Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

Canada

504??

EXPERIMENTS WITH MODIFIED REGRESSION ESTIMATORS FOR SMALL DOMAINS

M.A. Hidiroglou and C.E. Sarndal

Working Paper No. BSMD 85-002E

EXPERIMENTS WITH MODIFIED REGRESSION ESTIMATORS FOR SMALL DOMAINS

by

M.A. Hidiroglou, Statistics Canada

and

C.E. Sãrndal, Université de Montréal and Statistics Canada

July, 1984

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, 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 Monte Carlo study involving 500 samples.

Experiments with Modified Regression Estimators for Small Domains

by

M.A. Hidiroglou and C.E. Särndal

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 experiments 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.

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 Dagum, Hidiroglou, Morry, 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, ..., D; g=1, ..., G. Let N_{dg} be the size of U_{dg}.

Then the population size N can be expressed as

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

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 $n_{s_{d.}}$, $n_{s_{.g}}$ and $n_{s_{dg}}$. 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 of x_k for the k:th unit, and

x_k is known for all k=1, ..., N.

The following estimators of the domain total $t_d = \Sigma_{U_{d.}} y_k$ are compared.

The straight expansion estimator (EXP):

$$\hat{t}_{dEXP} = \frac{N}{n} \Sigma_{s_{d.}} y_k \qquad (2.2)$$

The poststratified estimator (POS) :

$$\hat{t}_{dPOS} = N_{d \cdot} \bar{y}_{s_{d \cdot}} \qquad (2.3)$$

where

$$\bar{y}_{s_{d \cdot}} = \Sigma_{s_{d \cdot}} y_k / n_{s_{d \cdot}}$$

is the mean of the $n_{s_{d \cdot}}$ y - values from the d:th domain. If $n_{s_{d \cdot}} = 0$ 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, REG and MRE have been investigated, the "Count" version and the " Ratio" version.

The formulas for the "Count" versions are:

Synthetic-Count estimator (SYN/C):

$$\hat{t}_{dSYN/C} = \sum_{g=1}^{G} N_{dg} \bar{y}_{s_{\cdot g}} \qquad (2.4)$$

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

Regression-Count estimator (REG/C):

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

where $\bar{y}_{s_{dg}}$ is the mean of y in s_{dg}, and $\hat{N}_{dg} = Nn_{s_{dg}}/n$. 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.

Modified Regression-Count estimator (MRE/C):

$$\hat{t}_{dMRE/C} = \sum_{g=1}^{G} \{N_{dg} \bar{y}_{s_{.g}} + F_d \hat{N}_{dg} (\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})\} \tag{2.6}$$

with

$$F_d = \begin{cases} E_d/n_{s_{d.}} & \text{if } n_{s_{d.}} \geq E_d \\ n_{s_{d.}}/E_d & \text{if } n_{s_{d.}} < E_d \end{cases}$$

where

$$E_d = E_{srs}(n_{s_{d.}}) = nN_{d.}/N$$

is the expected sample take, under simple random sampling, from the d:th domain.

The MRE/C estimator thus differs from the ordinary REG/C estimator in that the bias correction term receives a weight, F_d, which is bounded above by unity, and attains unity when the sample take equals its expectation. The theoretical justification for F_d is given in Section 5. Intuitively, the effect of F_d is to dampen the variance contributed by the correction term.

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

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

Synthetic-Ratio estimator (SYN/R):

$$\hat{t}_{dSYN/R} = \sum_{g=1}^{G} X_{dg} \hat{R}_g \tag{2.7}$$

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

$$\hat{R}_g = \Sigma_{s_{\cdot g}} y_k / \Sigma_{s_{\cdot g}} x_k$$

Regression - Ratio estimator (REG/R):

$$\hat{t}_{dREG/R} = \sum_{g=1}^{G} \{X_{dg} \hat{R}_g + \hat{N}_{dg} (\bar{y}_{s_{dg}} - \hat{R}_g \bar{x}_{s_{dg}})\} \tag{2.8}$$

Modified Regression - Ratio estimator (MRE/R):

$$\hat{t}_{dMRE/R} = \sum_{g=1}^{G} \{X_{dg} \hat{R}_g + F_d \hat{N}_{dg} (\bar{y}_{s_{dg}} - \hat{R}_g \bar{x}_{s_{dg}})\} \tag{2.9}$$

where F_d is defined as in the MRE/C estimator above.

