Survey Methodology, June 1985
Vol. 11, No. 1, pp. 65–77
Statistics Canada



# An Empirical Study of Some Regression Estimators for Small Domains

M.A. HIDIROGLOU and C.E. SÄRNDAL[1]

ABSTRACT

The synthetic estimator (SYN) has been traditionally used to estimate characteristics of small domains. Although it has the advantage of a small variance, it can be seriously biased in some small domains which depart in structure from the overall domains. Särndal (1981) introduced the regression estimator (REG) in the context of domain estimation. This estimator is nearly unbiased, however, it has two drawbacks; (i) its variance can be considerable in some small domains and (ii) it can take on negative values in situations that do not allow such values.

In this paper, we report on a compromise estimator which strikes a balance between the two estimators SYN and REG. This estimator, called the modified regression estimator (MRE), has the advantage of a considerably reduced variance compared to the REG estimator and has a smaller Mean Squared Error than the SYN estimator in domains where the latter is badly biased. The MRE estimator eliminates the drawback with negative values mentioned above. These results are supported by a Monte Carlo study involving 500 samples.

KEY WORDS: Small domains; regression estimation; modified regression estimator; bias; mean squared error.

## 1. INTRODUCTION

The synthetic estimator (SYN) has the advantage of a small variance, but the following disadvantages: (a) it can be badly biased in some domains, and ordinarily we do not know which ones; (b) consequently, a calculated coefficient of variation (cv), or a calculated confidence interval, is meaningless for such domains.

For the same model that underlies the SYN estimator one can create a nearly unbiased analogue, the generalized regression estimator (REG), which has the additional advantage that a standard design based confidence interval is easily computed for each domain estimate. A disadvantage with REG is that the estimated variance (and hence the cv and the width of the confidence interval) can be unacceptably large in very small domains. (This is, of course, a direct consequence of the shortage of observations in such domains.) Also, the REG can (although with small probability) take negative values in situations where such values are unacceptable.

It is therefore desirable to strike a balance between SYN and REG. Here, we report on an empirical study with one such compromise estimator, the modified regression estimator (MRE). It has a small (but noticeable) bias in those domains where the synthetic estimator is greatly biased; in other domains, the MRE is nearly unbiased. The MRE has the advantage of a considerably reduced variance compared to the REG estimator. In addition, the MRE has a smaller Mean Squared Error than the SYN estimator in domains where the latter is badly biased. Meaningful confidence intervals can also be easily constructed for the new MRE estimator.

[1] M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada, 5-C8, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6 and C.E. Särndal, Department of Mathematics and Statistics, University of Montréal, Montréal, Québec, Canada H3C 3J7.

The paper is structured as follows. In Section 2, some of the commonly used estimators for small areas such as the direct, post-stratified and synthetic estimators are reviewed as well as some of the regression estimators given by Särndal (1981, 1984). In Section 3, the proposed modified regression estimators are introduced and discussed. In Section 4, the properties of the modified regression estimators as well as some of the other estimators are studied through a Monte Carlo simulation using business tax data. Finally, Section 5 provides some general conclusions.

## 2. ESTIMATORS

Let the population $U = \{1, \ldots, k, \ldots, N\}$ be divided into $D$ non-overlapping domains $U_{1.}, \ldots, U_{d.}, \ldots, U_{D.}$. Let $N_{d.}$ be the size of $U_{d.}$. (In our empirical study, the domains are defined by a cross-classification of 4 industrial groupings with the 18 census divisions in the province of Nova Scotia. There were $D = 70$ non-empty domains, as described in Hidiroglou, Morry, Dagum, Rao and Särndal (1984).)

The population is further divided along a second dimension, into $G$ non-overlapping groups, $U_{.1}, \ldots, U_{.g}, \ldots, U_{.G}$.

The size of $U_{.g}$ is denoted $N_{.g}$. (In our study, the groups are based on Gross Business Income classes.) The cross-classification of domains and groups gives rise to $DG$ population cells $U_{dg}$; $d = 1, \ldots, D$; $g = 1, \ldots, G$. Let $N_{dg}$ be the size of $U_{dg}$.

Then the population size $N$ can be expressed as

$$N = \sum_{d=1}^{D} N_{d.} = \sum_{g=1}^{G} N_{.g} = \sum_{d=1}^{D} \sum_{g=1}^{G} N_{dg} \tag{2.1}$$

Let $s$ denote a sample of size $n$ drawn from $U$ by simple random sampling (srs). Denote by $s_{d.}$, $s_{.g}$ and $s_{dg}$ the parts of $s$ that happen to fall, respectively, in $U_{d.}$, $U_{.g}$ and $U_{dg}$.

The corresponding sizes, which are random variables, are denoted by $n_{d.}$, $n_{.g}$ and $n_{dg}$. Note that (2.1) holds for lower case $n$'s as well. The variable of interest, $y$ (= Wages and Salaries) takes the value of $y_k$ for the $k$:th unit (= unincorporated business tax filer). The auxiliary variable $x$ (= Gross Business Income) takes the value $x_k$ for the $k$:th unit, and $x_k$ is known for all $k = 1, \ldots, N$.

The following estimators of the domain total $t_d = \sum_{U_{d.}} y_k$ are compared, where $\sum_{U_{d.}}$ denotes the summation over the units in $U_{d.}$.

