the study, i.e., (i) post-stratified domain, (ii) synthetic, and (iii) sample size dependent estimators, we used the Keyfitz's (1957) variance estimator for estimating the variance of a ratio. The quantity ${}_aR_g$ defined by eq. (2.5) is a random variable but, for the purpose of variance estimation, ${}_aR_g$ is considered as a constant. Therefore our variances are conditional variances given a particular configuration of ${}_aR_g$.

Let ${}_a\tilde{Y}_{(m)}^{(r)}$ be the ratio adjusted estimate for small area 'a' from the Monte Carlo replicate 'r' when using small area estimation method 'm', then

$$ {}_a\tilde{Y}^{(r)}_{(m)} = {}_a\hat{Y}^{(r)}_{(m)} \left(\frac{{}_aP}{{}_a\hat{P}^{(r)}_{(m)}} \right) , \tag{3.1} $$

where ${}_aP$ is the known population of the small area 'a' and ${}_a\hat{Y}^{(r)}_{(m)}$ and ${}_a\hat{P}^{(r)}_{(m)}$ are respectively the unadjusted estimates of the characteristic total ${}_aY$ and the population ${}_aP$ from Monte Carlo replicate 'r' when using method 'm'. For estimating the variance, each PSU in the stratum is considered as a replicate in both NSR and SR areas. There were always 2 PSU's selected per stratum in NSR areas and the number of selected PSU's varied from 4 to 10 in SR areas. Let ${}_a\tilde{Y}^{(r)}_{(m)hi}$ and ${}_a\tilde{P}^{(r)}_{(m)hi}$ be respectively the contributions to the estimate ${}_a\tilde{Y}^{(r)}_{(m)}$ and the population ${}_aP$ from the ith PSU in stratum h from rth Monte Carlo replicate when using method 'm'. The contribution to the small area estimate from a particular PSU is simply the sum of the final weights of the records with the characteristic under study and belonging to the small area.

Now define the D-value for the ith selected PSU in stratum h for small area 'a' from Monte Carlo replicate r when using method m,

$$ {}_aD^{(r)}_{(m)hi} = {}_a\tilde{Y}^{(r)}_{(m)hi} - {}_aR^{(r)}_{(m)} \, {}_a\tilde{P}^{(r)}_{(m)hi} , \tag{3.2} $$

where the ratio ${}_aR^{(r)}_{(m)}$ is given by

$$ {}_aR^{(r)}_{(m)} = {}_a\tilde{Y}^{(r)}_{(m)} / {}_aP . \tag{3.3} $$

The variance of ${}_a\tilde{Y}^{(r)}_{(m)}$ denoted by ${}_a\hat{V}^{(r)}_{(m)}$, is then given by

$$ {}_a\hat{V}^{(r)}_{(m)} = \sum_{h \in \underset{\sim}{h}} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left({}_aD^{(r)}_{(m)hi} - {}_a\bar{D}^{(r)}_{(m)h} \right)^2 , \tag{3.4} $$

where

$$ {}_a\bar{D}^{(r)}_{(m)h} = \frac{1}{n_h} \sum_{i=1}^{n_h} {}_aD^{(r)}_{(m)hi} , $$

n_h = number of PSU's selected from stratum h, and

$$ \underset{\sim}{h} = [h \mid {}_aU_h \neq \emptyset] $$

i.e., $\underset{\sim}{h}$ is the set of all strata which have non-null intersection with the domain 'a'.

4. Description of the Monte Carlo Study

The Canadian Labour Force Survey (LFS) follows a multi-stage clustered area sampling design (Platek and Singh; 1976). The LFS sample has undergone a major redesign following the 81 census (Singh, Drew, and Choudhry; 1984). In simulating the LFS design, the stratification from the redesigned sample was adopted and the sampling frame was constructed from 1981 census 2B data which is 1/5 sample within Enumeration Areas (EA's). The auxiliary information was based on 1976 census counts. The small areas in our study are the census divisions (CD's) or grouped CD's in the province of Nova Scotia. Out of 18 CD's in the province, one CD coincided with the Labour Force Economic Region (ER) and therefore was not included in the study. The remaining 17 CD's were grouped into 13 small areas for which estimates of employed and unemployed were obtained using small area estimation procedures discussed in section 2.

