# EVALUATION OF SMALL AREA ESTIMATION TECHNIQUES FOR THE CANADIAN LABOUR FORCE SURVEY[1]

J.D. Drew, M.P. Singh, G.H. Choudhry[2]

Estimates from sample surveys are sometimes required for domains whose boundaries do not coincide with those of design strata. Taking the Canadian Labour Force Survey as an example of a survey utilizing a clustered sample design, some alternative small area estimation techniques available in the literature are evaluated empirically including synthetic, domain (simple and post-stratified) and composite estimators which are linear combinations of synthetic and post-stratified domain estimators. A sample dependent estimator which attaches weight to the post-stratified domain estimate depending on the amount of sample in the domain is proposed and its performance is also evaluated.

## 1. INTRODUCTION

With increasing emphasis on planning, administering and monitoring social and fiscal programs at local levels, there has been demand for more and good quality data at these levels from various municipal, provincial and federal government departments as well as from private institutions. The type of data required ranges from simple population counts to complex socio-economic variables such as employment, unemployment, income, houseing, proverty indices, health conditions and facilities etc. However, until recently not much attention had been paid to the development of sound statistical estimation techniques for small area data, with the notable exception of statistical demographers who for some time have been investigating the particular problem of small area population estimates, and who have identified several competing methods based on the use of administrative data and other sources.

[1] Presented at the Annual meetings of the American Statistical Association, Cincinnati, August 1982

[2] J.D. Drew, M.P. Singh and G.H. Choudhry, Census and Household Survey Methods Division, Statistics Canada.

A comprehensive review of existing small area (domain) estimation techniques along with their limitations is given by Purcell and Kish (1979). From the research done to date it is clear that there is not a unique best solution to the small area estimation problem. The choice of a particular method for small area estimation will depend on the data needs and on the richness and availablility of data sources, which differ from country to country, and within countries from one subject matter to another. Therefore, the classification of the type of small areas (domains) and examination of the data sources available in a particular context, followed by thorough investigation of the alternative small area estimation techniques for given situations, seems to be the most appropriate approach to development of small area data. In this context, we shall use the following classification of domains suggested by Purcell and Kish (1979) and point out the type of domain to which developments in this articles primarily refer.

(a) Planned domains - for which separate samples have been planned, designed, and selected. In the Canadian context, such domains for example may be economic or planning regions within a province or the province itself.

(b) Cross Classes - which cut across the sample design and the sample units (may also be referred to as characteristic domains); e.g., age/sex, occupation, industry.

(c) Unplanned Domains - that have not been distinguished at the time of sample design and thus may cut across the design strata or the primary sampling units (PSU's) within the strata. Examples of these in the Canadian context include Federal Electoral Districts, and Census Divisions or subdivisions, counties and manpower planning regions.

It should be noted that both types (a) and (c) refer to areal domains.

We consider this distinction of the domains into the above types important since the form of the estimator as well as its efficiency would depend upon the particular type of application. As pointed out by Purcell and Kish most of the developments in small area estimation techniques in the United States and elsewhere have concentrated on the domains of types (a) and (b). In Canada however, type (a) and (b) domains are not so problematic due to the type of design and the sizes of the national surveys, and the main emphasis has been on the data for the domains of type (c), with the possible exception of population counts using symptomatic data.

Investigations into the application and evaluation of small area estimation techniques for variables other than population started with the publication of synthetic estimates from the National Center for Health Statistics (1968). Since then a series of investigations (Gonzalez (1973), Gonzalez and Waksberg (1973), Schaible, Brock and Schnack (1977), Gonzalez and Hoza (1978) and others) have been carried out usingdata from the Current Population Survey in the application and evaluation of a particular synthetic estimator. Using a synthetic estimator whose form is different, studies were carried out by Purcell and Linacre (1976) aimed at production of estimates for Census divisions in Australia and by Ghangurde and Singh (1976, 1977, 1978) in the evaluation of synthetic estimates in the context of Canadian Labour Force Survey (LFS).

As remarked by Purcell and Kish (1979), the nature of the design in relation to the domains of interest has an important role to play in the choice of an estimator. The estimators considered in this article are geared to the Canadian LFS where the domains are unplanned domains (typec) and are of a size such that, had they been planned domains (type a), the reliability of regular unbiased survey estimates would be satisfactory without having to resort to small areas estimation techniques. Also in the LFS, primary sampling units are small (populations from 2,000 - 5,000) relative to the sizes of the domains of interest. This differs from the situation in the United States where the sizes of primary sampling units for most of the large scale surveys are larger, comparable in size to the small areas for which the estimates are desired.

In this article estimators are evaluated in the context of producing Census Division level estimates from the Labour Force Survey, using data from the 1971 and 1976 Censuses of Population and Housing in an auxiliary fashion. In addition to synthetic estimators, we evaluate post-stratified domain estimators which were considered earlier by Singh and Tessier (1976), and composite estimators which are linear combinations of the synthetic and the post-stratified domain estimators, similar to those considered by Schaible (1979) and Schaible, Brock and Schnack (1977). Also we propose and evaluate a new estimator which we call a sample dependent estimator, which is of the same form as the composite estimator, except the weight given to the synthetic component is a decreasing function of the amount of sample falling into the domain upto a critical point after which the estimator relies totally on the post-stratified domain component. Efficiencies of the small area estimators relative to the direct (or simple domain) estimator for the characteristics employed and unemployed were obtained in an empirical (Monte Carlo) study in which the LFS design was simulated using census data. The situations where both the design and the auxiliary information are up-to-date and where both are out-of-date were considered. We have also evaluated the bias of synthetic estimators for the characteristics employed and unemployed for Federal Electoral Districts.

## 2. DESCRIPTION OF ESTIMATION PROCEDURE

Consider a finite population consisting of N units, (e.g. households or dwellings in household surveys), divided into L design strata labelled 1, 2, ..., h, ...L. The stratification has been carried out on the basis of geographic and/or certain socio-economic characteristics, and the sample allocation ensures certain precision for estimates from individual strata. The problem considered is that of estimating the total of an x-variate for all those unites belonging to an unplanned areal domain (type c). We denote by 'a' the set of units belonging to the small area or domain of interest, thus the parameter to be estimated is the total of the x-variable in the domain 'a', which we denote by ${}_aX$.

