BENCHMARKING OF SMALL AREA ESTIMATORS

H.J. Mantel, A.C. Singh, and M. Bureau, Statistics Canada
H.J. Mantel, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6

KEY WORDS: Direct and indirect estimators, composite estimators, bias and MSE

1. Introduction

There exists a considerable body of research on small area estimation using cross-sectional survey data in conjunction with supplementary data obtained from census and administrative sources. A good collection of papers on this topic can be found in Platek, Rao, Särndal and Singh (1987). For large areas (or domains) direct estimators (*i.e.* estimators based only on sample data from the area of interest) are often used; however, indirect estimators, in which strength is borrowed from similar areas via a model containing auxiliary variables from the supplementary data, are often used for small areas. For repeated surveys it may also be beneficial to borrow strength over time; see Pfeffermann and Burck (1990) and Singh and Mantel (1991). Direct small area estimators, though (approximately) unbiased, are not reliable because of high variance. Indirect small area estimators are more reliable, though they may be somewhat biased.

A common problem in the application of small area techniques is that the individual small area estimates within a larger area do not add up to the direct estimator for the larger area. This problem can be resolved by benchmarking of the small area estimators with respect to the direct estimator for the larger area. This is desirable for at least three reasons: (*i*) the usual direct estimator for the larger area is approximately unbiased, whereas the aggregated small area estimators may be substantially biased, (*ii*) benchmarking gives rise to some robustification in that the average of the benchmarked small area estimators has good bias and variance properties, (*iii*) there will be internal consistency between published estimates for the larger area and the total of estimates of the individual small areas within it.

Three methods for benchmarking are proposed in the literature: (*i*) Battese, Harter and Fuller (1988) distribute the difference between the direct large area estimator and the sum of the small area estimators in proportion to the mean squared error (MSE) of each small area estimator. (*ii*) Pfeffermann and Barnard (1991) distribute the difference "optimally" using the full MSE matrix of the small area estimators. This method has an advantage for time series methods in that it can be built in as part of the Kalman filter algorithm (giving as a byproduct an estimate of the MSE matrix of the benchmarked estimators); see Pfeffermann and Burck (1990). (*iii*) Rao and Choudhry (1993) distribute the difference in proportion to the small area estimates, *i.e.* a simple ratio (or raking) adjustment is made.

In this paper we perform an empirical study using a synthetic population based on data from Statistics Canada's Survey of Employment, Payroll and Hours (SEPH) to compare the effect of benchmarking on various small area estimators. In particular, we compare, in a repeated sampling framework, the loss in efficiency due to benchmarking to the gain in efficiency due to "borrowing strength". Two types of indirect small area estimators are synthetic (in which small areas are assumed to be like a larger area) and composite (convex combinations of direct and synthetic estimators). For small area estimation we consider three types of composite estimators where the weights for the convex combination can be either (*i*) optimal (*i.e.* based on a correctly specified model), (*ii*) pseudo-optimal (*i.e.* based on an incorrect model), or (*iii*) based on some other working convention such as the one for sample size dependent weights.

2. Domain Estimation Methods

Let the vector of small area population totals, Y_a, $a = 1, \ldots A$, be denoted by $\underline{Y}$. Here we define briefly some well known small area estimators which we will use in our simulation study. Rao (1986), Särndal and Hidiroglou (1989) and Pfeffermann and Burck (1990) also contain a good survey of various small area estimators.

2.1. Direct Estimators

2.1.1 Expansion estimator

This method of estimation is defined by $\mathrm{EXP}_a = \sum_{i \in s_a} w_i y_i$ where s_a is the portion of the sample falling in small area a, and w_i is the survey weight for unit i. For stratified simple random sampling, which we use in our simulation study, we have

$$\mathrm{EXP}_a = \sum_k (N_k/n_k) \sum_{i \in s_{ka}} y_i \ , \qquad (2.1)$$

where s_{ka} denotes the set of n_{ka} sample units falling in the small area a and stratum k and n_k, N_k denote

respectively the sample and population sizes for the kth stratum. The above estimator is often unreliable because the random sample size n_{ka} may be small in expectation and could have high variability. Conditional on the realized sample size n_{ka}, $\mathbf{EXP}_a$ is biased; however, unconditionally, it is unbiased for Y_a.

