EVALUATION OF ALTERNATIVE SMALL AREA ESTIMATORS USING ADMINISTRATIVE RECORDS

M.A. Hidiroglou, M. Morry, E.B. Dagum, Statistics Canada
J.N.K. Rao, Carleton University, C.E. Särndal, Université de Montréal

1.0 INTRODUCTION

Due to increasing emphasis on planning and administering economic programs at local levels, there has been a demand for more and better quality data at these levels on a wide range of economic data. Such data available from surveys may not have adequate precision and hence there is an increasing demand on the use of administrative records to produce this data. Administrative sources, however, may not contain all the required information on a one-hundred percent basis. It may therefore be necessary to pool this information with the survey data. In the present context, a number of variables are available on 100% basis from one administrative source provided by Revenue Canada, whereas some of the variables of interest as well as variables common to an administrative source provided by Statistics Canada are available on a sample basis. The problem at hand is to use these various administrative files in conjunction with survey data to produce reliable small area estimates.

In this paper, some estimators for small areas are evaluated in the context of producing Census Division level by Major Industrial Division estimates, using the unincorporated data compiled at Statistics Canada and Revenue Canada. Several of the collected variables are candidates for small area estimation, but we will focus on Wages and Salaries. This variable is available on a 25% sample basis at Statistics Canada for the Gross Business Income range $25,000 to $500,000, but not available from the Revenue Canada file. Wages and Salaries are related to Gross Business Income (available from both sources) for certain industrial groupings. Hence, Gross Business Income can be used as auxiliary information, in order to obtain reliable small area estimates of total Wages and Salaries. In addition to the usual synthetic estimators proposed in the literature (Gonzalez 1973, Schaible 1979) composite estimators which are combinations of the synthetic estimator and the direct domain estimator are studied. In particular, the composite estimators proposed by Särndal (1981), and Fay and Herriot (1979) are investigated. Efficiences of the various proposed estimators relative to the direct estimators are obtained for Wages and Salaries via a simulation study in which the combined use of the two administrative files is mimicked using the Statistics Canada administrative file. This Statistics Canada file has the advantage of containing all the variables to be used for the small area estimation.

2.0 ESTIMATORS

Suppose that the population of size N consists of A mutually exclusive and exhaustive small areas labelled a=1,...,A. For each small area 'a', units are further classified into I mutually exclusive industrial groupings. Suppose that the area by industrial cross-classification can be further classified into G mutually exclusive and exhaustive income classes, labelled g=1,...,G. This labelling gives a three-way cross-classification into AIG cells with N_{aig} population members in the aig-th cell, with a corresponding sample count n_{aig} in a simple random sample of size n. For aggregation across a subscript, we replace that subscript by '.'; thus

$N_{a..} = \sum_{i=1}^{I} \sum_{g=1}^{G} N_{aig}$ is the population size for the a-th area. $N_{ai.} = \sum_{g=1}^{G} N_{aig}$ is the population size for the ai-th area by industry classification. Similarly, the sample aggregates $n_{a..}$ and $n_{ai.}$, are defined. The variable y will be used to denote Wages and Salaries while the variable x will denote the Gross Business Income.

For a particular sample, let t_m (ai) denote the estimate of total $Y_{ai..}$ for the a-th area and i-th industrial grouping for the m-th method of estimation. The various estimators for totals are next discussed.

2.1 Direct Estimators

The expansion estimator (EXP) utilizes only the sample data in the small area and industrial classification. For the a-th small area and i-th industry it is given by:

$$t_1(ai) = \frac{N}{n} \sum_{g=1}^{G} \sum_{k=1}^{n_{aig}} y_{aigk} = \frac{N}{n} y_{ai..} \qquad (2.1)$$

where y_{aigk} is the value of the k-th sampled unit within the aig-th cross-classification. The estimator t_1 (ai) is unbiased for $Y_{ai..}$, the population total for the (a,i)-th classification.

The expansion estimator can be improved using the known population domain sizes and observed sample domain sizes. This post-stratified estimator (POS) is given by:

$$t_2(ai) = \frac{N_{ai.}}{n_{ai.}} y_{ai..} \qquad (2.2)$$

for $n_{ai.} \geq 1$ and defined to be zero arbitrarily for $n_{ai.} = 0$. The estimator $t_2(ai)$ is unbiased for $Y_{ai..}$ if the probability of getting $n_{ai.} = 0$ is zero.