Let $a_h$ be the set of those units belonging to the domain which are in stratum h, then

$$a = \bigcup_{h=1}^{L} a_h. \qquad (2.1)$$

In practice the domain 'a' will have a non-null intersection with a certain number of design strata and if we denote by $\underset{\sim}{h}$ the set of such strata, then we have

$$a = \bigcup_{h \varepsilon \underset{\sim}{h}} a_h. \qquad (2.2)$$

The particular design under consideration follows a multi-stage clustered sample design which is self-weighting within each stratum with weight $W_h$ for stratum h.

For a particular given sample we can obtain the quantities:

$t_h$ = sample total of x-variate in stratum h,

and

${}_a t_h$ = sample total of x-variate in $a_h$

for h=1, 2, ..., L. Note that ${}_a t_h = 0$ for $h \not\in \underset{\sim}{h}$. Then the direct (or also referred to as design based or simple domain) estimator for the total of x-variate for those units in 'a' say ${}_a\hat{X}$, is given by:

$$\hat{X}_a = \sum_{h \varepsilon \underset{\sim}{h}} W_h \cdot {}_a t_h. \qquad (2.3)$$

It should be noted that the direct estimator (2.3) does not utilize any auxiliary information - all it requires is the identification of those sampled units which belong to the domain. Due to the clustered nature of the design, the sample falling in the domain may on occasion be very small or non-existent, generally resulting in high variance for this estimator.

The other estimators in this section rely in different fashions on auxiliary information for a variable y, which is often taken as the count of persons by population sub-groups (defined on the basis of age/sex etc.) from a recent census. These estimators are:

1) Post-stratified domain
2) Synthetic
3) Composite
4) Sample Dependent

Additionally estimators (2) - (4) rely to differing degrees on sample external to the domain.

For each of the above estimators, the adjustments based on the auxiliary information can be made either be applying separate adjustments to each stratum intersecting the domain, or by applying an overall adjustment for all strata intersecting the domain. Thus the estimators will be further classified as <u>separate</u> or <u>combined</u> depending on the level at which the adjustment is made. These estimators are denoted by ${}_a\hat{X}_{uv}$, where u is the level of adjustment with values:

u = s : separate
= c : combined

and v is the type of estimaror taking the following values:

v = p : post-stratified domain
= s : synthetic
= c : composite
= d : sample dependent

For example, ${}_a\hat{X}_{cs}$ denotes the <u>combined synthetic</u> estimator, etc.

## 2.1 Post-Stratified Domain Estimator

Define

$Y_{hg}$= total of the auxiliary y-variable for population sub-group g in group g in stratum h, and

${}_aY_{hg}$= total of the auxiliary y-variable for population sub-group g in $a_h$.

Further let ${}_a\tilde{Y}_{hg}$ be an unbiased estimated of ${}_aY_{hg}$ which would be formed analogously to the direct estimate defined in (2.1), except the characteristic being estimated in this case would be the auxiliary y-variable whose value is known for the set of sampled units (s) at some stage of sampling (whereas (2.1) is defined on the x-variate for the sample of ultimate units). In practice provided auxiliary y-variable information is available for them, sampling units at any stage down to the penultimate stage could be used.

Then the separate post-stratified domain estimator (for which adjustments are applied at the stratum level) is:

$$ {}_a\hat{X}_{sp} = \sum_{g} \sum_{h \in \underset{\sim}{h}} (W_h \cdot {}_at_{hg}) \frac{{}_aY_{hg}}{{}_a\tilde{Y}_{hg}} \qquad (2.4) $$

where ${}_at_{hg}$ is the sample total of the x-variate for population sub-group g in the intersection of domain 'a' and stratum h.

Similarly the combined post-stratified domain estimator (for which adjustments are applied at the domain level) is:

$$ {}_a\hat{X}_{cp} = \sum_{g} \sum_{h \in \underset{\sim}{a}} (W_h \cdot {}_at_{hg}) \frac{\sum_{h \in \underset{\sim}{a}} {}_aY_{hg}}{\sum_{h \in \underset{\sim}{a}} {}_a\tilde{Y}_{hg}} \tag{2.5} $$

The post-stratified domain estimator is unbiased except for the effect of ratio estimation bias, provided ${}_aY_{hg}$ is obtained at the same time as ${}_a\tilde{Y}_{hg}$ and using the same source such as census.

Estimators of the above type have been considered earlier by Singh and Tessier (1976) with a different choice of post-strata.

## 2.2 Synthetic Estimators

We consider separate and combined synthetic estimators defined respectively as follows:

$$ {}_a\hat{X}_{ss} = \sum_{g} \sum_{h \in \underset{\sim}{a}} (W_h \cdot t_{hg}) \frac{{}_aY_{hg}}{Y_{hg}} \tag{2.6} $$

$$ {}_a\hat{X}_{cs} = \sum_{g} \sum_{h \in \underset{\sim}{a}} (W_h \cdot t_{hg}) \frac{\sum_{h \in \underset{\sim}{a}} {}_aY_{hg}}{\sum_{h \in \underset{\sim}{a}} Y_{hg}} , \tag{2.7} $$

where $t_{hg}$ is the sample total for the x-variable for population sub-group g in stratum h.

The above synthetic estimator has been considred by Purcell and Linacre (1976) and also by Ghangurde and Singh, (1976, 1977, 1978) who developed expressions for its variance and bias and evaluated the estimator using census data and a super-population model. A different form of syntheticestimator was proposed earlier by the National Centre for Health Statistics (1968) and investigated by Gonzalez (1973), Gonzalez and Waksberg (1973) and Gonzalez and Hoza (1975, 1978) using data form the Current Population Survey.