**The straight expansion estimator (EXP):**

$$\hat{t}_{d\text{EXP}} = \frac{N}{n} \sum_{s_{d.}} y_k \tag{2.2}$$

**The poststratified estimator (POS):**

$$\hat{t}_{d\text{POS}} = N_{d.} \bar{y}_{s_{d.}} \tag{2.3}$$

where

$$\bar{y}_{s_{d.}} = \sum_{s_{d.}} \frac{y_k}{n_{d.}}$$

is the mean of the $n_{d.}$ $y$-values from the $d$:th domain. If $n_{d.} = 0$ we define the POS estimator to be zero (somewhat arbitrarily, since strictly speaking the estimator is then undefined). Neither the EXP nor the POS estimator are particularly advantageous. They serve mainly as benchmarks against which the behaviour of the following more efficient estimators will be compared.

Two versions of the SYN and REG have been investigated, the "Count" version and the "Ratio" version. The SYN estimator is based on the assumption that a given model holds for each group $g$. For the "Count" version a given model would lead to the assumption that the mean of each group is the same across all domains $d$. For the "Ratio" version, the implied model would be that the ratios of a given variable of interest over an auxiliary variable would be constant within a given group across all domains. If the assumption of homogeneity of domain characteristics does not hold within each group, the SYN estimators can be very biased. The REG estimation method as given by Särndal (1984) is motivated by the following requirements: (a) to obtain approximately design-unbiased estimates with simple variance estimates and easily calculable (and meaningful) confidence intervals; (b) to strengthen the estimates by involving sample data from all domains.

The formulas for the "Count" versions are:

**Synthetic-Count estimator (SYN/C):**

$$\hat{t}_{d\text{SYN/C}} = \sum_{g=1}^{G} N_{dg}\bar{y}_{s_{.g}} \tag{2.4}$$

where $\bar{y}_{s_{.g}}$ is the mean of $y$ in $s_{.g}$.

**Regression-Count estimator (REG/C):**

$$\hat{t}_{d\text{REG/C}} = \sum_{g=1}^{G} \{N_{dg}\bar{y}_{s_{.g}} + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})\} \tag{2.5}$$

where $\bar{y}_{s_{dg}}$ is the mean of $y$ in $s_{dg}$, and $\hat{N}_{dg} = Nn_{dg}/n$. Here, $\sum_{g=1}^{G} \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})$ is a bias correction term that ordinarily carries a considerable variance contribution.

The "Ratio" versions of the SYN and REG estimators are:

**Synthetic-Ratio estimator (SYN/R):**

$$\hat{t}_{d\text{SYN/R}} = \sum_{g=1}^{G} X_{dg}\hat{R}_g \tag{2.6}$$

with $X_{dg} = \sum_{U_{dg}} x_k$ and

$$\hat{R}_g = \frac{\sum_{s_{.g}} y_k}{\sum_{s_{.g}} x_k}$$

**Regression - Ratio estimator (REG/R):**

$$\hat{t}_{d\text{REG/R}} = \sum_{g=1}^{G} \{X_{dg}\hat{R}_g + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \hat{R}_g\bar{x}_{s_{dg}})\} \tag{2.7}$$

## 3. MODIFIED REGRESSION ESTIMATORS

Regression estimators introduced by Särndal (1984) were constructed by fitting a regression model to some auxiliary variables and using the resulting fitted model to create predicted values for the units in the population domain. Assuming that the sampling design, $p$, is an arbitrary one (not necessarily srs) with inclusion probabilities $\pi_k$ (first order) and $\pi_{kl}$ (second order), let the regression model be given by

$$E_\xi(y_k) = \underset{\sim}{x}_k'\underset{\sim}{\beta};\ V_\xi(y_k) = v_k$$

where the $y_k$ are independent random variables. An estimator of $\underset{\sim}{\beta}$ is

$$\hat{\underset{\sim}{\beta}} = \left(\sum_{s} \frac{\underset{\sim}{x}_k \underset{\sim}{x}_k'}{v_k \pi_k}\right)^{-1} \sum_{s} \frac{\underset{\sim}{x}_k y_k}{v_k \pi_k}$$

where it is assumed that the $v_k$ are known to multiplicative constant(s) that cancel when $\hat{\underset{\sim}{\beta}}$ is derived.

Following Särndal (1984), a nearly unbiased estimator of the unknown $d$-th domain total is given by

$$\hat{t}_{d\text{REG}} = \sum_{U_d.} \hat{y}_k + \sum_{s_d.} \frac{e_k}{\pi_k} \tag{3.1}$$

where $\hat{y}_k = \underset{\sim}{x}_k' \hat{\underset{\sim}{\beta}}$ is the $k$-th predicted value and $e_k = y_k - \hat{y}_k$ denotes the $k$-th residual.

We shall refer to $\sum_{U_d.} \hat{y}_k$ as *the synthetic term* of the estimator $\hat{t}_{d\text{REG}}$ and the second term, $\sum_{s_d.} e_k/\pi_k$, will be called the *correction* term.

If $s_d.$ is non-empty, an approximately unbiased alternative to the REG estimator (3.1) is given by

$$\hat{t}_{d\text{ALT}} = \sum_{U_d.} \hat{y}_k + N_d. \frac{\sum_{s_d.} \frac{e_k}{\pi_k}}{\hat{N}_d.} \tag{3.2}$$

where

$$\hat{N}_d. = \sum_{s_d.} \frac{1}{\pi_k}$$

is the estimated domain size.

The correction term now appears in the form of a ratio estimator,

$$\frac{\sum_{s_d.} \frac{e_k}{\pi_k}}{\sum_{s_d.} \frac{1}{\pi_k}},$$

multiplied by the known domain size $N_d.$ (obviously, $N_d.$ is known since the cell counts $N_{dg}$ are known).