3. Results from the empirical study

The hypothesis that we expected to verify was that the MRE estimator is situated, with respect to both bias and variance between the SYN and REG estimators. We expected on the part of the MRE estimators a rather small bias and a substantial decrease in variance and Mean Squared Error as compared to the REG estimators. These hypotheses were indeed borne out by the empirical results.

For the Monte Carlo study reported in Dagum et al (1984) 500 samples had been drawn from a Nova Scotia population of N=1678 unincorporated tax filers. The results in Table 1-6 are based on these same 500 samples. From these tables, the following conclusions emerge: (where conclusion C states the main new results, whereas A and B resumes what is known from earlier work Särndal and Rabäck (1983); Dagum 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.

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.

In summary we find that the overall performance of the MRE estimators is such that we suggest them as interesting alternatives for future applications of small area estimation. The recommended confidence interval procedure based on the MRE estimators is given in section 5.

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 very 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.

4. The REG estimation method

This section and the next contain a brief presentation of the theoretical arguments underlying the REG and MRE estimators. This material can be skipped by readers more interested in the empirical results already presented.

The REG estimation method is motivated by the following requirements: (a) to obtain approximately design unbiased estimates with simple variance estimates and easily calculated (and meaningful) confidence intervals; (b) to strengthen the estimates by involving sample data from all domains.

A regression model is fitted and the auxiliary variables are used to create predicted or "imputed" values for the units in the domain. We assume here more generally that the sampling design, p, is an arbitrary one (not necessarily srs) with inclusion probabilities Π_k (first order) and $\Pi_{k\ell}$ (second order).

We assume a regression model such that the y_k are independent (throughout) random variables with

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

As an estimator of $\underset{\sim}{\beta}$, use

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

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

Note that the estimator $\hat{\underset{\sim}{\beta}}$ pools together sample data from <u>all</u> domains.

Let the k:th predicted value be

$$\hat{y}_k = \underset{\sim}{x}_k' \hat{\underset{\sim}{\beta}}$$

and denote the k:th residual by

$$e_k = y_k - \hat{y}_k$$

Following Särndal (1981), we take

$$\hat{t}_{dREG} = \Sigma_{U_{d.}} \hat{y}_k + \Sigma_{s_{d.}} e_k/\Pi_k \tag{4.1}$$

as our nearly unbiased estimator of the unknown d:th domain total,

$$t_d = \Sigma_{U_{d.}} y_k \quad . \tag{4.2}$$

The first term of (4.1),

$$\hat{t}_{dSYN} = \Sigma_{U_{d.}} \hat{y}_k \tag{4.3}$$

can, by virtue of its form, be seen as a natural estimator of t_d .

However, (4.3) is biased, and the second term $\Sigma_{s_{d.}} e_k/\Pi_k$ of (4.1) is therefore added to remove the bias.

We shall call $\Sigma_{U_{d.}} \hat{y}_k$ <u>the synthetic term</u> of the estimator $\hat{t}_{dREG}$. (For the particular model (4.5) below, this term gives the original synthetic estimator, (2.4)).

The second term, $\Sigma_{s_{d.}} e_k/\Pi_k$, will be called the correction term.

The estimated variance under the sampling design p is

$$\hat{V}_p(\hat{t}_{dREG}) = \underset{s_{d.}}{\sum\sum_{k \neq \ell}} \Delta_{k\ell}\, e_k\, e_\ell/\Pi_k\, \Pi_\ell$$

where

$$\Delta_{k\ell} = \begin{cases} \Pi_k(1-\Pi_k) & \text{if } \ell = k \\ \Pi_{k\ell} - \Pi_k\,\Pi_\ell & \text{if } \ell \neq k \end{cases} \tag{4.4}$$

A 100 $(1-\alpha)$% (design-based) confidence interval for $\hat{t}_{dREG}$ is given by

$$\hat{t}_{dREG} \pm z_{1-\alpha/2}\, \{\hat{V}_p(\hat{t}_{dREG})\}^{\frac{1}{2}} .$$

These results were given in Särndal (1981).