The difference between the synthetic and post-stratified domain estimators can be readily seen by comparing (2.4) and (2.6). The post-stratified domain estimator uses only the sample falling into the domain (i.e., ${}_{a}t_{hg}$) and the adjustment factor is the ratio of the true to the estimated values for the y-variable for the domain and hence can take on values greater than or less than 1 (its expected value being unity). On the other hand the synthetic estimator uses the estimate from entire strata intersected by the domain (i.e. $w_h \cdot t_{hg}$ for $h \in \underset{\sim}{h}$) which is then deflated by adjustment factors specific to population subgroups. (i.e. the ratio of the y-variable for the domain to the y-variable for the entire stratum).

The synthetic estimator will suffer from bias depending on the degree of departure from the assumption of homogeneity for the x-variate between the domain and the larger area, namely $\underset{\sim}{h}$, withing sub-groups of the y-variable. In defining the above synthetic estimator, the larger area was restricted to those strata which form part of the domain as it was believed that such a choice would lead to less bias. In general however, $\underset{\sim}{h}$ need not be so restricted but it may include other neighbouring ares which are believed to satisfy the homogeneity assumption. Bias and mean square error of such estimators have been reported by some of the earlier referenced authors.

## 2.3 Composite Estimators

A composite estimator using the direct estimator and the synthetic estimator as the two components was suggested by Royall (1973) and others, and has been studied by Schaible (1978). Such an estimator minimises the chances of extreme situations (both in terms of bias and mean square error) and therefore may be preferred over either of its components. Synthetic estimators have a low variance by virtue of their use of data from a larger area to derive estimates for small are (domain), but for the same reason this introduces bias which could be quite large if as noted earlier, the assumption of homogeneity is not satisfied. On the other hand the simple domain estimator, which is unbiased, may have large variance particularly if the sample falling in the domain is very small. Empirical evidence of such relative performances of synthetic and direct estimators are available from Gozalez and Waksberg (1975)

Schaible, Brock and Schnack (1977), and Ghangurde and Singh (1977). The composite estimator considered here is obtained by replacing the direct estimator (2.3) by the post-stratified domain estimator which may be slightly biased but is generally more efficient than the direct estimator.

The two types of composite estimators: namely, separate and combined are formed as linear combinations of the corresponding post-stratified domain and synthetic estimators; viz,

$${}_a\hat{X}_{sc} = \alpha_1 \ {}_a\hat{X}_{sp} + (1 - \alpha_1) \ {}_a\hat{X}_{ss} \qquad (2.8)$$

and

$${}_a\hat{X}_{cc} = \alpha_2 \ {}_a\hat{X}_{cp} + (1 - \alpha_2) \ {}_a\hat{X}_{cs} \qquad (2.9)$$

The optimum values for $\alpha_1$ and $\alpha_2$ for minimum mse's are given by

$$\overset{*}{\alpha}_1 = \frac{\text{mse} \left[{}_a\hat{X}_{ss}\right] - E \left[{}_a\hat{X}_{ss} - {}_aX\right] \left[{}_a\hat{X}_{sp} - {}_aX\right]}{\text{mse} \left[{}_a\hat{X}_{ss}\right] + \text{mse} \left[{}_a\hat{X}_{sp}\right] - 2\, E \left[{}_a\hat{X}_{ss} - {}_aX\right] \left[{}_a\hat{X}_{sp} - {}_aX\right]} \qquad (2.10)$$

and a similar expression for $\overset{*}{\alpha}_2$.

Further, neglecting the covariance term in (2.10) under the assumption that this term will be small relative to mse $[{}_a\hat{X}_{ss}]$ and mse $[{}_a\hat{X}_{sp}]$, then the optimal weight $\overset{*}{\alpha}_1$ can be approximated by

$$\overset{**}{\alpha}_1 = \frac{\text{mse} \left[{}_a\hat{X}_{ss}\right]}{\text{mse} \left[{}_a\hat{X}_{ss}\right] + \text{mse} \left[{}_a\hat{X}_{sp}\right]} \qquad (2.11)$$

with a similar expression for $\overset{**}{\alpha}_2$, which was the approach to defining weights followed by Schaible (1978).

## 2.4 Sample Dependent Estimators

In practice the true values of $\alpha_1$* (or $\alpha$*$_2$) used as the weight in the composite estimator will not be available as they involve population variances and covariances, which would have to be estimated from the sample. Further calculation of the covariance term in (2.10), in particular, may be quite complex and thus one may have to resort to an approximate value $\alpha_1$** (or $\alpha_2$**) which would require simply the estimated mse's of the two component estimators or an estimate of the ratio of the two mse's. In either case there estimates would introduce a certain amount of instability in the weight used, thus affecting the performance of the composite estimator.

The sample dependent estimator (Drew and Choudhry, (1979)) which is a particular case of a composite estimator, depends on the outcome of the given sample and is quite simple to compute. It is constructed using the result that the performance of the post-stratified domain estimator depends upon the proportion of the sample falling in the domain. If the proportion of the sample within the domain is 'reasonably large' then the sample dependent estimator is the same as the post-stratified domain estimator, otherwise it becomes a composite estimator with gradual increasing reliance (in the sense of increasing weight) on the synthetic estimator as the size of the sample in the domain decreases. Thus the separate sample dependent estimator (i.e., constructed at the stratum level) is given by

$$\hat{X}_{a\,sd} = \sum_g \sum_{h \ni \underset{\sim}{h}} \lfloor \delta_{hg} \; W_h \, . \, t_{a\,hg} \frac{{}_aY_{hg}}{\tilde{Y}_{a\,hg}} + (1 - \delta_{hg}) \, W_h \, . \, t_{hg} \frac{{}_aY_{hg}}{\tilde{Y}_{a\,hg}} \rfloor \quad (2.12)$$

where

$$\delta_{hg} = 1, \text{ if } {}_a\tilde{Y}_{hg} / {}_aY_{hg} \geq K_o \, ,$$

$$= \frac{1}{K_o} \; \frac{{}_a\tilde{Y}_{hg}}{{}_aY_{hg}} \, , \qquad \text{otherwise.}$$