Estimators which use auxiliary information, such as counts or totals provided by administrative files, will be classified into (a) synthetic estimators, (b) generalized regression estimators, and (c) mixtures of synthetic and direct estimators.

2.2 Synthetic Estimators

For synthetic estimators, it is assumed that small area population means or proportions, for a given characteristic obtained across areas and for given subgroups of the population, are approximately equal to the over-all mean. The size of bias of the synthetic estimators will depend on the departure from this assumption. The problems associated with synthetic estimators have been well documented by Gonzalez (1973), Gonzalez and Hoza (1978), Levy (1971) and Schaible (1979). Two types of synthetic estimators are studied: the count-synthetic and the ratio-synthetic. The count-synthetic estimator (COUNT SYN) requires the knowledge of N_{aig}, the counts for a given small area, industrial grouping and income class cross-classification, obtained from the larger administrative file. It is given by:

$$t_3(ai) = \sum_{g=1}^{G} N_{aig}\, \bar{y}_{.ig.}, \qquad (2.3)$$

where

$$\bar{y}_{.ig.} = \left(\sum_{a=1}^{A} \sum_{k=1}^{n_{aig}} y_{aigk}\right) / \left(\sum_{a=1}^{A} n_{aig}\right)$$

$$= y_{.ig.} / n_{.ig}$$

is the over-all sample mean of Wages and Salaries for the i-th industrial grouping and g-th income class.

The ratio-synthetic estimator (RATIO SYN) requires totals of Gross-Business Income for a given small area, industrial grouping and income class cross-classification, X_{aig}. It is given by

$$t_4(ai) = \sum_{g=1}^{G} X_{aig.}\, (\bar{y}_{.ig.}/\bar{x}_{.ig.}) \qquad (2.4)$$

where $\bar{x}_{.ig.}$ is the overall sample mean of Gross Business Income for the i-th industry grouping and g-th income class.

2.3 Generalized Regression Estimators

Särndal (1981) proposed asymptotically design unbiased estimators that incorporate auxiliary information through the generalized regression technique (or Design/Model technique). In the two special cases included in our study, the estimators yielded by this technique (which can be adapted to any sampling design) are biased-corrected versions of the synthetic estimators (2.3) and (2.4).

In the special case of simple random sampling, the generalized regression estimator is of the form

$$t_{REG}(ai) = \sum_{g=1}^{G} \sum_{k=1}^{N_{aig}} \hat{y}_{aigk}$$

$$+ \frac{N}{n} \sum_{g=1}^{G} \sum_{k=1}^{n_{aig}} \hat{e}_{aigk}, \qquad (2.5)$$

where $\hat{y}_{aigk} = \sum_{j=1}^{p} \hat{b}_j\, z_{aigjk}$ is the predicted value of y_{aigk} resulting from the fit of a regression model of the form

$$y_{aigk} = \sum_{j=1}^{p} b_j\, z_{aigjk} + \varepsilon_{aigk}$$

with error term ε_{aigk},

and $\hat{e}_{aijk} = y_{aigk} - \hat{y}_{aigk}$

are the residuals. Here z_{aigjk} is the value for the j-th auxiliary variable (j=1,...,p) on the k-th unit in the (aig)-th cell. Estimates $\hat{b}_j$ may be obtained using generalized least squares procedures taking into account the distribution assumptions behind ε_{aijk} and the sample design weights.

The generalized regression estimator corresponding to the model

$$y_{aigk} = b_{ig} + \varepsilon_{aigk}$$

$$E(\varepsilon_{aigk}) = 0,\ V(\varepsilon_{aigk}) = \sigma^2_{ig}$$

(the ε's are assumed independent throughout) will be referred to as REG COUNT and is given by