Of a particular interest in our application are estimators that arise from the general formulas (4.1) and (4.3) in cases where the model is formulated in terms of G groups that cut across the domains. Two such models are now examined.

A. Model leading to "count" estimators.

Assume that, for $g = 1, \ldots, G$,

$$E_\xi\,(y_k) = \beta_g;\ V_\xi(y_k) = \sigma_g^2\ ;\ k \in U_{.g} \tag{4.5}$$

We find

$$\hat{\beta}_g = (\Sigma_{s_{.g}}\, y_k/\Pi_k)/(\Sigma_{s_{.g}}\, 1/\Pi_k) = \tilde{y}_{s_{.g}} ,$$

say, and the estimator of $t_{d.}$ becomes

$$\hat{t}_{dREG/C} = \sum_{g=1}^{G} \{N_{dg}\, \tilde{y}_{s_{.g}} + \hat{N}_{dg}\, (\tilde{y}_{s_{dg}} - \tilde{y}_{s_{.g}})\} \tag{4.6}$$

with

$$\hat{N}_{dg} = \Sigma_{s_{dg}}\, 1/\Pi_k$$

and

$$\tilde{y}_{s_{dg}} = (\Sigma_{s_{dg}}\, y_k/\Pi_k)\, /\hat{N}_{dg}$$

In the special case of simple random sampling (srs), (4.6) reduces to the REG/C formula (2.5)

Also, under srs, the synthetic (first) term of (4.6) becomes the SYN/C estimator (2.4).

Note that the population cell counts N_{dg} must be known in (4.6).

B. <u>Model leading to "ratio" estimators</u>

Let, for g=1, ..., G

$$E_\xi(y_k) = \beta_g\, x_k \;;\; V_\xi(y_k) = \sigma_g^2\, x_k,\; k \epsilon U_{.g} \tag{4.7}$$

We obtain

$$\hat{\beta}_g = \frac{\sum_{g=1}^{G} \hat{N}_{dg}\, \tilde{y}_{s_{dg}}}{\sum_{g=1}^{G} \hat{N}_{dg}\, \tilde{x}_{s_{dg}}}$$

and

$$\hat{t}_{dREG/R} = \sum_{g=1}^{G} \{X_{dg} \hat{\beta}_g + \hat{N}_{dg} (\tilde{y}_{s_{dg}} - \hat{\beta}_g \tilde{x}_{s_{dg}})\} \tag{4.8}$$

where the totals $X_{dg} = \Sigma_{U_{dg}} x_k$ are required auxiliary information.

It is easy to see that in the special case of srs, then (4.8) becomes the REG/R formula (2.8) included in our study, while the synthetic term $\sum_{g=1}^{G} X_{dg} \hat{\beta}_g$ becomes the SYN/R formula (2.7).

5. The MRE estimation method

If $s_{d.}$ is non-empty, an approximately unbiased alternative to the REG estimator (4.1) is given by

$$\hat{t}_{dALT} = \Sigma_{U_{d.}} \hat{y}_k + N_{d.} \frac{\Sigma_{s_{d.}} \dot{e}_k/\Pi_k}{\hat{N}_{d.}} \tag{5.1}$$

where

$$\hat{N}_d = \Sigma_{s_{d.}} 1/\Pi_k$$

is the estimated domain size.

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

$$\frac{\Sigma_{s_{d.}} e_k/\Pi_k}{\Sigma_{s_{d.}} 1/\Pi_k},$$

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

The size $n_{s_{d.}}$ 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, were substantial (say, $E_d \geq 50$), then it is practically certain that the realized sample take, $n_{s_{d.}}$, will not be exceedingly small. For example,under srs, values $n_{s_{d.}} \leq 30$ will hardly ever occur.