The size $n_d.$ being random, the ratio form will serve to reduce the variance of the correction term. The effect will be particularly noticeable in domains where the average of the residuals is clearly away from zero (that is, in domains where the model does not fit well).

If the expected sample take in the domain, $E_d = E_p(n_d.) = \sum_{U_d} \pi_k$, were substantial (say, $E_d \geq 50$), then it is practically certain that the realized sample take, $n_d.$, will not be exceedingly small. For example, under srs, values $n_d. \leq 30$ will hardly ever occur. In such situations, the nearly unbiased estimator (3.2) can be recommended as is. It should realize important efficiency gains over (3.1), notably in domains where the model does not fit as well. But in practice one often encounters domains that are so small that the expected sample take $E_d$ does not exceed 5. This is true for a number of domains in our study. In such cases, realized sample takes $n_d.$ between zero and five are very likely. Our empirical work has confirmed the intuitively obvious fact that the residual correction will, in these small domains, contribute greatly to the variance, whether the correction appears in its straight form, $\sum_{s_d.} e_k/\pi_k$, as in (3.1), or in its ratio form , $N_d.(\sum_{s_d.} e_k/\pi_k)/(\sum_{s_d.} 1/\pi_k)$, as in (3.2).

To counteract this inflated variance contribution, we modify the correction term of (3.2) in a way implying that we settle for a small bias (in domains where the model fits less well) in exchange for a reduced variance contribution when the realized sample take $n_d.$ is lower than expected (and it is assumed that the expected sample take is already low in itself).

The form of the new correction term will be determined by the relation between realized sample take $n_{d.}$, and expected sample take $E_d$. The correction term $\sum_{s_{d.}} e_k/\pi_k$ will be multiplied by $(\hat{N}_d/N_d)$ when $n_{d.} < E_d$ and by $(N_d/\hat{N}_d)$ otherwise. The resulting correction term using this adaptive "dampening factor" will have the effect of not "over-correcting" the synthetic term when some of the residuals $e_k$ behave as outliers for small $n_{d.}$'s. The "over-correcting" may have the effect of greatly underestimating a domain $d$, yielding negative values when only positive values are acceptable, or conversely greatly overestimating the domain.

The resulting estimator, the modified regression estimator (MRE), incorporating these two types of realizations of $n_{d.}$, is

$$\hat{t}_{d\text{MRE}} = \sum_{U_{d.}} \hat{y}_k + F_d \sum_{s_{d.}} \frac{e_k}{\pi_k} \tag{3.3}$$

where

$$F_d = \begin{cases} \dfrac{N_{d.}}{\hat{N}_{d.}} \text{ when } n_{d.} \geq E_d \\ \dfrac{\hat{N}_{d.}}{N_{d.}} \text{ when } n_{d.} < E_d \end{cases}$$

It can be shown that (3.3) is nearly unbiased conditionally on $n_{d.}$, as long as $n_{d.} \geq E_d$. For $n_{d.} < E_d$, the MRE has some conditional bias, which tends to increase the more $n_{d.}$ falls short of its expected value. At the same time, the MRE estimator is being pushed towards its synthetic term, thus benefitting from the stability (low variance) of the synthetic term. Unconditionally, the MRE estimator given by (3.3) will have a certain small bias, but a much reduced variance compared with the REG estimator.

We note a final point in favour of MRE estimator. As a result of its considerable variance in very small domains, the REG estimator will, with a small but positive probability, take values extremely removed from the true value $t_{d.}$. The value of the REG may even be negative, which is, of course, unacceptable for a variable (such as Wages and Salaries) which is by definition non-negative. Negative values of the REG estimate can occur when there exists large negative residuals $e_k$ in the correction term of (3.1), and are especially likely when $n_{d.} < E_d$. The new MRE estimator virtually eliminates this occurence of negative estimates. In practice, if by a remote possibility the MRE takes a negative value, we recommend to redefine the MRE estimator as being equal to the always positive SYN estimator.

A natural formula for estimating the variance of (3.2) is

$$\hat{V}_p(\hat{t}_{d\text{ALT}}) = \left(\frac{N_{d.}}{\hat{N}_{d.}}\right)^2 \sum_{\substack{k \neq \ell \\ \epsilon\, s_{d.}}} \sum \Delta_{k\ell} \frac{(e_k - \bar{e}_{s_{d.}})(e_\ell - \bar{e}_{s_{d.}})}{\pi_k \pi_\ell} \tag{3.4}$$

where

$$\bar{e}_{s_{d.}} = \frac{\sum_{s_{d.}} \frac{e_k}{\pi_k}}{\sum_{s_{d.}} \frac{1}{\pi_k}}$$

and

$$\Delta_{k\ell} = \begin{cases} 1 - \pi_k & \text{if } \ell = k \\ 1 - \dfrac{\pi_k \pi_\ell}{\pi_{k\ell}} & \text{if } \ell \neq k. \end{cases}$$

We propose that the same formula may serve well to estimate the variance of the MRE estimator (3.3). It is true that (3.3) differs from (3.2) when the realized sample take falls short of the expected sample take; however, it is not foreseen that the difference will be great enough to cause serious distortion in the validity of a confidence interval for $t_d$ centred on $\hat{t}_{d\text{MRE}}$ using (3.4) as the estimated variance.