In each Monte Carlo sample (replicate), the LFS design was simulated through all stages of sampling and a total of 1,000 Monte Carlo samples were selected independently. The primaries and secondaries (where it was not the final stage) were selected based on census population or dwelling counts, while the final stage of sampling was a systematic sample of households. The PSU's in the NSR areas were selected on the basis of 1976 census population counts and all other selections were based on 1981 census households counts. During estimation, the final weights were obtained by performing two iterations of raking ratio procedure where each iteration consisted of total population 15+ adjustment at the sub-provincial level followed by population adjustment by age/sex at the province level. The sub-provincial areas were census metropolitan areas (CMA's) and non-CMA's within ER's. The age/sex groups were defined as: ages 15-24, 25-54, and 55+ for male and female populations. The population counts at the sub-provincial area level and the counts by age/sex at the province level were the 1981 census counts. The CD which coincided with one of the Labour Force ER's was excluded from the study only after the final weights had been obtained using raking ratio estimation procedure. The auxiliary information for the purpose of obtaining small area estimates was based on both 1976 Census (population by age/sex at the domain level) and 1981 Census (total population 15+ at the domain level). The same age/sex groups were defined at domain level as those at the province level for obtaining the final weights using the raking ratio estimation procedure.

5. Analysis of Results

5.1 Evaluation of Small Area Estimators

We denote by ${}_a\tilde{Y}^{(r)}_{(m)}$ the estimate of total ${}_aY$ for small area 'a' from the rth Monte Carlo replicate when using the method m. Then the percent absolute relative bias of method m for small area 'a' is given by

$$
\begin{aligned}
{}_aARB_{(m)} &= 100 \left| \frac{1}{1000} \sum_r {}_a\tilde{Y}^{(r)}_{(m)} \Big/ {}_aY - 1 \right| \\
&= \left| \frac{1}{10} \sum_r {}_a\tilde{Y}^{(r)}_{(m)} \Big/ {}_aY - 100 \right| ,
\end{aligned}
\tag{5.1}
$$

and the average over all small areas is

$$
\overline{ARB}_{(m)} = \frac{1}{A} \sum_a {}_aARB_{(m)} , \tag{5.2}
$$

where A is the number of small areas in the study and is equal to 13. The mean square error of method m for small area 'a' is defined as

$$
{}_aMSE_{(m)} = \frac{1}{1000} \sum_r \left({}_a\tilde{Y}^{(r)}_{(m)} - {}_aY \right)^2 . \tag{5.3}
$$

and the percent root mean square error of method m for small area 'a' is

$$
{}_aRMSE_{(m)} = 100 \left({}_aMSE_{(m)} \right)^{1/2} \Big/ {}_aY . \tag{5.4}
$$

The average percent root mean square error of method m over all areas will be

$$
\overline{RMSE}_{(m)} = \frac{1}{A} \sum_a {}_aRMSE_{(m)} . \tag{5.5}
$$

The average percent absolute relative biases and the average percent root mean square errors of the small area estimators for the LF characteristics employed and unemployed are presented in table 1 for auxiliary variables: (a) total population 15+, and (b) population by age/sex. The synthetic estimator (SYN^*) in table 1 was constructed by using data from the whole ER, whereas the synthetic estimator (SYN) utilizes the data from only those strata which are intersected by the small area under consideration, i.e., ${}_aU_h \neq \emptyset$. The following observations are made:

(i) Effect of Auxiliary Variable

The performance of population by age/sex as auxiliary variable is uniformly superior to total population for all estimators when estimating employment, although the improvement is only marginal. For unemployed, it makes very little difference whether total population or population by age/sex was used as auxiliary variable.