In such situations, the nearly unbiased estimator (5.1) can be recommended as is. It should realize important efficiency gains over (4.1), notably in domains where the model fits 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 $n_{s_{d.}}$ 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, $\Sigma_{s_{d.}} e_k/\Pi_k$, as in (4.1), or in its ratio form, $N_{d.} (\Sigma_{s_{d.}} e_k/\Pi_k)/(\Sigma_{s_{d.}} 1/\Pi_k)$, as in (5.1).

To counteract this inflated variance contribution, we modify the correction term of (5.1) 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 $n_{s_{d.}}$ 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, $n_{s_{d.}}$, and expected sample take,

$$E_d = E_p(n_{s_{d.}}) = \Sigma_{U_{d.}} \Pi_k .$$

More specifically, when $n_{s_{d.}} < E_d$, let us multiply the correction term of (5.1) by a "dampening factor" chosen as $(\hat{N}_{d.}/N_{d.})^2$. That is, instead of

$$\frac{N_{d.} \Sigma_{s_{d.}} e_k/\Pi_k}{\hat{N}_{d.}}$$

the correction term, for $n_{s_{d.}} < E_d$, will be

$$\left(\frac{\hat{N}_{d.}}{N_{d.}}\right)^2 \frac{N_{d.} \Sigma_{s_{d.}} e_k/\Pi_k}{\hat{N}_{d.}} = \frac{\hat{N}_{d.}}{N_{d.}} \Sigma_{s_{d.}} e_k/\Pi_k .$$

When $n_{s_{d.}} \geq E_d$, we see no reason to change the correction term. The resulting estimator incorporating these two types of realizations of $n_{s_{.d}}$ is

$$\hat{t}_{dMRE} = \Sigma_{U_{d.}} \hat{y}_k + F_d \Sigma_{s_{d.}} e_k/\Pi_k \qquad (5.2)$$

where

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

In the case of simple random sampling, and under the models (4.5) and (4.7) respectively, we then obtain the MRE/C estimator, (2.6), and the MRE/R estimator, (2.9).

It can be shown that (2.6) and (2.9) are nearly unbiased conditionally on $n_{s_{d.}}$, as long as $n_{s_{d.}} \geq E_d$. For $n_{s_{d.}} < E_d$, the MRE has some conditional bias, which tends to increase the more $n_{s_{d.}}$ falls short of its expected value. At the same time the MRE estimator is being pushed towards the SYN estimator, thus benefitting from the stability (low variance) of the SYN estimator. Unconditionally, the MRE estimator (5.1) will have a certain small bias, but a much reduced variance compared with the REG estimators, as shown empirically by our Tables 1-6.

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 (4.1), and are especially likely when $n_{s_{d.}} < E_d$. The new MRE estimator virtually eliminated the occurrence (in the series of 500 samples that we drew) of negative estimates. In practice, if by a remote possibility the MRE takes a negative value, we recommend (as done in the results shown in Tables 1-6) to redefine the MRE estimator as being equal to the always positive SYN estimator.

A natural formula for estimating the variance of (5.1) is

$$\hat{V}_p(\hat{t}_{dALT}) = \left(\frac{N_{d.}}{\hat{N}_{d.}}\right)^2 \sum_{\substack{k \neq \ell \\ \in s_{d.}}}\sum \Delta_{k\ell} \, (e_k - \bar{e}_{s_{d.}})(e_\ell - \bar{e}_{s_{d.}})/\Pi_k \, \Pi_\ell \qquad (5.3)$$

where

$$\bar{e}_{s_{d.}} = (\Sigma_{s_{d.}} e_k)/n_{s_{d.}} ,$$

and $\Delta_{k\ell}$ is defined by (4.4). We propose that the same formula may serve well to estimate the variance of the MRE estimator (5.2). It is true that (5.1) differs from (5.2) when the realized sample take falls short of the expected; however, we do not foresee the difference to be great enough to cause serious distortion in the validity of a confidence interval for t_d centred on $\hat{t}_{dMRE}$ and using (5.3) as the estimated variance.