In the case of simple random sampling, and assuming for $g = 1, \ldots, G$,

$$E_\xi(y_k) = \beta_g;\ V_\xi(y_k) = \sigma_g^2;\ k \in U_{.g}, \tag{3.5}$$

we find

$$\hat{\beta}_g = \frac{\sum_{s.g} y_k}{n_{.g}} = \bar{y}_{s.g},$$

leading to the "Count estimator" whose modified version (MRE/C) is

$$\hat{t}_{d\text{MRE/C}} = \sum_{g=1}^{G} \{N_{dg}\bar{y}_{s_g.} + F_d\hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s.g})\} \tag{3.6}$$

where $E_d$ in the formula for $F_d$ is now given by

$$E_d = E_{\text{srs}}(n_{d.}) = \frac{nN_{d.}}{N}$$

with

$$\hat{N}_{dg} = n_{dg}\left(\frac{N}{n}\right)$$

and

$$\bar{y}_{s_{dg}} = \begin{cases} \dfrac{\sum_{s_{dg}} y_k}{n_{dg}} & \text{for } n_{dg} \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The MRE/C estimator will have some bias, which is, however, ordinarily much less than that of the SYN/C estimator.

The underlying model assumptions which lead to the "ratio estimator", whose modified version is denoted as MRE/R, are for $g = 1, \ldots, G$,

$$E_\xi(y_k) = \beta_g x_k;\ V_\xi(y_k) = \sigma_g^2 x_k,\ k \in U_{.g}.$$

The MRE/R estimator is then, in the case of simple random sampling,

$$\hat{t}_{d\text{MRE/R}} = \sum_{g=1}^{G} \{X_{dg}\hat{R}_g + F_d\hat{N}_{dg}(\bar{y}_{s_{dg}} - \hat{R}_g\bar{x}_{s_{dg}})\} \tag{3.7}$$

where

$$\hat{R}_g = \frac{\sum_{d=1}^{D} \hat{N}_{dg}\bar{y}_{s_{dg}}}{\sum_{d=1}^{D} \hat{N}_{dg}\bar{x}_{s_{dg}}},$$

and

$$X_{dg} = \sum_{U_{dg}} x_k.$$

Drew, Singh and Choudhry (1982) provided small domain estimators which, although not derived by a regression approach, have some similarity to the ones given in this paper. Their "count" version is

$$\hat{t}_{d\text{KNO/C}} = \sum_{g} N_{dg}\{W'_{dg}\bar{y}_{s_{dg}} + (1 - W'_{dg})\bar{y}_{s_{.g}}\} \tag{3.8}$$

while their "ratio" version is

$$\hat{t}_{d\text{KNO/R}} = \sum_{s_g} X_{dg}\left\{W'_{dg}\frac{\bar{y}_{s_{dg}}}{\bar{x}_{s_{dg}}} + (1 - W'_{dg})\frac{\bar{y}_{s_{.g}}}{\bar{x}_{s_{.g}}}\right\} \tag{3.9}$$

where

$$W'_{dg} = \begin{cases} \dfrac{n_{dg}}{E_{dg}} \text{ if } n_{dg} \leq E_{dg} \\ 1 \quad \text{otherwise} \end{cases}$$

with $E_{dg} = n(N_{dg}/N)$. In the present context, if $W'_{dg}$ in (3.8) is replaced by

$$W''_{dg} = \begin{cases} \left(\dfrac{n_{d.}}{E_d}\right)\left(\dfrac{n_{dg}}{E_{dg}}\right) & \text{if } n_{d.} < E_d \\ \left(\dfrac{E_d}{n_{d.}}\right)\left(\dfrac{n_{dg}}{E_{dg}}\right) & \text{if } n_{d.} \geq E_d \end{cases}$$

we obtain $\hat{t}_{d\text{MRE/C}}$.

## 4. RESULTS FROM THE EMPIRICAL STUDY

In order to study the properties of the estimators discussed in the preceding sections, a simulation was undertaken. The province of Nova Scotia was chosen as our population with $N = 1678$ sampling units (unincorporated tax filers). The variable of interest, $y$, is Wages and Salaries. We use a single auxiliary variable, $x$, namely, Gross Business Income. It is assumed that $x_1, \ldots, x_N$ are known.

Domains of the population were formed by a cross-classification of four industrial groups with eighteen regions. The industrial groups were Retail (515 units), Construction (496 units), Accommodation (114 units) and Others (553 units). The overall correlation coefficients between Wages and Salaries and Gross Business Income were 0.42 for Retail, 0.64 for Construction, 0.78 for Accommodation and 0.61 for Others. The regions were the 18 Census Divisions of the province. This produced 70 non-empty domains (out of the four times 18 domains, two combinations had no units). Thus, 70 domain totals $t_d$ are to be estimated every time a sample is drawn.

For the Monte Carlo simulation, 500 simple random samples, $s$, each of size $n = 419$, were selected from the population of $N = 1678$ units. The selected sample units were classified into type of industry and Census Division. The population could have been divided along a second dimension, say income groups. But for the purposes of this study, all the taxfilers were considered as belonging to one income group ($G = 1$).

The results are summarized for each small area within the industrial groups RETAIL and ACCOMMODATION using tables and graphs. For the tables (1-4), summary statistics are the relative conditional bias and mean squared error. The eight graphs, one for each of the eight estimators, are given in figure 1. In each graph, there are eighteen vertical 'distribution bands', one for each of the eighteen Census Divisions for the industrial group RETAIL. The upper and lower points of each distribution band correspond, respectively, to the 90:th and 10:th percentile of the distribution of the 500 values of $(\hat{t}_{d.} - t_{d.})/t_{d.}$. Consequently, a distribution band placed roughly symmetrically about the zero line indicates that the corresponding estimator is approximately unbiased for the domain of interest; otherwise, the estimator is biased for the domain. The shorter the band, the smaller the variance of the estimator in the domain. The abscissa measures the mean sample take for the domain.