(ii) Relative Performance of Estimators

(a) Bias Consideration

The biases of the two synthetic estimators SYN and SYN^*, and the SPREE are very large for both employed and unemployed. When comparing the bias of SPREE with that of the synthetic estimators, SPREE has larger bias for unemployed. The bias of synthetic estimator increases when data is used from larger area as is seen from a comparison of SYN with SYN^*. The bias of post-stratified domain estimator (PSD) is quite small, and is less than the unbiased direct (EXP) estimator which in fact is observed only because of Monte Carlo variance of the estimator. The bias of modified regression (MRE) estimator suggested by Hidiroglou and Sarndal (1984) is 1.5 times that of sample size dependent (SSD) estimator of Drew, Singh, and Choudhry (1982) for both employed and unemployed, but the biases of both of these estimators are negligible.

(b) Efficiency Considerations

There are very large gains for all the estimators over the direct estimator when estimating employment and the gains for unemployment, although not as much as for employment, are also very substantial. As expected, the two synthetic estimators and the SPREE have the smallest average percent root mean square errors, but these estimators also have very high biases. The sample size dependent (SSD) estimator and modified regression (MRE) estimator both gain over the post-stratified domain (PSD) estimator and the

gains for MRE estimator are more due to its increased reliance on the synthetic (SYN) component, but the differences are small. From the point of view of efficiency, these estimators (SSD and MRE) compare well with the synthetic (SYN) estimator.

5.2 Evaluation of Variance Estimator

We have ${}_a\hat{V}^{(r)}_{(m)}$ as the estimated variance of ${}_a\tilde{Y}^{(r)}_{(m)}$, the ratio adjusted estimate for small area 'a' from rth Monte Carlo replicate 'r' when using method m. The percent standard deviation (SD) of method m for small area 'a' is given by

$$ {}_aSD_{(m)} = 100\,({}_a\bar{\hat{V}}_{(m)})^{1/2}/{}_a\bar{\tilde{Y}}_{(m)}, \tag{5.6} $$

where

$$ {}_a\bar{\hat{V}}_{(m)} = \frac{1}{1000} \sum_r {}_a\hat{V}^{(r)}_{(m)} \quad , \quad \text{and} $$

$$ {}_a\bar{\tilde{Y}}_{(m)} = \frac{1}{1000} \sum_r {}_a\tilde{Y}^{(r)}_{(m)} \quad . $$

${}_a\bar{\tilde{Y}}_{(m)}$ will be very close to ${}_aY$ when the bias for method m is negligible.

Now we define the average percent standard deviation of method m over all small areas as

$$ \bar{SD}_{(m)} = \frac{1}{A} \sum_a {}_aSD_{(m)} \quad . \tag{5.7} $$

Also, from 1,000 Monte Carlo replicates, we can obtain the Monte Carlo variance of ${}_a\tilde{Y}_{(m)}$ given by

$$ {}_aV^*_{(m)} = \frac{1}{1000} \sum_r ({}_a\tilde{Y}^{(r)}_{(m)} - {}_a\bar{\tilde{Y}}_{(m)})^2 \quad , \tag{5.8} $$

and as before obtain the average Monte Carlo percent standard deviation ($\bar{SD}^*$) for method m over all areas as