$$t_5(ai) = \sum_{g=1}^{G} \{N_{aig}\, \bar{y}_{.ig.}$$

$$+ \frac{N}{n} n_{aig}\, (\bar{y}_{aig.} - \bar{y}_{.ig.})\} \qquad (2.6)$$

where $\bar{y}_{aig.} = y_{aig.}/n_{aig.}$

The generalized regression estimator corresponding to the model

$$y_{aigk} = b_{ig}\, X_{aigk} + \varepsilon_{aigk},$$

$$E(\varepsilon_{aigk}) = 0,\ V(\varepsilon_{aigk}) = \sigma^2_{ig}\, x_{aigk}$$

will be referred to as REG RATIO and is given by

$$t_6(ai) = \sum_{g=1}^{G} \{X_{aig.} \frac{\bar{y}_{.ig.}}{\bar{x}_{.ig.}} + \frac{N}{n} n_{aig}$$

$$(\bar{y}_{aig.} - \frac{\bar{y}_{.ig.}}{\bar{x}_{.ig.}}\, \bar{x}_{aig.})\} \qquad (2.7)$$

Note that the synthetic estimators (2.3) and (2.4) appear as the first terms of (2.6) and (2.7) respectively. In other words, the estimators $t_5(ai)$ and $t_6(ai)$ correct the bias in the count-synthetic and ratio-synthetic estimators, respectively. Rao (1984) noted that these estimators can be expressed as a convex combination of the direct and synthetic estimators.

2.4 Mixtures of Direct and Synthetic Estimators

Since estimators t_3 (ai) and t_4 (ai) do not use the small area means $y_{aig.}$ directly, it is natural to look for estimators that are weighted averages of t_3 (ai) or t_4 (ai) with t_1 (ai). The optimal composite estimator of this form is given by

$$t_{opt}\ (ai) = c\ t_m\ (ai) + (1-c)\ t_1\ (ai) \quad (2.8)$$

where (m=3,4) and the optimal weight c is obtained by minimizing the MSE (t_{opt} (ai)), (Schaible et al, (1979)). The estimation of c from sample data, however, is unreliable due to difficulties in estimating MSE of the biased synthetic estimators t_3 (ai) or t_4 (ai).

The empirical Bayes approach is an alternative to the above-mentioned methods. It provides sample-based weights that reflect the uncertainty of a linear regression fit over small area means. This method was applied by Fay and Herriot (1979) as a means to estimate income for small places in the U.S.A.

The empirical Bayes approach can be summarized as follows. Suppose that $\bar{y}_{ai.} \mid \bar{Y}_{ai.} \sim$ ind N $(\bar{Y}_{ai.}, D_{ai})$ and $\bar{Y}_{ai.} \sim$ ind N $(\underset{\sim}{\bar{X}}_{ai.}\ \underset{\sim}{b}_i, U_i)$, where $\bar{Y}_{ai.}$ is the population mean in the a-th area and i-th industrial grouping, $\underset{\sim}{\bar{X}}_{ai.} = (\bar{X}_{ail}, \ldots, \bar{X}_{aip})$ is the 1Xp vector of population means of auxiliary variables in the ai-th cell and $\underset{\sim}{b}_i$ is the pX1 vector of regression parameters associated with the i-th industrial grouping, and U_i measures the uncertainty in the linear fit to $Y_{ai.}$. The sampling variances D_{ai} are assumed to be known, but U_i is estimated from the marginal distribution of $\bar{y}_{ai.}$ by solving the following nonlinear equation in U_i:

$$\sum_{a=1}^{A} (\bar{y}_{ai.} - \bar{y}^*_{ai.})^2 / (U_i + D_{ai}) = A-p \quad (2.9)$$

where $\bar{y}^*_{ai.} = \underset{\sim}{\bar{X}}_{ai.} (\underset{\sim}{X}_i^T \underset{\sim}{V}_i^{-1} \underset{\sim}{X}_i^{-1}) \underset{\sim}{X}_i^T \underset{\sim}{V}_i^{-1} \underset{\sim}{\bar{y}}_i$

and $\underset{\sim}{V}_i$ is a diagonal matrix with a-th diagonal element $v_{ai} = D_{ai} + U_i$, $\underset{\sim}{\bar{y}}_i = (\bar{y}_{1i.}, \ldots, \bar{y}_{ai.})^T$. The resulting estimator of U_i is denoted by $\hat{U}_i$.