2.1.2 Separate ratio estimator

If X_{la}, the small area total of a suitable covariate, is known for some post-strata indexed by l, then the efficiency of the estimator $\mathbf{EXP}_a$ could be improved upon by exploiting this knowledge. We define

$$\mathbf{SRAT}_a = \Sigma_l X_{la} \hat{Y}_{\exp,la} / \hat{X}_{\exp,la}, \qquad (2.2)$$

where $\hat{Y}_{\exp,la}$ is the expansion estimator for the total of y in small area a by post-stratum l. In our simulation study later we take the post-strata to be the intersection of design strata with small areas. When the covariate x is a constant then the estimator, also called post-stratified and denoted by $\mathbf{POST}_a$, is both conditionally and unconditionally unbiased; however, $\mathbf{SRAT}_a$ would generally be slightly biased. These estimators may also not be sufficiently reliable because of the possibility of n_{ka}'s being small in expectation. If $\hat{X}_{\exp,la} = 0$, the above estimators are not defined. In practice, some ad hoc value such as 0 is often chosen for $\hat{Y}_{\exp,la} / \hat{X}_{\exp,la}$ when $\hat{X}_{\exp,la} = 0$. In the simulation study presented in this paper, we set $\hat{Y}_{\exp,la} / \hat{X}_{\exp,la} = \hat{Y}_{\exp,l} / \hat{X}_{\exp,l}$ whenever $\hat{X}_{\exp,la} = 0$.

2.1.3 Combined ratio estimator

An alternative to the separate ratio estimator is the combined ratio estimator,

$$\mathbf{CRAT}_a = X_a \mathbf{EXP}_a / \hat{X}_{\exp,a} \qquad (2.3)$$

When the covariate x_i is a constant then the estimator will be denoted by $\mathbf{HAJEK}_a$. $\mathbf{CRAT}_a$ would generally be slightly biased. If $\hat{X}_{\exp,a} = 0$ then the above estimators are not defined. In practice, some ad hoc value such as 0 is often chosen for $\mathbf{EXP}_a / \hat{X}_{\exp,a}$ when $\hat{X}_{\exp,a} = 0$. In our simulation study presented later, we set $\mathbf{EXP}_a / \hat{X}_{\exp,a} = \hat{Y}_{\exp} / \hat{X}_{\exp}$ whenever $\hat{X}_{\exp,a} = 0$.

2.1.4 Generalized regression estimator (GREG)

In this method a linear regression model is assumed to relate the individual level variate values y_i to a vector of covariates x_i. These covariates would need to be known for each sampled unit and domain totals would also be required. The sample data can be used to estimate the regression parameter and a synthetic estimator of the domain totals is then constructed. However, there may be some local lack of fit of the global regression model and this is accounted for by a direct estimate of the domain sum of residuals from the regression. The estimator is

$$\mathbf{GREG}_a = x_a^T \hat{\beta} + N_a \bar{e}_a \qquad (2.4)$$

where $\hat{\beta} = (\Sigma_s (x_i x_i^T)/(v_i \pi_i))^{-1} (\Sigma_s (x_i y_i)/(v_i \pi_i))$, $\bar{e}_a = \hat{e}_{\exp,a} / \hat{N}_{\exp,a}$, $e_i = y_i - x_i^T \hat{\beta}$, x_a is the domain a total of the covariate vectors x_i, v_i are pre-specified regression weights and π_i is the survey weight for unit i. This version of generalized regression estimation, with a synthetic $\hat{\beta}$, was proposed by Särndal and Hidiroglou (1989). When the sample size in domain a is 0 we take $\bar{e}_a = 0$. $\bar{e}_a$ would be relatively stable when the regression model accounts for a large proportion of the variability in y.