Table 1. Industrial group: RETAIL. Areas: 18 census divisions in Nova Scotia. Mean sample take and mean of each of eight estimators over 500 repeated simple random samples from the entire population. Column three shows the true domain total.

AREA	MEAN SAMPLE TAKE	TRUE TOTAL	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	1.762	68.07	66.88	59.06	76.43	69.80	66.35	88.77	74.13	66.55
2	5.448	412.07	412.27	396.34	265.68	369.11	401.91	299.53	377.63	402.46
3	3.896	186.98	183.77	188.38	171.85	183.51	187.73	184.24	185.71	186.82
4	3.024	109.48	110.91	103.85	125.47	115.09	110.99	123.43	113.67	110.31
5	5.932	241.54	241.85	244.54	292.61	253.09	242.72	272.83	249.15	242.72
6	7.628	282.39	277.06	280.53	360.64	301.31	285.91	311.54	289.20	283.19
7	8.610	479.30	488.89	483.18	400.65	465.80	482.20	392.56	465.20	483.09
8	5.642	213.01	207.96	210.36	286.23	233.83	218.61	264.78	226.62	215.49
9	24.640	1118.48	1114.32	1117.59	1094.44	1120.20	1123.81	1108.75	1122.39	1124.25
10	8.920	401.75	392.64	393.87	461.31	409.27	397.77	436.22	403.70	396.86
11	8.346	392.58	380.66	385.95	423.35	397.29	391.23	431.89	399.77	392.06
12	10.576	689.80	692.31	689.24	503.99	654.19	687.09	561.83	665.66	688.99
13	0.478	20.35	19.52	8.45	32.68	27.62	21.14	40.60	32.10	21.18
14	2.798	77.05	79.53	74.45	102.29	85.18	76.81	95.20	84.37	78.90
15	4.212	163.10	173.08	161.19	208.24	173.63	162.42	212.64	174.16	162.03
16	2.244	76.74	78.66	72.93	133.37	96.92	79.10	148.89	101.31	78.52
17	23.950	1100.05	1093.61	1093.40	1080.25	1096.25	1097.59	1043.01	1091.87	1097.88
18	0.542	20.00	21.49	9.29	32.68	26.71	18.77	33.32	27.05	18.72

Table 2. Industrial group: RETAIL. Areas: 18 census divisions in Nova Scotia. Variance of each of eight estimators over 500 repeated simple random samples from the entire population.

AREA	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	3214	2129.8	26.54	695.6	1396.8	34.27	733.9	1485.2
2	42683	24424.2	352.51	10900.9	17289.6	444.65	9087.9	14317.1
3	10480	6870.5	128.02	2585.3	4220.0	138.93	2336.7	3790.4
4	5635	3632.8	68.29	716.9	1186.3	62.41	1191.3	1856.7
5	14593	9691.5	391.23	4966.5	7374.2	315.62	3943.0	5985.4
6	12304	5694.4	591.87	3071.9	4285.2	406.25	1704.9	2519.6
7	34843	18008.9	727.94	9224.1	13469.6	638.36	11844.7	17259.7
8	12064	8640.7	411.29	3267.4	5024.1	301.15	3350.0	4990.0
9	73103	40520.6	5208.32	24070.9	29280.2	4983.69	21319.8	25850.9
10	22052	9390.4	1012.98	5837.7	7927.8	821.91	5372.8	7262.9
11	23424	12486.6	832.94	6729.8	9595.6	804.73	7854.0	11085.3
12	46669	21917.7	1155.15	12385.1	17116.1	1333.52	11710.4	16547.1
13	635	102.7	8.63	42.6	228.8	12.27	150.1	784.3
14	3872	2848.1	55.60	1190.4	2145.3	48.75	1322.6	2347.2
15	8034	3514.8	211.82	1784.8	2811.4	196.64	1867.2	2941.7
16	3248	2117.3	109.05	1158.9	2516.0	126.90	1140.2	2656.4
17	81332	47805.1	5121.65	29001.2	35297.4	4435.96	27445.5	33197.7
18	1003	192.7	8.63	142.3	654.5	8.26	135.1	636.7

Table 3. Industrial group: RETAIL. Areas: 18 census divisions in Nova Scotia. Mean Squared Error of each of eight estimators over 500 repeated simple random samples from the entire population.