From the tables and graphs, the following conclusions emerge: (where conclusion C states the main new results, whereas A and B resume what is known from earlier work Särndal and Råbäck (1983); Hidiroglou et al. (1984)).

A. The SYN/C and SYN/R estimators are badly biased in some domains, namely, in those domains where the underlying model fits poorly. However, they consistently have an attractively low variance, compared to the other alternatives. The Mean Squared Error of the two SYN estimators will consequently be very large in domains with large bias (poor model fit); by contrast, the Mean Squared Error is small in domains with little bias (good model fit).

B. The REG/C and REG/R estimators are essentially unbiased. Their variance, although usually much lower than that of the EXP and POS estimators, is consistently much higher than that of the SYN/C and SYN/R estimators. In the smallest domains, none of the unbiased estimators (EXP, POS, REG/C, REG/R) is attractive from the variance point of view; this is especially true for the REG estimators. This problem is remedied by the two MRE modifications of the REG estimators.

C. The two MRE estimators, MRE/C and MRE/R, are negligibly biased when the SYN estimators happen to be nearly unbiased (e.g., RETAIL, area 17); otherwise the MRE estimators have a certain bias, which, however, is ordinarily much less pronounced than that of the SYN estimators (e.g., RETAIL, area 2). The MRE estimators have considerably smaller variance and Mean Squared Error, in all domains, than the REG estimators. This tendency is particularly pronounced in the smaller domains. In comparison with the SYN estimators, we find that the MRE estimators (as expected) still have a larger variance in virtually all domains. However, the Mean Squared Error of the MRE estimators is smaller than that of the SYN estimators in domains where the latter are badly biased. In Table 6 we see, for example, that the MRE/R estimator has a smaller Mean Squared Error than that of the SYN/R in 9 out of 16 small areas. The obvious explanation is that in domains where the SYN estimator is greatly biased, the $(\text{bias})^2$ constitutes an extremely large contribution to the Mean Squared Error of the SYN, whereas for the MRE estimators, the $(\text{bias})^2$ is not very important. Since we do not know which domains create the large biases, the goal of producing reliable estimates in all domains is on the whole better served by the MRE method of estimation.

**Table 1**

Mean Sample Take and Relative Bias of Each of Eight Estimators over 500 Repeated Simple Random Samples from the Entire Population Industrial Group: RETAIL; 18 Census Divisions in Nova Scotia.

| Area | Mean Sample Take | Estimator EXP | POS | SYN/C | MRE/C | REG/C | SYN/R | MRE/R | REG/R |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.76 | −0.02 | −0.13 | 0.12 | 0.02 | −0.03 | 0.30 | 0.09 | −0.02 |
| 2 | 5.45 | 0.00 | −0.04 | −0.36 | −0.10 | −0.02 | −0.27 | −0.08 | −0.02 |
| 3 | 3.90 | −0.02 | 0.01 | −0.08 | −0.02 | 0.00 | −0.01 | −0.01 | 0.00 |
| 4 | 3.02 | 0.01 | −0.05 | 0.15 | 0.05 | 0.01 | 0.13 | 0.04 | 0.04 |
| 5 | 5.93 | 0.00 | 0.01 | 0.21 | 0.05 | 0.00 | 0.13 | 0.03 | 0.00 |
| 6 | 7.63 | −0.02 | −0.01 | 0.28 | 0.07 | 0.01 | 0.10 | 0.02 | 0.00 |
| 7 | 8.61 | 0.02 | 0.01 | −0.16 | −0.03 | 0.01 | −0.18 | −0.03 | 0.01 |
| 8 | 5.64 | −0.02 | −0.01 | 0.34 | 0.10 | 0.03 | 0.24 | 0.06 | 0.01 |
| 9 | 24.64 | 0.00 | 0.00 | −0.02 | 0.00 | 0.00 | −0.01 | 0.00 | 0.01 |
| 10 | 8.92 | −0.02 | −0.02 | 0.15 | 0.02 | −0.01 | 0.09 | 0.00 | −0.01 |
| 11 | 8.35 | −0.03 | −0.02 | 0.08 | 0.01 | 0.00 | 0.10 | 0.02 | 0.00 |
| 12 | 10.58 | 0.01 | 0.00 | −0.27 | −0.05 | 0.00 | −0.18 | −0.03 | 0.00 |
| 13 | 0.48 | −0.04 | −0.58 | 0.61 | 0.36 | 0.04 | 1.00 | 0.58 | 0.04 |
| 14 | 2.80 | 0.03 | −0.03 | 0.33 | 0.11 | 0.00 | 0.24 | 0.10 | 0.02 |
| 15 | 4.21 | 0.06 | −0.01 | 0.28 | 0.06 | 0.00 | 0.30 | 0.07 | −0.01 |
| 16 | 2.24 | 0.03 | −0.05 | 0.74 | 0.26 | 0.03 | 0.94 | 0.32 | 0.02 |
| 17 | 23.95 | −0.01 | −0.01 | −0.02 | 0.00 | 0.00 | −0.05 | −0.01 | 0.00 |
| 18 | 0.54 | 0.07 | −0.54 | 0.63 | 0.34 | −0.06 | 0.67 | 0.35 | −0.06 |

**Table 2**

Mean Squared Error of Each of Eight Estimators over 500 Repeated Simple Random Samples from the Entire Population Industrial Group: RETAIL; 18 Census Divisions in Nova Scotia.