The empirical Bayes estimator (EB) of $Y_{ai.}$ is given by:

$$t_7\ (ai) = N_{ai} \left\{ \frac{\hat{U}_i}{\hat{U}_i + D_{ai}}\ \bar{y}_{ai.} + \frac{D_{ai}}{\hat{U}_i + D_{ai}}\ \bar{y}^*_{ai.} \right\}. \quad (2.10)$$

Efron and Morris (1971, 1972) suggested a modification of t_7(ai) since the latter could perform poorly for some individual components (ai). The modification uses a restricted estimator (EB/M) given by:

$$t_8(ai) = t_6(ai) \quad \text{if } t_1(ai) - d \le t_7(ai) \le t_1(ai) + d$$

$$= t_1(ai) - d \text{ if } t_7(ai) < t_1(ai) - d$$

$$= t_1(ai) + d \text{ if } t_7(ai) > t_1(ai) + d \quad (2.11)$$

where $d = (N_{ai.}\ D_{ai})^{1/2}$ is usually used.

Using the empirical Bayes technique, it must be noted that the computation of the sampled-based weights is complex. Consequently, it may be difficult to evaluate their design bias and design variance by analytical methods like the ones provided by Särndal. For this reason, Monte Carlo simulation is a convenient route to study the properties of different methods for small area estimation.

2.5 Variance Estimation

Estimates of variance for the synthetic estimators t3(ai) and t_4(ai) can be readily provided. However, since their mean square error can be much larger than the variance, no variance expressions for these estimators will be given.

For the expansion estimator t_1(ai) and the regression estimators t_5(ai) and t_6(ai), the form for the estimator of variance is:

$$v[t_m(ai)] = \frac{N(N-n)}{n(n-1)} \left[n_{ai.}(n_{ai.}-1)\ s^2_{ai} + \frac{n_{ai.}}{n} \left(1 - \frac{n_{ai.}}{n}\right) z^2_{ai.} \right]; \ (m=1,5,6) \quad (2.12)$$

where s^2_{ai} and $\bar{z}_{ai.}$ are the estimated domain variance and mean for the variable z_{aigk} and the variable z_{aigk} is given by y_{aigk} for t_1(ai), $y_{aigk} - \bar{y}_{.ig.}$ for t_4(ai), and $y_{aigk} - \hat{b}_{ig}\, x_{aigk}$ for t_6(ai). For t_2(ai), the variance expression (2.12) does not have the '2 nd' term. For domains with no sample units, we have defined t_2(ai) to be equal to zero.

For the empirical Bayes estimator, an estimator of variance is given by

$$v[t_7(ai)] = N^2_{ai.} \left\{ \left[P_{ai} + (1-P_{ai})\ w_{aai} \right]^2 D_{ai} + (1-P_{ai})^2 \sum_{b(\neq a)} w^2_{abi}\ D_{bi} \right\}, \quad (2.13)$$

where $P_{ai} = \hat{U}_i/(\hat{U}_i + D_{ai})$ and w_{abi} is the

ab-th element of $X_{\sim i}\,(X_{\sim i}^T V_{\sim i}^{-1} X_{\sim i})^{-1} X_{\sim i}^T V_{\sim i}^{-1}$.
The variance estimator (2.13) is obtained by treating the $\hat{U}_i$ as fixed numbers.

DESCRIPTION OF THE SIMULATION STUDY

In order to study the properties of the various estimators discussed in the previous section, a simulation was undertaken. This simulation mimicked the use of administrative data arising from several sources and their subsequent combination to yield small area estimates. Since the Statistics Canada administrative file had all the required information, it was used as the file for drawing the samples required for the simulation.

The province of Nova Scotia was chosen as the population of tax filers for which the simulation would be undertaken for several reasons. Firstly, we have a sufficient number of observations (1678) in the population of unincorporated tax filers whose Gross Business Income belonged to the range of $25 K to $500 K to carry out a simulation which could be handled in terms of computer time. Secondly, we have a sufficiently good span of correlations between Wages and Salaries, and Business Income between the various major industrial groupings, to assess the use of Business Income as an auxiliary variable. The small areas of interest were the 18 Census Divisions within Nova Scotia. The major industrial groups studied within these areas were Retail (515 units in the population), Construction (496 units in the population), Accommodation (114 units in the population) and the remaining industries grouped into Others (553 units in the population). The relative sub-domain sizes (Census Division by major industrial group classification) varied between 0.06% to 6.79%.