AREA	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	3209	2206.8	96	697.1	1397.0	462	769.2	1484.5
2	42598	24623.0	21782	12725.4	17358.3	13110	10256.2	14300.8
3	10469	6859.8	357	2592.2	4212.2	146	2333.7	3782.9
4	5626	3657.2	324	746.9	1166.2	257	1206.5	1853.7
5	14554	9681.2	2999	5090.0	7360.9	1294	3993.0	5974.9
6	12308	5686.5	6713	3423.5	4289.0	1255	1747.8	2515.2
7	34265	17988.0	6912	9397.8	13451.0	8161	12019.7	17239.6
8	12066	8630.5	5772	3694.2	5045.4	2981	3528.7	4986.1
9	72974	40440.3	5776	24025.7	29250.1	5068	21292.5	25832.6
10	22091	9433.7	4559	5892.6	7927.7	2009	5365.9	7272.3
11	23519	12505.6	1778	6738.6	9578.2	2348	7890.0	11063.4
12	46588	21874.1	35310	13558.5	17084.8	17454	12222.8	16514.1
13	635	244.3	161	95.4	228.9	422	287.9	783.4
14	3871	2849.1	692	1254.2	2141.1	378	1373.5	2346.0
15	8028	3511.5	2249	1892.0	2806.2	2651	1985.8	2937.0
16	3245	2127.6	3316	1563.8	2516.6	5333	1741.6	2654.3
17	81211	47753.7	5503	28957.6	35232.8	7681	27457.6	33135.0
18	1003	306.9	169	187.1	654.7	186	184.6	637.0

TABLE 4. Industrial group: ACCOMMODATION. Areas: 16 census divisions in Nova Scotia. Mean sample take and mean of each of eight estimators over 500 repeated samples from the entire population. Column 3 shows the true domain total.

AREA	MEAN SAMPLE TAKE	TRUE TOTAL	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	0.252	19.45	19.61	4.90	17.81	18.19	19.32	26.48	24.80	19.73
2	1.370	90.72	84.84	71.26	113.41	99.78	92.10	113.83	100.39	92.80
3	1.016	27.50	29.06	20.27	32.82	30.00	28.51	30.75	29.05	28.23
4	0.226	7.42	6.71	1.68	5.00	5.48	6.92	6.28	6.48	7.05
5	2.040	139.31	143.97	120.85	168.40	151.06	143.39	164.85	147.81	140.08
6	1.488	69.42	72.13	60.45	81.24	76.47	71.53	71.54	70.70	70.05
7	1.526	194.74	196.79	160.08	139.02	172.96	192.05	135.74	171.04	191.50
8	1.538	140.36	144.68	113.88	81.24	116.66	138.87	103.17	124.69	138.22
9	6.828	446.87	451.18	439.83	507.08	457.72	445.50	500.06	456.04	444.68
10	1.258	54.29	53.97	40.22	76.24	63.69	55.85	70.60	61.32	55.53
11	3.056	146.00	152.29	143.02	220.26	177.25	157.54	204.80	169.32	155.27
12	1.802	142.74	145.32	120.51	131.22	136.23	139.03	109.97	128.65	138.29
14	1.044	187.28	191.06	125.35	90.60	144.68	174.23	127.91	158.50	175.39
15	1.540	225.03	217.79	172.41	177.76	194.87	206.23	191.90	200.64	206.61
17	3.084	237.99	221.49	225.87	231.83	234.60	237.19	204.33	222.34	230.69
18	0.516	12.91	13.47	5.96	54.99	54.28	20.67	51.28	50.57	19.35

TABLE 5. Industrial group: ACCOMMODATION. Areas: 16 census divisions in Nova Scotia. Variance of each of eight estimators over 500 repeated simple random samples from the entire population.