Similarly the combined sample dependent estimator (i.e. constructed at the domain level) is given by

$$\hat{X}_{acd} = \sum_g \lfloor \delta_g \left( \sum_{h \in \underset{\sim}{h}} W_h \, . \, t_{ahg} \right) \frac{\sum_{h \in \underset{\sim}{h}} Y_{ahg}}{\sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg}}$$

$$+ (1 - \delta_g) \left( \sum_{h \in \underset{\sim}{h}} W_h \, . \, t_{hg} \right) \frac{\sum_{h \in \underset{\sim}{h}} Y_{ahg}}{\sum_{h \in \underset{\sim}{h}} Y_{hg}} \rfloor \qquad (2.13)$$

where

$$\delta_g = 1, \text{ if } \sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg} \Big/ \sum_{h \in \underset{\sim}{h}} Y_{ahg} \geq K_o$$

$$= \frac{1}{K_o} \sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg} \Big/ \sum_{h \in \underset{\sim}{h}} Y_{ahg}, \qquad \text{otherwise}$$

The ratios

$$\tilde{Y}_{ahg} / Y_{ahg} \quad \text{and} \quad \sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg} \Big/ \sum_{h \in \underset{\sim}{h}} Y_{ahg}$$

indicate the over- or under-representation of the population sub-group at the individual stratum or domain level with respect to auxiliary information for the y-variable, conditional upon the selected sample.

Values of ratios greater than or equal to 1 signify that, conditional onthe given sample (s), the representation of the population sub-groups for the auxiliary y-variable is better than or as good as its unconditional representation had the domain been sampled independently at the same rate as the stratum.

The value of $K_o$ may be appropriately chosen. In this study the efficiency of sample dependent estimator has been investigated for two specific values of $K_o$ namely 1.0 and 0.5.

Holt, Smith and Tomberlin (1979) under the prediction approach derived an estimator (which relies on synthetic and direct estimates) where the weight attached to the direct component depends only on the sample falling into the domain. Sarndal (1981) proposed an alternative estimator in which the weight attached to the direct component depends on the sample in the domain relative to the sample in the larger area.

## 3. DESCRIPTION OF THE EMPIRICAL STUDY

### 3.1 Simulation of the LFS Design

The LFS follows a multi-stage area sampling design (see Platek and Singh, (1976)). Within each of the 10 provinces of Canada, two principal area types are identified - the Self-Representing Units (SRU's) which correspond to cities generally of 15,000 or more population, and the Non Self-Representing Units (NSRU's) which correspond to smaller urban centers and rural areas. In the SRU's, cities are divided into compact areal strata with populations of 15,000 each, within which a two stage sample of clusters (similar to blocks) and dwelling is selected.

In NSRU's, Economic Regions, of which there are from 1-10 per province, form the starting point. These are stratified into 1-5 strata with populations from 30,000 to 80,000 using census data for 7 broad industryclassifications. Within strata, primary sampling units (PSU's) from 2,000 - 5,000 in population are formed. The second stage in the rural portions of PSU's corresponds to 1971 Census Enumeration Areas (i.e., EA's), with populations of roughly 500, whereas in urban portions all urban centers are selected with certainty. The last two stages correspond to clusters and dwellings.

In simulating the LFS design two cases were examined: (i) the case whereboth the sample design and the auxiliary information are up-to-date, and(ii) the case where both are out-of-date.

For (i), the sample design, the auxiliary information, and the study variables were all based on 1971 census data. Counts of persons (15+) cross-classified by age/sex, and Labour Force status were retrieved at the EA level. In NSRU's, for each replication in the Monte Carlo study independent samples of primaries and secondaries were selected based on census population or dwelling counts. Within rural EA's and urban centers, the final two stages of sampling were simulated by random samples of persons. In SRU's, EA's comprising the areal strata were known, but there after the LFS design was independent of the census. Hence for the purposes of the study, EA's were randomly partitioned into 'clusters' having a size distribution corresponding to that for LFS clusters. For each replication, a sample of 'cluster' and a random sample of persons within were selected.

## 3.2 Choice of Population Sub-Groups

The estimators defined in section 2 utilize auxiliary information for population sub-groups. Since the LFS is redesigned only decennially, it would be desirable to base the population sub-groups on information collected in the mid-decade as well as decennial census, so that the auxiliary information could be updated mid-way through the life of the survey. This ruled out such variables as industry or occupation, leaving various cross-classifications of basic demographic variables as the possible choices for population sub-groups.

For the variables marital status, age and sex, the Automatic Interaction Detection (AID) procedure, due to Sonquist and Margan (1964) was used ona sample of census data from across Canada to derive optimal population sub-groups, separately for each Labour Force characteristic. Results ofthe AID analysis showed that for unemployed, no population sub-groups accounted for more than 2% of the variation, while for the characteristics employed and not in Labour Force the following sub-groups accounted for approximately 25% of the variation: (i) age 15-16 and 65+; (ii) age17-64, sex female; (iii) age 17-64, sex male. Further splitting of these sub-groups did not result in significant additional gains.

In addition to estimators based on the above population sub-groups, estimators based on total population 15+, and on dwelling counts were also considered. Dwelling count data were included due to the possibilities which exist for up-to-date dwelling information being available intercensally at the required level of detail. It might be noted that the estimators using population 15+ and dwelling counts are both special cases of the general formulation where the number of population sub-groups equals 1.

## 3.3 Evaluation of Efficiency of Small Area Estimators

In the Monte Carlo study, we have considered 16 Census Divisions (CD's) and 11 Federal Electoral Districts (FED's) in the province of Nova Scotia and 7 FED's from elsewhere in Canada. (There are altogether 18 CD's in the province of Nova Scotia, but two of 18 CD's correspond to complete LFS strata and therefore were omitted from the study). Due to the multi-stage nature of the design and larger number of domains in the study, the computational costs involved were high and it was decided to use only 100 replications.