| Area | Estimator EXP | POS | SYN/C | MRE/C | REG/C | SYN/R | MRE/R | REG/R |
|---|---|---|---|---|---|---|---|---|
| 1 | 3,209 | 2,206 | 96 | 697 | 1,397 | 462 | 769 | 1,484 |
| 2 | 42,598 | 24,623 | 21,782 | 12,725 | 17,358 | 13,110 | 10,256 | 14,380 |
| 3 | 10,469 | 6,853 | 357 | 2,592 | 4,212 | 146 | 2,333 | 3,782 |
| 4 | 5,626 | 3,657 | 324 | 746 | 1,186 | 257 | 1,206 | 1,853 |
| 5 | 14,554 | 9,681 | 2,999 | 5,090 | 7,360 | 1,294 | 3,993 | 5,974 |
| 6 | 12,308 | 5,686 | 6,713 | 3,423 | 4,289 | 1,255 | 1,747 | 2,515 |
| 7 | 34,865 | 17,988 | 6,912 | 9,387 | 13,451 | 8,161 | 12,019 | 17,239 |
| 8 | 12,066 | 8,630 | 5,772 | 3,694 | 5,045 | 2,981 | 3,528 | 4,986 |
| 9 | 72,974 | 40,440 | 5,776 | 24,025 | 29,250 | 5,068 | 21,292 | 25,832 |
| 10 | 22,091 | 9,433 | 4,559 | 5,832 | 7,927 | 2,009 | 5,365 | 7,272 |
| 11 | 23,519 | 12,505 | 1,778 | 6,738 | 9,578 | 2,348 | 7,890 | 11,063 |
| 12 | 46,588 | 21,874 | 35,310 | 13,558 | 17,084 | 17,454 | 12,222 | 16,514 |
| 13 | 635 | 244 | 161 | 95 | 228 | 422 | 287 | 783 |
| 14 | 3,871 | 2,849 | 692 | 1,254 | 2,141 | 378 | 1,373 | 2,346 |
| 15 | 8,088 | 3,511 | 2,249 | 1,892 | 2,806 | 2,651 | 1,985 | 2,937 |
| 16 | 3,245 | 2,127 | 3,316 | 1,563 | 2,516 | 5,333 | 1,741 | 2,654 |
| 17 | 81,211 | 47,753 | 5,503 | 28,957 | 35,232 | 7,681 | 27,457 | 33,136 |
| 18 | 1,003 | 306 | 169 | 187 | 654 | 186 | 184 | 637 |

**Table 3**

Mean Sample Take and Relative Bias of Each of Eight Estimators over 500 Repeated Samples from the Entire Population
Industrial group: ACCOMMODATION; Areas: 16 Census Divisions in Nova Scotia.

| Area | Mean Sample Take | Estimator EXP | POS | SYN/C | MRE/C | REG/C | SYN/R | MRE/R | REG/R |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.25 | 0.01 | −0.75 | −0.08 | −0.06 | −0.01 | 0.36 | 0.28 | 0.01 |
| 2 | 1.37 | −0.06 | −0.21 | 0.25 | 0.10 | 0.02 | 0.25 | 0.11 | 0.02 |
| 3 | 1.02 | 0.06 | −0.26 | 0.19 | 0.09 | 0.04 | 0.12 | 0.06 | 0.03 |
| 4 | 0.23 | −0.10 | −0.77 | −0.33 | −0.26 | −0.07 | −0.15 | −0.13 | −0.05 |
| 5 | 2.04 | 0.03 | −0.13 | 0.21 | 0.08 | 0.03 | 0.18 | 0.06 | 0.01 |
| 6 | 1.49 | 0.04 | −0.13 | 0.17 | 0.10 | 0.03 | 0.03 | 0.02 | 0.01 |
| 7 | 1.53 | 0.01 | −0.18 | −0.29 | −0.11 | −0.01 | −0.30 | −0.12 | −0.02 |
| 8 | 1.54 | 0.03 | −0.19 | −0.42 | −0.17 | −0.01 | −0.26 | −0.11 | −0.02 |
| 9 | 6.83 | 0.01 | −0.02 | 0.13 | 0.02 | 0.00 | 0.12 | 0.02 | 0.00 |
| 10 | 1.26 | −0.01 | −0.26 | 0.40 | 0.17 | 0.03 | 0.30 | 0.13 | 0.02 |
| 11 | 3.06 | 0.04 | −0.02 | 0.51 | 0.21 | 0.08 | 0.40 | 0.16 | 0.06 |
| 12 | 1.80 | 0.02 | −0.16 | −0.08 | −0.05 | −0.03 | −0.23 | −0.10 | −0.03 |
| 14 | 1.04 | 0.02 | −0.33 | −0.52 | −0.23 | −0.07 | −0.32 | −0.15 | −0.06 |
| 15 | 1.54 | −0.03 | −0.23 | −0.21 | −0.13 | −0.08 | −0.15 | −0.11 | −0.08 |
| 17 | 3.08 | −0.07 | −0.05 | −0.03 | −0.01 | 0.00 | −0.14 | −0.07 | −0.03 |
| 18 | 0.52 | 0.04 | −0.54 | 3.26 | 3.20 | 0.60 | 2.97 | 2.92 | 0.50 |

**Table 4**

Mean Squared Error of Each of Eight Estimators over 500 Repeated Simple Random Samples from the Entire Population
Industrial Group: ACCOMMODATION; Areas: 16 Census Divisions in Nova Scotia.