For the direct, synthetic and generalized regression estimators, we have considered two procedures: (1) G=3 income classes given by $25 K - $50 K, $50 K - $150 K, and $150 K - $500 K; (2) G=1 given by $25 K - $500 K. For the empirical Bayes estimation procedure, only the $25 K - $500 K income class was considered. 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. For the empirical Bayes procedure, the regression fit between Wages and Salaries and Gross Business Income was done within each major industrial grouping, and an intercept term was allowed to enter into the model. Two versions of the empirical Bayes estimator were obtained: (i) For one (EB/S) the sample estimate of variance, D_{ai}, for each subdomain mean of Wages and Salaries was used, (ii) for the other (EB/P), the population variance, D_{ai}, for each subdomain mean of Wages and Salaries was used. The ratio of MSE for versions (i) and (ii) provides a measure of increase in the MSE due to estimating D_{ai}. In addition, for both versions, the restricted estimator given by (2.11) was also computed: those modified versions are denoted as EB/SM for (i) and EB/PM for (ii). Empirical Bayes estimators could not be computed for cells with fewer than 2 observations; for these cells the REG RATIO estimator with three income was used. This modification is labelled as NEB.

For the Monte Carlo simulation we selected 500 samples, each of size 429, from the target population of 1,678 companies (unincorporated) in Nova Scotia. The expected number of sample observations in a subdomain ranged from 0.25 for the smallest to 29.3 for the largest. The main findings are discussed with respect to (a) relative percentage bias of estimators; (b) relative percentage efficiency; (c) relative percentage bias of the variance estimators; (d) coverage rate of confidence intervals; (e) coefficient of variation measures for the various estimators. The relative percentage bias of $t_m(ai)$ is computed as

$$\overline{RB}\left[t_m(ai)\right] = \frac{1}{A}\sum_{a=1}^{A}\left|\frac{\bar{t}_m(ai) - Y_{ai..}}{Y_{ai..}}\right| X\ 100$$

$$= \frac{1}{A}\sum_{a=1}^{A}\frac{\left|\bar{B}\left[t_m(ai)\right]\right|}{Y_{ai..}}\ X\ 100$$

by averaging over the small areas where $\bar{t}_m(ai) = \sum_{r=1}^{500} t_m^{(r)}(ai)/500$, and $t_m^{(r)}(ai)$ is the value of $t_m(ai)$ for the r-th Monte-Carlo sample (r=1,...,500) and $Y_{ai..}$ is the (known) population total for the ai-th cell.

The relative percentage efficiency of $t_m(ai)$ with respect to the direct estimator EXP is computed as

$$\overline{Eff}\left[t_m(ai)\right] = \frac{1}{A}\sum_{a=1}^{A}\left\{\frac{\overline{MSE}\left[t_1(ai)\right]}{\overline{MSE}\left[t_m(ai)\right]}\right\}^{1/2} X\ 100,$$

m=1, ..., 8 where $\overline{MSE}\left[t_m(ai)\right] = \sum_{r=1}^{500}$

$$\left[t_m^{(r)}(ai) - Y_{ai..}\right]^2 / 500,\ m=1,\ \ldots,\ 8$$

is the Monte Carlo approximation of the MSE of $t_m(ai)$.

The relative percentage bias of the variance estimator $v\left[t_m(ai)\right]$ is given by

$$\overline{RB}\ v\left[t_m(ai)\right] = \frac{1}{A}\sum_{a=1}^{A}\left\{\frac{\bar{v}\left[t_m(ai)\right]}{\overline{MSE}\left[t_m(ai)\right]} - 1\right\} X\ 100$$

where $\bar{v}\left[t_m(ai)\right] = \sum_{r=1}^{500} v^{(r)}\left[t_m(ai)\right] / 500.$

The confidence coverage rate for the estimators $t_m(ai)$, m=1,2,5,6,7 is evaluated as

$$\bar{P}_m(i) = \sum_{a=1}^{A}\sum_{r=1}^{500} I_m^{(r)}(ai) / (500A),\ \text{where}$$

$I_m^{(r)}(ai)$ = 1 if the 100 (1 - α)%

confidence interval given by

$t_m^{(r)}(ai) \pm z_{\alpha/2} \{v^{(r)}|t_m(ai)|\}^{\frac{1}{2}}$ contains

the true total $Y_{ai..}$, and zero otherwise. Here, $v^{(r)}|t_m(ai)|$ is the variance estimate of $t_m(ai)$ for the r-th Monte-Carlo sample and $z_{\alpha/2}$ is the upper $\alpha/2$ - point of N(0, 1) - variate.