$$ \bar{SD}^*_{(m)} = \frac{1}{A} \sum_a {}_aSD^*_{(m)} \tag{5.9} $$

where

$$ {}_aSD^*_{(m)} = 100 ({}_aV^*_{(m)})^{1/2} / {}_a\tilde{\bar{Y}}_{(m)} $$

The average percent standard deviation of method m obtained from the variance estimator, i.e., $\overline{SD}_{(m)}$ and from the Monte Carlo variance, i.e., $\overline{SD}^*_{(m)}$ are given in table 2 for employed and unemployed for the following methods - post-stratified domain (PSD), synthetic (SYN), sample size dependent (SSD), and modified regression (MRE). We observe that the synthetic estimator has slight positive bias which simply confirms the known fact that Keyfitz variance estimator overestimates the design variance when sampling is without replacement. But the variance of synthetic estimator will be of little use when there is large bias associated with this estimator. For SSD and MRE estimators, the average performance of the variance estimator for employed is quite satisfactory but unemployed variance has small negative bias. For SSD estimator, the biases of the standard deviations in individual small areas varied from less than 1% to 11% for both employed and unemployed. Out of 13 areas in the study, the biases of standard deviations of SSD estimator were negative for both employed and unemployed in 7 areas and positive for both employed and unemployed in 4 areas. In the remaining two areas, these biases were positive for employed and negative for unemployed. There were only 4 areas where the magnitude of biases were very large and these were from those 7 areas which had negative biases for both employed and unemployed. In these 4 areas very large biases in the standard deviation of PSD estimator were also observed for both employed and unemployed. An examination of the composition of the small areas revealed that these were the only 4 areas which did not contain any complete stratum. One of these 4 areas was only one partial stratum and the remaining three consisted of 2 partial strata each. The worst bias was for the standard deviation of PSD estimator in the area which consisted of only one partial stratum (standard deviation was underestimated by 45.8% for employed and 31.7% for unemployed) and these biases were also large for the other three areas. The large negative biases for these areas are obtained because Keyfitz's variance estimator results in estimated variance equal to zero whenever there is only one replicate falling in the small area from the selected sample. Moreover the estimate and the corresponding estimated

variance area both zero for PSD estimator when there is no sample in the small area. When there is only one replicate in the area, the synthetic approach to variance estimation suggested by Drew, Singh, and Choudhry (1982) may be followed. For small areas with no sample, the PSD estimator and the corresponding variance are not very meaningful. The performance of Keyfitz's variance estimator should be satisfactory in all other cases. The standard deviations of the MRE estimator in individual areas was not examined as it is expected that these should be similar to the standard deviations of the SSD estimator. The variance of the synthetic (SYN) estimator is simply the design variance but this will not measure the total error of the small area estimate because the bias is not included.

5.3 Effect of Survey Design on Small Area Estimation

The performance of alternate small area estimation methods also depends on the particular survey design, e.g., level of stratification and PSU sizes. If the survey design features deep geographic stratification then it is more likely that most of the small areas will contain some complete strata and as a result certain minimum sample size will be guaranteed for the small area.

Let us partition the population of a given small area 'a' (i.e., ${}_aP$) into two components: ${}_{a1}P$, the population of the small area in complete strata, and ${}_{a2}P$, the population of small area in partial strata. If the proportion of population in partial strata (i.e., ${}_{a2}P/{}_aP$) is small, then in most cases simple estimators, e.g., PSD will yield reliable estimates for the small area. Another factor affecting the reliability of the estimates is the population of the small area in the partial strata as a proportion of the total population of the partial strata, i.e, ${}_{a2}P/{}_aP_{h*}$, where ${}_aP_{h*}$ is the population of all the strata which are only partially contained in the small area. The proportion should be large for reliable small area estimates because it will be more likely that the sample will fall in the small area. Considering both factors, the product ${}_{a2}P/{}_aP.\ {}_aP_{h*}/{}_{a2}P\ (= {}_aP_{h*}/{}_aP)$ should be small for reliable small area estimates, i.e., the population of the partial strata should be small as compared to the population of the small area. We divided the 13 small areas in our study into 2 groups on the basis of the values of the ratio ${}_aP_{h*}/{}_aP$. Group 1 consisted of 6 areas with small values for ${}_aP_{h*}/{}_aP$ and group 2 consisted of the remaining 7 areas. The average standard errors for group 1 were smaller than those for group 2. The relative biases of synthetic and SPREE estimators were also smaller for group 1.

6. Conclusion

The post-stratified domain (PSD) estimator, although on the average considerably more efficient as compared to the simple domain (EXP) estimator, performs very poorly in those domains which consist of only partial strata. The two synthetic estimators (SYN and SYN*) and the SPREE had very high biases for both employed and unemployed. The efficiencies of the two composite estimators considered in this study (SSD and MRE) are comparable to each other but the MRE estimator has more bias as compared to the SSD estimator due to its increased reliance on the synthetic component. When considering both bias and the efficiency of the estimators, the sample size dependent (SSD) estimator would seem to be preferable to other estimators examined in the context of Labour Force Survey. The sampling rates in Nova Scotia are relatively high and the magnitudes of relative biases and efficiencies of these estimators may be different in other provinces where the sampling rates are low, e.g., Quebec and Ontario.