Census Divisions and Federal Electoral Districts, it should be noted, comprise networks of geo-statistical and geo-political areas respectively across Canada. There are approximately 300 of each, with the populations of Federal Electoral Districts being fairly uniform in the range 80,000 to 120,000, while those of Census Divisions, which often correspond to local levels of government or counties, vary greatly.

We have reported results only for the 16 Census Divisions in Nova Scotia. Results were similar for other unplanned domains considered.

If we let ${}_{a}\hat{X}_{m(r)}$ be the estimate of total ${}_{a}X$ (i.e. the total for the x-variable for the domain 'a') for the r'th replicate, for small area estimation method m, then the average mean square error for the method m over the 16 domains in the study was calculated as:

$$\text{Avg mse } (m) = \frac{1}{16} \sum_{a} \sum_{r=1}^{100} ( {}_{a}\hat{X}_{m(r)} - {}_{a}X)^{2}/100 \quad . \qquad (3.1)$$

The efficicency of the small area estimator (m) relative to the direct estimator, say method $m_0$ was obtained as:

$$\text{Eff } (m \text{ vs } m_0) = \frac{\text{Avg mse } (m_0)}{\text{Avg mse } (m).} \qquad (3.2)$$

### 3.4 Evaluation of Bias of Synthetic Estimators

Since the composition of the LFS frame and the Federal Electoral Districts were known for all of Canada in terms of both 1971 and 1976 census units, it was possible to compute exact biases of the synthetic estimators based on census data. The following cases were considered: (i) design and auxiliary information up-to-date (in which case the design, adjustment factors and x-variables were all based on the 1971 census); and (ii) design and auxiliary information out-of-date (in which case the design and adjustment factors were based on the 1971 census, but the x-variables were based on the 1976 census).

Let ${}_aB_{ss}$ and ${}_aB_{cs}$ denote the biases of the separate and combined synthetic estimates for unplanned domain 'a', then we have

$$ {}_aB_{ss} = \sum_g \sum_{h \in \underset{\sim}{h}} (X_{hg} \frac{{}_aY_{hg}}{Y_{hg}} - {}_aX_{hg}) \qquad (3.3)$$

and

$$ {}_aB_{cs} = \sum_g ( \sum_{h \in \underset{\sim}{h}} X_{hg} \frac{\sum_{h \in \underset{\sim}{h}} {}_aY_{hg}}{\sum_{h \in \underset{\sim}{h}} Y_{hg}} - \sum_{h \in \underset{\sim}{h}} {}_aX_{hg} ) \qquad (3.4)$$

where ${}_aY_{hg}$ and $Y_{hg}$ are defined as in section 2, and where $X_{hg}$ and ${}_aX_{hg}$ are similarly defined for the x-variable (based on the census.

Relative absolute biases at the province level were obtained by summing the absolute biases over individual FED's and dividing by the provincial total for the x-variable.

## 4. ANALYSIS OF RESULTS

### 4.1 Efficiency considerations: Auxiliary Information up-to-date

In this part of the empirical (Monte Carlo) study, data used for simulation of the design and the auxiliary variables used in estimation refer to the same period as those of the study variable; i.e., to the 1971 census. Efficiencies of the four small area estimators are presented relative to the direct estimator in Table 1, for separate and combined levels of construction, and for each of the following auxiliary variables - dwellings, total population (15+), and population by age/sex groups. Census Divisions in the province of Nova Scotia whose populations range from 3,885 to 39,260 were used as the unplanned domains (type c) for the purpose of the study. The following observations can be made:

(i) Separate vs Combined Estimator: The level of construction of estimator does not have much impact on the efficiencies of synthetic estimators for both the characteristics employed and unemployed. For the post-stratified domain estimator for employed, however, the combined form is approximately twice as efficient as the separate. This is likely due to the effect of the clustering in the sample design being more accentuated with the separate estimator.

Since the post-stratified domain estimator was less efficient in its separate form, a similar result was anticipated for the composite estimator and hence, only the combined composite estimator was considered. On the other hand, the separate form of the sample dependent estimator was found to rely slighlty more on the synthetic component, leaving the efficiencies unaffected by the level of construction.

(ii) Effect of Auxiliary Information: The performance of population by age/sex as an auxiliary variable is uniformly superior, although only marginally so, to the total (15+) population for all four estimators using auxiliary information. Further, both these variables out-perform the dwelling count as an auxiliary variable.

In actual survey situations, the choice of population by age/sex as the auxiliary variable may be desirable also from the point of viex of correcting estimates for biases due to non-response and undercoverage as both factors may be dependent on age and sex.

(iii) Compararison among the estimators: For unemployed, performance of composite estimator with optimum $\alpha_2^*$ chosen for the characteristic unemployed is marginally superior to the other estimators irrespective of the level of construction, and the choice of auxiliary variable does not seem to have appreciable impact on any of the estimators. For employed, the situation is not that clear, however the sample dependent estimator shows an edge over other estimators and particularly so with population by age/sex as the auxiliary variable.

## 4.2 Efficiency Considerations: Auxiliary information out-of-date

In this part of the study whereas the design and auxiliary information were based on 1971 census results, the study variable was based on the 1976 census. As can be seen from table 2, although for unemployed the use of small area estimation techniques showed larger gains relative to the direct estimator (than in the up-to-date case), considerably smaller gains were observed for employed, which would likely be due to the reduced correlation between the study variable and the auxiliary information as both design and auxiliary information become out-of-date. Also in this case, the efficiency of the synthetic estimator is higher for both of the characteristics measured.