A measure of average coefficient of variation is given by

$$\overline{cv}\,|t_m(ai)| = \frac{1}{A}\sum_{a=1}^{A} \frac{\{\overline{MSE}|t_m(ai)|\}^{\frac{1}{2}}}{Y_{ai..}} \times 100$$

Our empirical results, utilizing the above stated measures, are as follows:

(a) Bias of estimates. Table 1 examines the performance of the estimators $t_m(ai)$ in terms of percent relative bias. The unbiased estimator EXP, and the approximately unbiased estimators REG COUNT and REG RATIO show negligible relative bias (<5%), excepting that it is slightly higher for REG COUNT and REG RATIO in the case of Accommodation with G=3 (6.9 and 6.0 respectively). The latter may be due to the smaller number of observations in the 3 income classes (for Accommodation) which we used to estimate the bias. In the other cases, there is a very little difference in the bias for the 1 domain and 3 domains, for the COUNT SYN and RATIO SYN. The POS estimator has a large relative bias for "Acommodation" but this is due to a non negligible probability of getting no sampled units in the cell. The empirical Bayes estimators have significant relative (8% to 38%) bias with the most bias showing for the smallest industrial group in terms of observations, namely "Accommodation." As expected, both the synthetic estimators have the largest bias (as large as 80%) followed by the four empirical Bayes estimators.

(b) Relative Percentage Efficiency of estimators. All the estimators are significantly more efficient than the expansion estimator, EXP. The division of the income classes into 3 domains as opposed to 1 domain does not significantly improve the efficiency of the unbiased or the approximately unbiased estimators, except in the case of the REG COUNT estimator for Construction, Accommodation and Others, and the REG RATIO for Retails. The estimators using the auxiliary variable x, (RATIO SYN and REG RATIO) are substantially more efficient than those using only the counts (COUNT SYN and REG COUNT). This is especially true for the industrial grouping "Accommodation" where the correlation between Wages and Salaries and Gross Business Income is fairly high. The RATIO SYN is the most efficient, whereas REG RATIO (G=3) and the empirical Bayes estimators have comparable efficiency excepting that the latter was somewhat more efficient for "Retail." The difference in efficiency among the four empirical Bayes estimators is not significant. (See Table 2.)

Table 1: Percentage Relative Bias for the Estimator

Division	No. of income classes	Estimators					
		EXP	POS	COUNT SYN	RATIO SYN	REG COUNT	REG RATIO
Retail	1	2.3	8.5	20.7	32.4	1.1	1.6
Construction	1	1.9	5.4	17.3	15.7	1.3	0.9
Accommodation	1	3.6	26.5	58.4	41.4	3.4	2.8
Others	1	1.7	3.2	33.7	26.8	1.3	1.0
Retail	3	2.3	8.5	26.8	27.5	1.5	1.5
Construction	3	1.9	5.4	17.8	16.0	1.1	0.9
Accommodation	3	3.6	26.5	44.3	39.4	6.9	6.0
Others	3	1.7	3.2	27.7	26.6	1.4	1.0

Division	No. of income classes	NEB/S	NEB/SM	NEB/P	NEB/PM
Retail	1	17.6	17.3	18.9	17.2
Construction	1	11.6	10.5	10.5	8.3
Accommodation	1	38.0	36.5	36.7	31.4
Others	1	21.2	17.2	21.0	14.9

Table 2: Percentage Efficiency of the Estimator Relative to EXP.