The Keyfitz's (1957) variance estimator for the variance of ratio can be applied for estimating the variance of small area estimators which incorporate population adjustment at the small area level. But the variance estimator breaks down when there is only one replicate in the small area from the selected sample. In this case, the synthetic approach to variance estimation suggested by Drew, Singh, and Choudhry (1982) may be followed. When there is no sample in the small area, model based small area estimates are constructed and the bias of the resulting estimate depends on the validity of the model assumptions. Robust small area estimates may be obtained from models using different types of symptomatic variables and survey data from other small areas.

The authors are grateful to R. Platek, J.N.K. Rao, G.B. Gray, M.P. Singh, J.D. Drew, and the reviewer for useful suggestions.

Table 1: Average Percent Absolute Relative Bias (ARB) and Average Percent Root Mean Square Error (RMSE) for Employed and Unemployed by Method

Method	Auxiliary Variable +	Average % ARB		Average % RMSE	
		Employed	Unemployed	Employed	Unemployed
EXP	-	0.50	0.94	25.35	35.92
PSD	1	0.29	0.63	7.07	26.47
	2	0.33	0.58	6.96	26.61
SYN	1	2.26	7.03	5.47	20.39
	2	2.05	6.48	5.39	20.19
SYN*	1	3.76	7.87	5.18	15.68
	2	2.85	7.70	4.51	15.74
SSD	1	0.32	1.02	6.12	23.81
	2	0.31	1.01	6.01	23.79
MRE	1	0.51	1.52	5.92	22.98
	2	0.48	1.44	5.82	22.95
SPREE	-	1.56	11.32	3.69	17.34

+ 1 = total population 15+
2 = population by age/sex groups

Table 2: Comparison of Average Percent Standard Deviation of Small Area Estimators from Keyfitz Variance Estimator and Monte Carlo Variance for Employed and Unemployed

Method †	Average % S.D. From			
	Keyfitz Variance Estimator		Monte Carlo Variance	
	Employed	Unemployed	Employed	Unemployed
PSD	6.32	23.84	7.08	26.40
SYN	5.05	18.66	4.80	18.22
SSD	6.00	22.57	6.11	23.68
MRE	5.86	21.98	5.90	22.78

† Auxiliary variable is total population 15+

REFERENCES

1. Deming, W.E., and Stephan, F.F. (1940), "On a least square adjustment of a sampled frequency table when the expected marginal totals are known", Annals of Mathematical Statistics, Vol. 11, pp. 427-444.

2. Drew, J.D., Singh, M.P., and Choudhry, G.H. (1982), "Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey", Survey Methodology, Vol. 8, No. 1, pp. 17-47.

3. Ghangurde, P.D., and Singh, M.P. (1977), "Synthetic estimation in Periodic Household Surveys", Survey Methodology, Vol. 3, No. 2, pp. 152-181.

4. Ghangurde, P.D., and Singh, M.P. (1978), "Evaluation of Efficiency of Synthetic Estimates", Proceedings of American Statistical Association, Social Statistics Section, pp. 52-61.

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

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

7. Hidiroglou, M.A., and Sarndal, C.E. (1984), "Experiments with modified regression estimators for small domains", Statistics Canada Report.

8. Keyfitz, N. (1957), "Estimates of sampling variance where two units are selected from each stratum", Journal American Statistical Association, Vol. 52, pp. 503-510.

9. Laake, P. (1978), "An evaluation of synthetic estimates of employment", Scandinavian Journal of Statistics, Vol. 5, pp. 57-60.

10. Laake, P. (1979), "A prediction approach to sub-domain estimation in infinite population", Journal American Statistical Association, Vol. 74, pp. 355-358.

11. Purcell, N.J., and Kish, L. (1979), "Estimation for small domains", Biometrics, Vol. 35, pp. 365-384.

12. Singh, M.P., Drew, J.D., and Choudhry, G.H. (1984), "Post '81 censal redesign of the Canadian Labour Force Survey", Survey Methodology, Vol. 10, No. 2, pp. 127-140.

13. Verma, R., and Basavarajappa, K.G. (1985), "Recent developments in the estimation of population for small areas by regression methods", Statistics Canada Report.