AREA	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	1144	71	6.6	6	25	8.2	16	164
2	7447	4713	362.8	550	1078	213.6	363	723
3	877	391	20.1	157	241	13.4	114	162
4	155	10	1.4	2	17	1.7	2	6
5	15209	8068	1247.3	2137	3220	620.5	1138	1788
6	5242	3834	113.8	990	2193	50.0	395	793
7	21235	7594	464.5	1359	3015	227.6	1253	2944
8	14081	6049	113.8	1563	4024	108.7	703	1765
9	50689	27873	6368.2	11318	14371	3754.3	7711	10006
10	2223	796	108.4	274	663	51.1	101	279
11	10517	5776	853.2	4158	7035	410.8	2213	3594
12	16814	10011	411.4	1107	1934	171.1	934	1820
14	51560	21851	322.4	6419	14013	448.2	2365	4945
15	59273	38689	2631.5	9657	17801	1664.3	3674	6309
17	29419	25114	1465.6	3018	4763	633.3	1882	3167
18	286	51	291.8	401	5574	135.3	228	4528

Table 6. Industrial group: ACCOMMODATION. Areas: 16 census divisions in Nova Scotia. Mean Squared Error of each of eight estimators over 500 repeated simple random samples from the entire population.

AREA	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	1142	283	9	7	25	58	44	164
2	7467	5082	877	631	1077	747	455	726
3	878	442	48	163	242	24	116	163
4	155	43	7	6	17	3	3	6
5	15200	8392	2091	2270	3230	1271	1208	1785
6	5239	3906	253	1038	2193	54	396	792
7	21197	8781	3569	1831	3016	3709	1812	2948
8	14071	6738	3608	2122	4018	1492	947	1766
9	50606	27867	9980	11413	14344	6575	7779	9991
10	2219	993	590	362	665	317	151	280
11	10535	5774	6366	5126	7154	3867	2752	3673
12	16787	10485	543	1148	1944	1245	1130	1836
14	51471	25644	9669	8221	14155	3972	3189	5077
15	59207	41381	4861	10548	18119	2759	4262	6636
17	29632	25211	1501	3023	4754	1765	2123	3214
18	286	99	2062	2112	5623	1607	1646	4561

REFERENCES

Dagum, E.B., Hidiroglou, M.A., Morry, M., Rao, J.N.K. and Särndal, C.E. (1984) Evaluation of alternative small area estimators using administrative data. Paper to be presented at ASA meetings, Philadelphia, August, 1984.

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.

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

APPENDIX

Figure 1 contains in a nutshell the more detailed information in Tables 1-3. The figure consists of eight graphs, one for each of the eight estimators. 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 in question; 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. For the estimators having small or negligible bias (EXP, POS, REG and MRE), the graphs thus convey the message of a decreasing variance as the mean sample take increases; this, of course, confirms our intuition.

Some other observations:

1. The SYN/C and SYN/R are seen to have considerable bias in some domains; however, they have, in all domains, a small variance in comparison to the other estimators.

2. REG/C and REG/R have smaller variances than EXP and POS, with exception made for the smallest domains. In the smallest domains, none of the unbiased estimators (EXP, POS, REG/C, REG/G) is attractive from a variance point of view; this is especially true for the REG estimators.

3. This problem is remedied by the two MRE modifications of the REG estimators. For the MRE estimators, the bias is small and the variance constrained to 'within reasonable limits', even in the smallest domains. Thus, the two MRE estimators present the best overall image of the estimators compared.

Figure 1: Industrial Group: RETAIL. Areas: 18 census divisions in Nova Scotia. Distribution band of relative error for selected estimators - abscissa represents mean sample take.