Division	No. of income classes	Estimators				
		POS	COUNT SYN	RATIO SYN	REG COUNT	REG RATIO
Retail	1	1.35	2.33	2.00	1.52	1.30
Construction	1	1.35	2.56	3.03	1.54	2.13
Accommodation	1	1.40	1.75	3.45	1.30	2.78
Others	1	1.20	1.52	1.92	1.25	1.56
Retail	3	1.35	2.06	2.22	1.54	1.49
Construction	3	1.35	2.76	2.94	2.00	2.08
Accommodation	3	1.40	2.86	3.45	2.13	2.78
Others	3	1.20	1.82	1.92	1.41	1.54

Division	No. of income classes	NEB/S	NEB/SM	NEB/P	NEB/PM
Retail	1	1.72	1.72	1.82	1.82
Construction	1	2.04	2.04	2.17	2.13
Accommodation	1	2.70	2.78	2.38	2.44
Others	1	1.45	1.54	1.52	1.54

(c) Bias of variance estimators. Among the unbiased or approximately unbiased procedures, POS displays the highest bias with the MSE being systematically underestimated, especially in the smallest industrial group, Accommodation. The bias associated with REG COUNT and REG RATIO is essentially neglibible, with the exception of significant negative bias (-11% and -15%) introduced in estimating the variance for "Accommodation" in the presence of three income domains. The empirical Bayes procedure shows a smaller bias in the estimated variance when the population variance D_{ai} is used instead of the sample variance $\hat{D}_{ai}$ in the estimation, but the underestimation is still high (-6% for Construction to -37% for Accommodation). (See Table 3.)

Table 3: Percent Relative Bias of the Variance Estimators.

Division	No. of income classes	Estimators			
		EXP	POS	REG COUNT	REG RATIO
Retail	1	0.6	-38.7	0.9	2.5
Construction	1	1.9	-39.4	1.7	0.05
Accommodation	1	-2.7	-66.6	0.3	-4.0
Other	1	2.8	-33.6	2.4	2.4
Retail	3	0.6	-38.7	1.0	0.8
Construction	3	1.9	-39.4	1.3	-1.5
Accommodation	3	-2.7	-66.6	-11.1	-14.6
Others	3	2.8	-33.6	1.2	1.0
		EB/S			EB/P
Retail	1	-29.8			-19.5
Construction	1	-30.8			- 5.9
Accommodation	1	-51.8			-37.4
Others	1	-59.4			-34.5

d) Coverage rates. The coverage rates for the confidence intervals are shown in Table 4 for the estimators that have variance estimators are associated with them, for nominal levels of 90% and 95%. All the coverage rates fall short of their desired nominal level. The differences between the 1 domain and 3 domains cases for EXP, POS, REG COUNT and REG RATIO are small. The coverage rates for EXP, REG COUNT and REG RATIO are approximately equal, and range from 79% to 85% (nominal 90%). The post-stratified estimator (POS) falls significantly short of its nominal level, (as low as 68.8% compared to the nominal 90% in others, for instance). The coverage rates for the empirical Bayes procedure are much less than their nominal levels (as low as 18% for Accommodation compared to nominal 90%), implying that the associated variance formula (2.13) is not satisfactory. The coverage rate, however, is somewhat improved (28% to 67% for nominal level 90%) when the population variance D_{ai} is used (EB/P) instead of the estimated variance $\hat{D}_{ai}$ (EB/S). We are at present, exploring alternative variance estimators such as the jack-knife.

Table 4: Percent Coverage Rates for the Estimates.
Nominal Levels 90% and (in brackets) 95%.

Division	No. of income classes	Estimation			
		EXP	POS	REG COUNT	REG RATIO
Retail	1	84.0(87.3)	78.5(82.2)	85.1(89.6)	85.3(91.3)
Construction	1	82.3(85.1)	74.3(78.9)	81.3(86.9)	82.9(87.8)
Accommodation	1	81.9(84.6)	73.5(74.5)	81.3(84.6)	77.9(81.3)
Others	1	77.4(80.4)	68.8(73.7)	80.8(85.6)	79.7(84.2)
Retail	3	83.4(87.3)	77.7(82.2)	84.8(88.3)	86.2(87.9)
Construction	3	80.5(85.1)	73.6(78.9)	80.3(85.4)	80.3(86.9)
Accommodation	3	79.3(84.6)	72.0(74.5)	78.5(80.9)	79.0(81.1)
Others	3	80.0(80.4)	70.9(73.7)	81.2(83.5)	80.9(82.8)
		EB/S			EB/P
Retail	1	61.7(66.0)			66.9(72.0)
Construction	1	54.5(58.7)			66.6(71.5)
Accommodation	1	17.7(20.0)			28.3(32.0)
Others	1	32.7(36.1)			43.2(46.1)

(e) Coefficient of variation measure. Table 5 presents the values of the coefficient of variation measure for the different estimators in the four industry groups. Using the expansion estimator as the standard against which the performance of the others is measured, it is evident that all the other estimators reduce the error in the estimation. On the basis of this measure, RATIO SYN is the best in all the four industry groups followed by empirical Bayes. The REG RATIO has a somewhat higher coefficient of variation measure than the empirical Bayes ones. The estimators using the x-auxiliary information have a significantly smaller coefficient of variation measure than those using the counts.