| Area | Estimator EXP | POS | SYN/C | MRE/C | REG/C | SYN/R | MRE/R | REG/R |
|---|---|---|---|---|---|---|---|---|
| 1 | 1,142 | 283 | 9 | 7 | 25 | 58 | 44 | 164 |
| 2 | 7,467 | 5,082 | 877 | 631 | 1,077 | 747 | 455 | 726 |
| 3 | 878 | 442 | 48 | 163 | 242 | 24 | 116 | 163 |
| 4 | 155 | 43 | 7 | 6 | 17 | 3 | 3 | 6 |
| 5 | 15,200 | 8,392 | 2,091 | 2,270 | 3,230 | 1,271 | 1,208 | 1,785 |
| 6 | 5,239 | 3,906 | 253 | 1,038 | 2,193 | 54 | 396 | 792 |
| 7 | 21,197 | 8,781 | 3,569 | 1,831 | 3,016 | 3,709 | 1,812 | 2,948 |
| 8 | 14,071 | 6,738 | 3,608 | 2,122 | 4,018 | 1,492 | 947 | 1,766 |
| 9 | 50,606 | 27,867 | 9,980 | 11,413 | 14,344 | 6,575 | 7,779 | 9,991 |
| 10 | 2,219 | 993 | 590 | 362 | 665 | 317 | 151 | 280 |
| 11 | 10,535 | 5,774 | 6,366 | 5,126 | 7,154 | 3,867 | 2,752 | 3,673 |
| 12 | 16,787 | 10,485 | 543 | 1,148 | 1,944 | 1,245 | 1,130 | 1,836 |
| 14 | 51,471 | 25,644 | 9,669 | 8,221 | 14,155 | 3,972 | 3,189 | 5,077 |
| 15 | 59,207 | 41,381 | 4,861 | 10,548 | 18,119 | 2,759 | 4,262 | 6,636 |
| 17 | 29,632 | 25,211 | 1,501 | 3,023 | 4,754 | 1,765 | 2,123 | 3,214 |
| 18 | 286 | 99 | 2,062 | 2,112 | 5,623 | 1,607 | 1,646 | 4,561 |