## 4.3 Consideration of Bias

Given that the post-stratified domain estimator will generally have negligible bias, the bias of both the composite and sample dependent estimators would generally be smaller than that of the synthetic estimator, i.e. stemming only from the degree of reliance on the synthetic component. Hence the bias of synthetic estimator was investigated in detail. Using the total population (15+) as the auxiliary variable, the relative bias for the characteristics employed and unemployed were computed and are given in Table 3 for the ten provinces. Theses biases refer to the case where the unplanned domains are Federal Electoral Districts and the study variables are based on 1976 census data, while the survey design and adjustment factors (synthetic weights) are based on the 1971 census. Biases were also computed using age/sex sub-groups as the auxiliary variable and were found to follow similar trends while being marginally smaller. It is observed from this table, with the exception of the two smaller provinces, namely P.E.I. (for unemployed) and N.B. (for employed), that the relative bias of separate synthetic estimator is smaller than that of the combined synthetic estimator for both the characteristics under study. This confirms the intuitive feeling that the higher the level at which synthetic estimator is constructed, the higher would be the resultant bias in general, due to weakening of the assumption of homogeneity.

Biases were also computed for the case when both the study variable and the auxiliary information referred to the 1971 census. Biases for this case while slightly lower, followed similar trends to those in Table 3.

While the bias of the synthetic estimator was fairly small on average, it can be observed from Table 4 that it exceeded 10% in 13 and 19 (out of 279) FED's when the auxiliary information was up-to-date and out-of-date respectively. Further, in about half the instances for which the bias exceeded 10% for the up-to-date case the bias also exceeded 10% for the later time period when the auxiliary information was out-of-date. This suggests that for domains with a known high bias at the time to which the auxiliary information refers, less use should be made of the synthetic estimator. For instance, with the sample dependent estimator the value of $K_o$ could be set lower in such cases. However there is still the danger of bias in the synthetic estimator from category (ii) type cases in Table 4 which cannot be identified when deriving current estimates during the intercensal period.

## 4.4 Efficiency vs Bias in Overall Choice of Estimator

The synthetic estimator is generally highly biased and at the same time highly efficient. Therefore, in the search for a reasonable estimator for small areas, the question is to what extent one can reduce the effect of the synthetic estimator's bias, without sacrificing too much on its efficiency, in order to obtain a 'reasonable level of confidence' in the final estimate. At the same time it is also important to determine the reliance on the synthetic estimator without introducing too many computational complexities. Looking from this perspective in the context of the Labour Force Survey, one should strive for small area estimators whose performance for unplanned domains is comparable to that of simple survey estimates for planned domains, and amongst estimators meeting this criterion, more emphasis should be on reducing bias than on improving efficiency, especially if the differences in efficiencies are minor.

Average variances of the unbiased design estimator for the planned domains (say $\hat{X}$), comparable in size to the unplanned domains were obtained analogously to the average mse defined in (3.1). The efficiencies of the synthetic, composite and sample dependent estimators relative to the usual survey estimate for the planned domain i.e. X were also obtained. These efficiencies ranged from 1.08 to 1.17 for unemployed, and 1.22 to 1.47 for employed, hence all three estimators meet the above mentioned criterion. Since the sample dependent estimator makes use of the synthetic estimators whenever there is not 'sufficient' sample in the domain, its bias would depend upon the weight attached to the synthetic estimator component and this can be controlled by a proper choice of $K_0$. Table 5 presents the $(1-\delta)$ values, averaged over 100 replicates with $K_0 = 0.5$ and $K_0 = 1.0$ for the separate sample dependent estimator using total population 15+ as the auxiliary variable for each of the Census Divisions (unplanned domains) in this study. These average $(1-\delta)$ values indicate the degree of reliance of the sample dependent estimator on the synthetic component. As expected, domains consisting primarily of partial strata tend to place increased reliance on the synthetic component. Nevertheless, that reliance remains quite small. For example, with $K_0 = 1$ the highest value it assumes is .28 for Census Division 218.

Also as expected, the average $(1-\delta)$ values for $K_0 = 0.5$ are lower than those for $K_0 = 1.0$, implying the lower the value of $K_0$ chosen, the lower would be the value of $(1-\delta)$ and consequently less reliance (weight) on the synthetic component of the sample dependent estimator. However as illustrated in Table 1, a trade-off between bias and efficiency is involved since lower choices of $K_0$ also result in reduced efficiency. The above values of $K_0$ provide a reasonable degree of confidence for the type of domains discussed here. In general, however, other values of $K_0$ may be chosen depending upon e.g. the size of the domain, sample size, strata sizes and their geographical configurations with respect to the domain.

## 4.5 Concluding Remarks

1. The use of population by age/sex fares uniformly better than the other auxiliary variables, although gains over total population (15+) are mariginal.

2. The post-stratified domain estimator although more efficient as compared to the simple domain estimator, performs poorly as compared to the other three small area estimators investigated.

3. From the point of view of bias, the separate estimator has smaller relative bias as compared to the combined synthetic estimator. Further while average biases tend to be fairly small and tend to increase only slightly when the auxiliary information became out-of-date, biases for individual domains can be very high and change dramatically, frustrating efforts to identify 'outliers' where reduced reliance on syntheic estimators should be made.

4. The combined composite estimator constructed as a linear combination of post-stratified and synthetic estimators is more efficient than either of its component estimators although only marginally so, as compared to the synthetic component, for optimum value of $\alpha$. Its bias would depend upon the weight attached to the synthetic component since the bias of the post-stratified estimator would generally be

negligible. Further, as the computation of the optimum $\alpha$ is quite involved, in practice only an estimated value of $\alpha$ may be used, resulting in a decrease in efficiency of this estimator.

5. The synthetic, composite and sample dependent estimator with $K_0 = 1$ are all more or less equally efficient, and out-perform the un biased design based estimator for planned domains.

6. Since the bias of the separate synthetic estimator is smaller than that of the combined synthetic estimator, the separate sample dependent estimator would result in smaller relative bias as compared to the combined sample dependent estimator. The bias of the separate to the combined sample dependent estimator. The bias of the separate post-stratified domain component can be controlled by collapsing those strata for which the intersection with the domain is very small. Thus considering all the three aspects, bias, mean square error and the computational complexities, the sample dependent estimator constructed at the stratum level using population by age and sex would seem to be a better choice.