Table 5: Percent Coefficient of Variation Measure for the Estimators.

Division	No. of income classes	Estimators					
		EXP	POS	COUNT SYN	RATIO SYN	REG COUNT	REG RATIO
Retail	1	60	44	23	35	38	52
Construction	1	55	41	20	18	35	25
Accommodation	1	101	71	66	44	89	65
Others	1	58	50	36	28	47	36
Retail	3	60	44	29	29	40	44
Construction	3	55	41	20	18	27	26
Accommodation	3	101	71	52	44	78	66
Others	3	58	50	30	29	41	37

Division	No. of income classes	NEB/S	NEB/SM	NEB/P	NEB/PM
Retail	1	35	35	33	33
Construction	1	25	25	23	24
Accommodation	1	51	51	53	53
Others	1	37	35	36	36

CONCLUSIONS

Our study confirms the results obtained by Särndal and Rabäck (1983), as far as their approximately unbiased procedures (REG COUNT and REG RATIO) are concerned, viz., the use of auxiliary information can be used to advantage to produce estimators for small areas with calculable variance estimates. For these approximately unbiased estimates, coverage rates fall short of their nominal level, especially for small domains. The RATIO SYN is the most efficient estimator in terms of MSE, followed by empirical Bayes and the REG RATIO.

In terms of bias of the estimates, coverage rates of the confidence interval and bias of the variance estimates, the REG RATIO appears superior to the empirical Bayes estimator and the post-stratified estimator. However, improved variance estimates (confidence intervals) for the empirical Bayes, such as bootstrap, jack-knife, and the variance estimator proposed by Morris (1983) need to be examined.

Further work on empirical Bayes procedures under the model appropriate for REG RATIO (section 2.3) are currently being eveloped along the lines of Battese and Fuller (1984). The performance of the procedures, conditionally given the domain sample sizes, is also under investigation.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Terry Gigantes, Director-General, for initiating and supporting this project. We are also most thankful to Gerry Horner and Conrad Bordeleau for coding the tax file to the Census Division level. The computer programs for the simulation were written in SAS by Claude Vaillancourt with the aid of George Kriger.

REFERENCES

Battese, G.E. and Fuller, W.A. (1984). An error Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Survey Section, Statistical Laboratory, Iowa State University, Ames.

Efron, B., and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators - Part I: The Bayes case. J. Amer. Statist. Assoc., 66, 807-815.

Efron, B., and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators - Part II: The empirical Bayes case. J. Amer. Statist. Assoc., 67, 130-139.

Fay III, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James Stein procedures to census data. J. Amer. Statist. Assoc., 74, 405-410.

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

Gonzalez, M.E. and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimates. Amer. Statist. Assoc. 73. 7-15.

Levy, P. (1971). The use of mortality data in evaluating synthetic estimates. 1971 Proceedings of the Social Statistics Section, American Statistical Association, 328-331.

Morris, C.N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. J. Amer. Statist. Assoc., 78, 47-55.

Rao, J.N.K. (1984). Some thoughts on small area estimation. Statistics Canada report.

Särndal, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for non-response. Bull. Int. Stat. Ind., 49:1 494-513 (Proceedings, 43rd Session, Buenos Aires).

Särndal, C.E. and Räbäck, G. (1983). Variance Reduction and Unbiasedness for Small Domains Estimators. Statistical Review, 5, 33-40.

Schaible, W.L. (1979). A composite estimator for small area statistics. In Synthetic Estimates for Small Areas (C. Steinberg, ed.) pp. 36-53. National Institute on Drug Abuse Research Monograph 24. U.S. Government Printing Office, Washington, D.C.

For additional information, see also

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. 18-25 August.