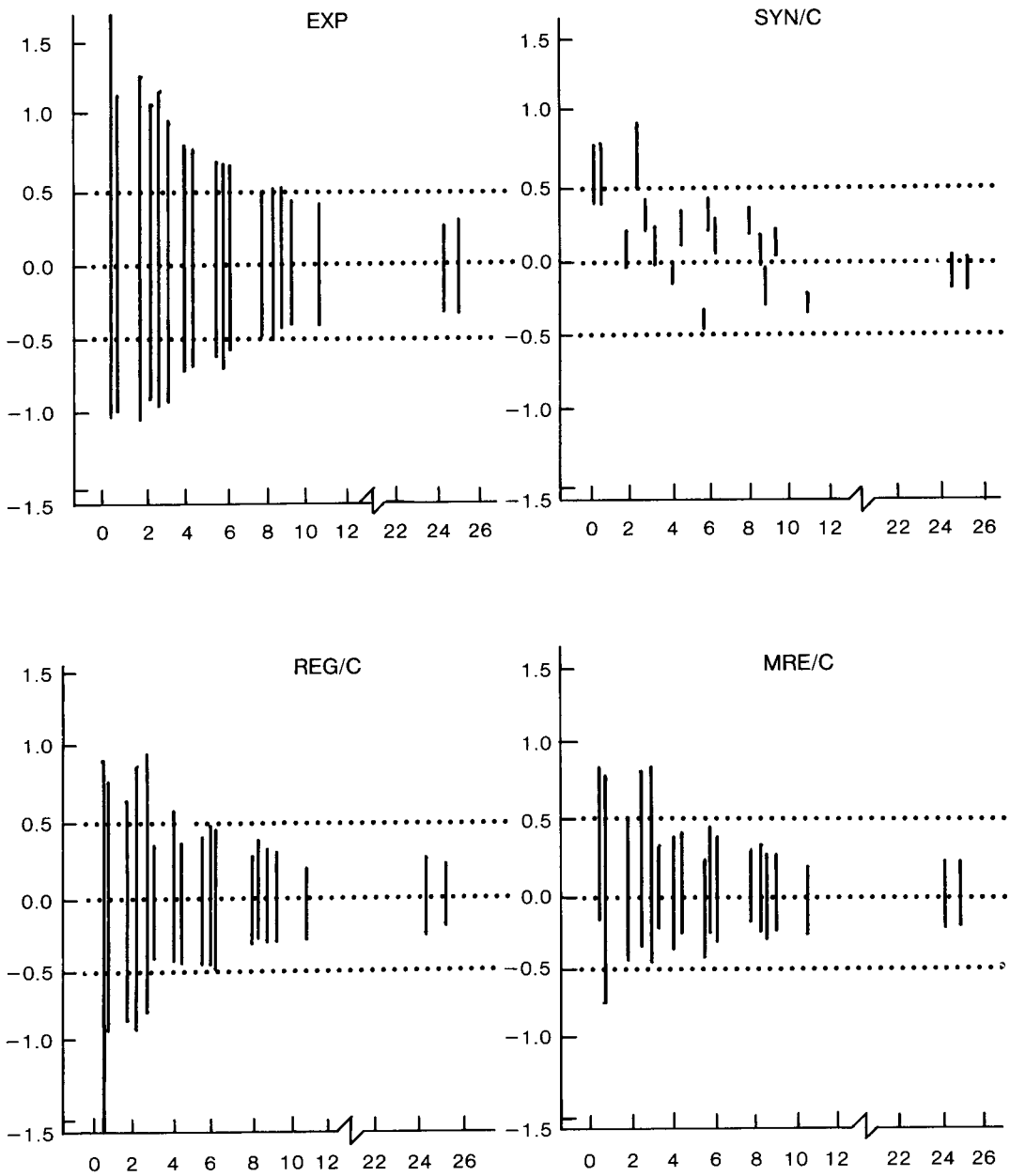
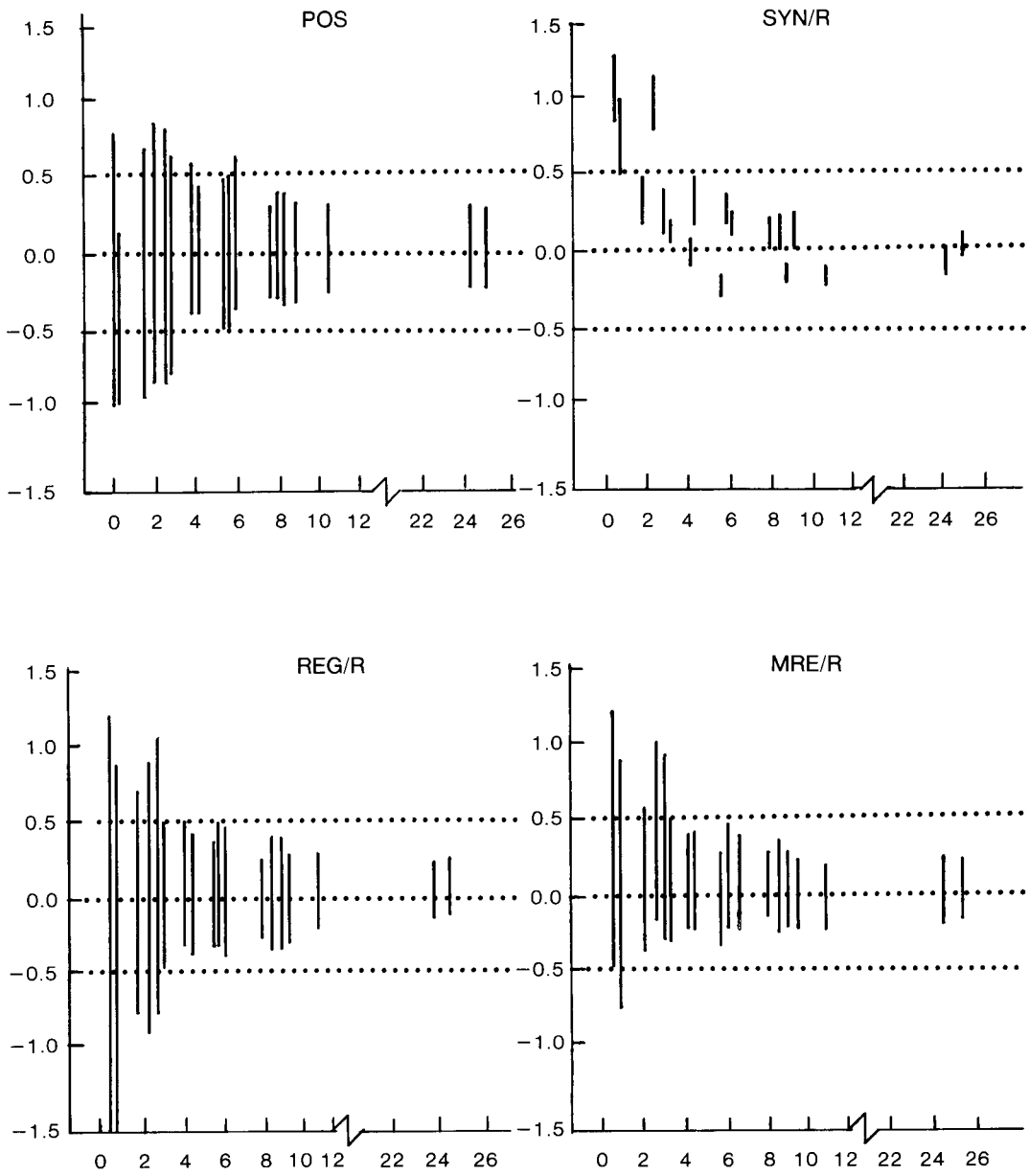



**Figure 1:** Distribution band of relative error for selected estimators — abscissa represents mean sample take. Industrial Group: RETAIL. Areas: 18 Census Divisions in Nova Scotia.

Figure 1 (continued)



## 5. CONCLUSIONS

In summary we find that the overall performance of the MRE estimators is such that we suggest them as promising alternatives for future applications of small area estimation. The recommended confidence interval procedure based on the MRE estimators is given in section 3.

We think that the MRE method presented here involves a simple mechanism for steering the estimates slightly in the direction of the stable SYN estimators, when the sample take is less than expected. This goal is also manifested (but attained by different means) in such other attempts as the empirical Bayes (Fay and Herriot, 1979) and sample-dependent (Drew, Singh, and Choudhry 1982) methods of estimation.

## REFERENCES

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,* 8, 17-47.

FAY, R.E. and HERRIOT, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74, 269-277.

HIDIROGLOU, M.A., MORRY, M., DAGUM, E.B., RAO, J.N.K. and SÄRNDAL, C.E. (1984). Evaluation of alternative small area estimators using administrative data. Paper presented at ASA meetings, Philadelphia, August, 1984.

SÄRNDAL, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustments for nonresponse. *Bulletin of the International Statistical Institute,* 49:1, 494-513. (proceedings, 43rd session, Buenos Aires).

SÄRNDAL, C.E. and RÅBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review,* 1983:5 (Essays in honour of T.E. Dalenius), 33-40.

SÄRNDAL, C.E. (1984). Design-Consistent versus Model-Dependent Estimation for Small Domains. *Journal of the American Statistical Association,* 79, 624-631.