## 5. FUTURE DIRECTION OF INVESTIGATION:

The study reported in this paper has focussed on evaluation of certain small area estimation methods using only census and survey data, in the context of the LFS, primarily for unplanned domains (type c). The estimators examined made use of synthetic and post-stratified domain estimators in different ways in an attempt to strike a balance between bias and mean square error. Below we point to directions which future investigations might take in efforts to develop statistically sound techniques for small area data in the Canadian context.

In the context of the Labour Force Survey, since the small area estimation methods for the unplanned domains have out-performed the unbiased design based estimates for comparable planned domains, it would be desireaable to extend this investigation to certain small planned domains (type a) as well. In particular the sample dependent estimator considered here and other similar estimators discussed in the literature will be further investigated for the Labour Force characteristics. In addition these investigations should also be extended to other smaller surveys conducted by Statistics Canada for which small area data are in demand. Further work on development of methods of variance estimation to be used in practice for these estimators is also needed.

Other estimators which seem to be promising are the Structure Preserving Estimators (SPREE) suggested by Purcell and Kish (1980). In this approach the estimation process, specified by the association structure (i.e. the relationship between y and x variables at some previous time at domain level) and the allocation structure (i.e. the current relationship at the larger area level), preserves the earlier relationship present in the association structure without interfering with current information in the allocation structure. In the Canadian context, for characteristics for which large scale surveys (such as the Labour Force Survey) are undertaken regularly, it would seem the short term demand for data for domains of the size of FED's or Census Divisions may be met through the use of refined estimation techniques (and pooling of estimates over a period of time) utilizing census and servey data alone. However, for meeting such demands in the longer term and for other types of data based on smaller surveys and other types and sizes of domains, all three sources of data namely census, surveys and administrative files would have to be fully explored. Multi-variate linear regression estimators of the type considered by Ericksen (1974) and Gonzalez and Hoza (1978) using data from all three sources should be studied in detail for their bias, mean square error and the computational complexities. Each of the three sources, with limitations of their own, when put together offer considerable potential for improvements in the sense that the weaknesses of one source can be the strengths of another. Hence there is reason for optimism that statistically sound techniques exploiting the strengths of data from different sources in an integrated fashion hold the future key to good quality small area data for a large variety of subject matters.

## ACKNOWLEDGEMENT

Discussions with Mr. R. Platek have been beneficial in the finalization of this paper.

Table 1. Efficiencies of Small Area Estimators Relative to Direct Estimator
- Nova Scotia Census Divisions (Auxiliary data up-to-date).

|  |  |  | ESTIMATOR |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  | Sample Dependent |  |
| Characteristic | Auxiliary Variable | Level of Construction | Post-Stratified Domain | Synthetic | Composite ( $\alpha^*_2$=0.223) | $K_o$=0.5 | $K_o$=1.0 |
| Employed | Dwelling | combined | 4.58 | 10.17 | 10.92 | 9.17 | 10.42 |
|  | Population | " | 4.92 | 10.75 | 10.58 | 10.50 | 11.67 |
|  | Population by age/sex | " | 5.08 | 10.83 | 11.25 | 11.17 | 12.25 |
|  | Dwelling | separate | 2.75 | 10.50 | - | 9.58 | 10.50 |
|  | Population | " | 2.83 | 10.92 | - | 10.58 | 11.42 |
|  | Population by age/sex | " | 2.83 | 11.00 | - | 11.00 | 11.75 |
| Unemployed | Dwelling | combined | 1.33 | 1.70 | 1.75 | 1.40 | 1.55 |
|  | Population | " | 1.36 | 1.70 | 1.75 | 1.43 | 1.58 |
|  | Population by age/sex | " | 1.36 | 1.70 | 1.75 | 1.43 | 1.58 |
|  | Dwelling | separate | 1.30 | 1.69 | - | 1.48 | 1.58 |
|  | Population | " | 1.33 | 1.69 | - | 1.51 | 1.61 |
|  | Population by age/sex | " | 1.33 | 1.69 | - | 1.51 | 1.61 |

Table 2. Efficiencies of Small Area Estimators Relative to Direct Estimator - Nova Scotia Census Divisions (Auxiliary data out-of-date)

| | | | ESTIMATOR | | |
|---|---|---|---|---|---|
| Characteristic | Auxiliary Variable | Level of Construction | Post-Stratified Domain | Synthetic | Sample Dependent<br>$(K_0=1.0)$ |
| Employed | population | combined | 3.47 | 4.73 | 4.07 |
| | population by age/sex | " | 3.60 | 4.73 | 4.23 |
| Unemployed | population | combined | 1.44 | 2.19 | 1.68 |
| | population by age/sex | " | 1.46 | 2.21 | 1.69 |

Table 3. % Average Relative Absolute Bias of Synthetic Estimators for FED's (using out-of-date, 15+ population as auxiliary information

| Characteristic | Employed | | Unemployed | |
|---|---|---|---|---|
| Level of Construction | Separate | Combined | Separate | Combined |
| Province | | | | |
| Newfoundland | 1.20 | 2.19 | 1.75 | 3.17 |
| Prince Edward Island | 3.54 | 5.62 | 3.71 | 2.80 |
| Nova Scotia | 0.87 | 1.64 | 1.25 | 1.87 |
| New Brunswick | 2.95 | 2.54 | 6.35 | 7.04 |
| Quebec | 2.53 | 3.87 | 3.87 | 4.51 |
| Ontario | 2.17 | 3.52 | 3.01 | 4.25 |
| Manitoba | 1.42 | 2.28 | 2.22 | 3.44 |
| Saskatchewan | 2.41 | 2.65 | 4.35 | 4.85 |
| Alberta | 5.24 | 7.08 | 5.12 | 10.33 |
| British Columbia | 1.73 | 2.71 | 2.72 | 4.20 |