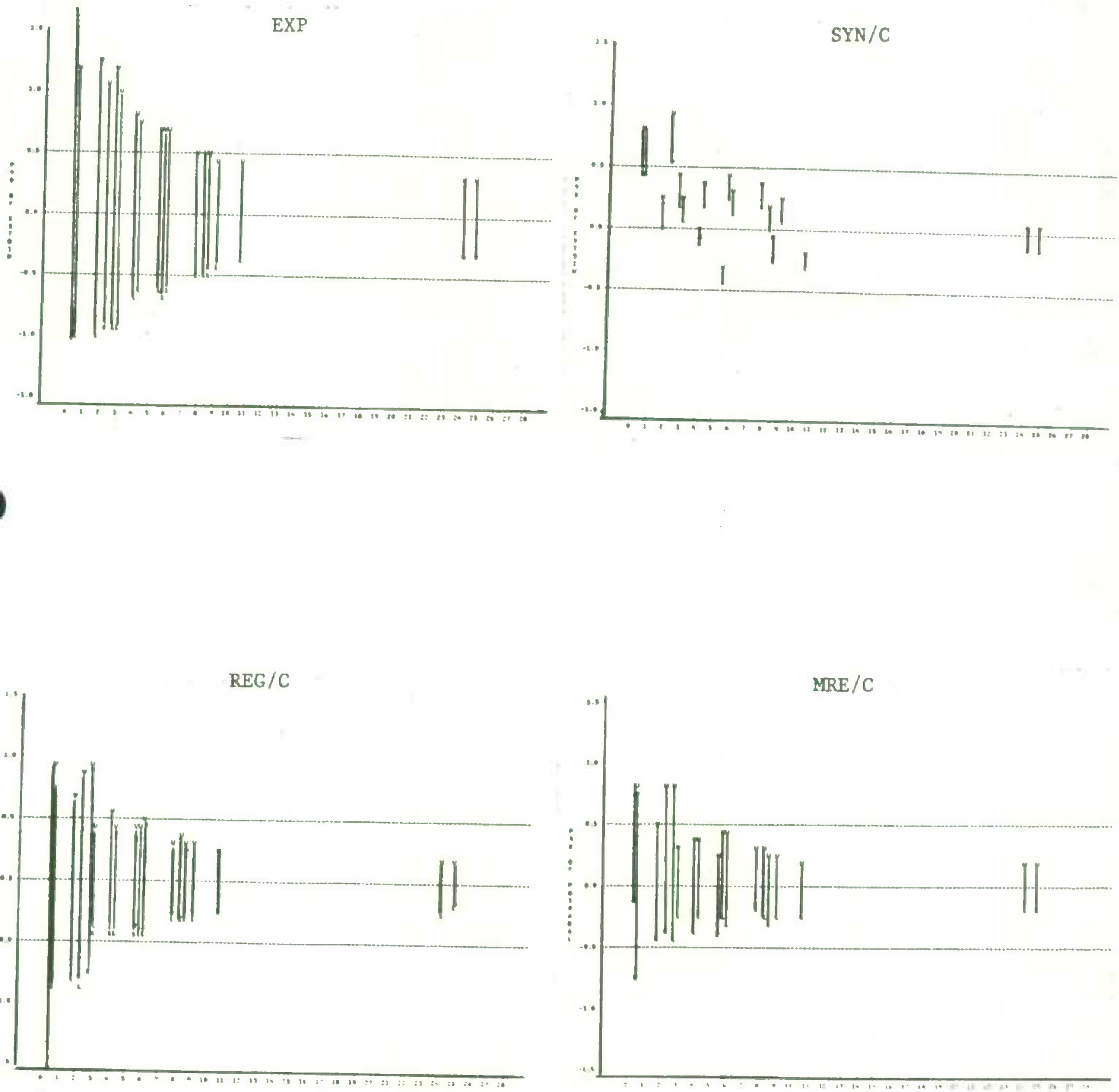

Figure 1: Industrial Group: RETAIL. Areas: 18 census divisions in Nova Scotia. Distribution band of relative error for selected estimators - abscissa represents mean sample take.