Table 4. FED's with Biases of Separate Synthetic Estimator for Unemployed Exceeding 10%

| | Category (i) | | | Category (ii) | | | Category (iii) | |
|---|---|---|---|---|---|---|---|---|
| | % rel bias | | | % rel bias | | | % rel Bias | |
| FED | Up-to-date | Out-of-date | FED | Up-to-date | Out-of-date | FED | Up-to-date | Out-of-date |
| 102 | 12.25 | -3.90 | 414 | 1.57 | 17.78 | 301 | 12.71 | 25.85 |
| 104 | -12.52 | 6.23 | 426 | -7.38 | -11.78 | 304 | -11.29 | -17.57 |
| 411 | -13.10 | -0.37 | 450 | 1.93 | -11.35 | 412 | -10.07 | -15.80 |
| 436 | 10.48 | -0.48 | 455 | 2.67 | 15.46 | 438 | 12.15 | 14.22 |
| 474 | 18.35 | -6.04 | 460 | -7.74 | 20.80 | 501 | 10.52 | 16.83 |
| 806 | 15.15 | -0.20 | 504 | 7.19 | 41.05 | 579 | 15.50 | 10.59 |
| | | | 527 | 9.59 | 11.76 | 818 | 10.83 | 39.41 |
| | | | 605 | 3.63 | 14.29 | | | |
| | | | 701 | 4.90 | 17.74 | | | |
| | | | 804 | 0.78 | 15.85 | | | |
| | | | 813 | -0.48 | -15.46 | | | |

Category (i): bias exceeds 10% only when auxiliary information is up-to-date.

(ii): bias exceeds 10% only when auxiliary information is out-of-date.

(iii): bias exceeds 10% in both the cases.

Table 5. Average Reliance of Separate Sample Dependent Estimator on Synthetic Component: Nova Scotia Census Divisions.

| | Reliance $(1-\delta)$ on Synthetic Component | | Proportion of Census Division Population | |
|---|---|---|---|---|
| Census Division | $K_o$=0.5 | $K_o$=1.0 | Partial Strata | Complete Strata |
| 201 | .04 | .15 | 1.00 | - |
| 202 | .15 | .20 | .70 | .30 |
| 203 | .01 | .12 | 1.00 | - |
| 204 | .12 | .22 | 1.00 | - |
| 205 | .04 | .14 | 1.00 | - |
| 206 | .03 | .08 | .37 | .63 |
| 207 | .05 | .07 | .26 | .74 |
| 210 | .06 | .09 | .35 | .65 |
| 211 | .04 | .05 | .13 | .87 |
| 212 | .06 | .10 | .52 | .48 |
| 213 | .03 | .16 | 1.00 | - |
| 214 | .04 | .16 | 1.00 | - |
| 215 | .11 | .21 | 1.00 | - |
| 216 | .05 | .15 | 1.00 | - |
| 217 | .01 | .01 | .03 | .97 |
| 218 | .18 | .28 | 1.00 | - |

## REFERENCES

[1] Drew, J.D. and Choudhry, G.H. (1979), "Small Area Estimation", Technical Report, Census and Household Surveys Methods Division, Statistics Canada.

[2] Ghangurde, P.D. and Singh, M.P. (1976), "Synthetic estimation in the LFS", Technical Report, Household Surveys Development Division, Statistics Canada.

[3] Ghangurde, P.D. and Singh, M.P. (1977), "Synthetic Estimates in Periodic Household Surveys", Survey Methodology, Vol. 3, No. 1, 152-181, Statistics Canada.

[4] Ghangurde, P.D. and Singh, M.P. (1978), "Evaluation of Efficiency of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 53-61.

[5] Gonzalez, M.E. (1973), "Use and Evaluation of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 33-36.

[6] Gonzalez, M.E. and Waksberg, J. (1973), "Estimation of the Error of Synthetic Estimates", paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

[7] Gonzalez, M.E. (1975), "Small Area Estimation of Unemployment", Proceedings of the American Statistical Association, Social Statistics Section, 437-460.

[8] Gonzalez, M.E. and Hoza, C. (1978), "Small Area Estimation with Application to Unemployment and Housing Estimates", Journal of the American Statistical Association 73, 7-15.

[9] Holt, T., Smith, T.M.F. and Tomberlin, T.J. (1979), "A Model Based Approach to Estimation for Small Sub-groups of a Population", Journal of

[10] National Center for Health Statistics (1968), "Synthetic State Estimates of Disability", P.H.S. Publication No. 1759, U.S. Government Printing Office, Washington, D.C.

[11] Platek, R. and Singh, M.P. (1976), "Methodology of the Canadian Labour Force Survey," Catalogue No. 71-526, Statistics Canada.

[12] Purcell, N.J. and Linacre, S. (1976), "Techniques for the Estimation of Small Area Characteristics", paper presented at the 3rd Australian Statistical Conference, Melbourne, Australia.

[13] Purcell, N.J. and Kish, L. (1979), "Estimation for Small Domains", Biometrics 35, 365-384.

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

[15] Royall, R.M. (1978), "Prediction models in Small Area Estimation", NIDA Workshop on Synthetic Estimates, Princeton, N.J.

[16] Sarndal, C.E. (1981), "When Robust Estimation is not an obvious answer: The case of the Synthetic Estimator versus Alternatives for Small Areas", Proceedings of the American Statistical Association, Survey Research Section.

[17] 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, 1017-1021.

[18] Schaible, W.L. (1978), "Choosing Weights for Composite Estimators for Small Area Statistics", Proceedings of the American Statistical Association, Survey Research Section, 741-746.

[19] Schaible, W.L. (1979), "A Composite Estimator for Small Area Statistics", in Synthetic Estimates for Small Areas (J. Steinberg, Ed.) National Institute on Drug Abuse Research Monograph No. 24, U.S. Government Printing Office, Washington, D.C., 36-53.

[20] Singh, M.P. and Tessier, R. (1975), "Some Estimators for Domain Totals", Journal of The American Statistical Association 71, 322-325.

[21] Sonquist, J.N. and Morgan J.A. (1964), "The Detection of Interaction Effects", Monograph no. 35, Survey Research Center, Institute for Social Research, University of Michigan.