2.2 Composite Estimators

2.2.1 Sample size dependent estimator

If the observed sample size in small area a is small then we may consider a convex combination of a direct estimator and a synthetic estimator (*e.g.* $x_a^T \hat{\beta}$ of (2.4)). Using sample size dependent weights, we have

$$\mathbf{SSD}_a = (1 - \lambda_a) \hat{Y}_{\mathrm{syn},a} + \lambda_a \hat{Y}_{\mathrm{dir},a} \qquad (2.5)$$

where $\lambda_a = 1$ if $\hat{N}_{\exp,a} \geq N_a$ and $\lambda_a = (\hat{N}_{\exp,a} / N_a)^d$ otherwise, and d is assigned some suitable value such as 1 or 2.

2.2.2 Empirical best linear unbiased estimator (EBLUP)

An alternative to sample size dependent smoothing of small area estimators is to use the empirical Bayes approach of Fay and Herriot (1979) or the more general best linear unbiased predictor (BLUP) approach (see *e.g.* Battese, Harter, and Fuller (1988), and Pfeffermann and Barnard (1991)). It is assumed that $\underline{Y} = F\underline{\alpha} + \underline{\nu}$ where the ν_as are small area effects and F is a matrix of regressors. The model for the small area estimators is then $\hat{\underline{Y}}_{\mathrm{dir}} = F\underline{\alpha} + \underline{\nu} + \underline{\varepsilon}$ where ε_a is an observation error term. The BLUP under this model is

$$\mathbf{BLUP} = \Lambda \hat{\underline{Y}}_{\mathrm{dir}} + (I - \Lambda) F \hat{\underline{\alpha}} \qquad (2.6)$$

where $\Lambda = V(V+W)^{-1}$, V and W are, respectively, the MSE matrices of $\hat{\underline{Y}}_{dir}$ and $F\hat{\underline{\alpha}}$, and $\hat{\underline{\alpha}}$ is the generalized least squares estimate of $\underline{\alpha}$. The mean squared error of **BLUP** is given by $V - V(V+W)^{-1}V$. The variance components V and W would need to be estimated, a survey based estimate would be used for V and then W would be estimated conditional on the estimated V using Henderson's method; more details are given in Section 3. When V and W are replaced by estimates the resulting estimator is termed empirical **BLUP** or **EBLUP**. When the model for the direct estimators is correctly specified the resulting estimator would be called optimal, otherwise it would be called pseudo-optimal.

2.3 Benchmarking

It is sometimes desirable that small domain estimators should add up to direct estimators for certain larger domains containing them. One simple possibility, presented by Rao and Choudhry (1993) is to make a ratio adjustments within each larger area. We will indicate this ratio adjusted constrained estimator by the prefix **CR_** (*e.g.* **CR_EBLUP** for the adjusted **EBLUP**). A second approach, following Pfeffermann and Barnard (1991), and which we will indicate by the prefix **CD_**, is based on the MSE (dispersion) matrix for the small area estimators. If the constraint is expressed as $L^T\underline{Y} = \underline{c}$, with $\underline{c}$ a fixed, known constant, then the minimum MSE linear unbiased estimator is

$$\hat{\underline{Y}} + \Gamma L(L^T\Gamma L)^{-1}(\underline{c} - L^T\hat{\underline{Y}}) \qquad (2.7)$$

where $\Gamma = \mathbf{MSE}(\hat{\underline{Y}})$. The third approach, suggested by Battese, Harter and Fuller (1988), and denoted by the prefix **CV_**, is given by (2.7) with the off diagonal elements of Γ set to zero.

3. Simulation Study

The methods described in Section 2 were compared empirically by means of a Monte Carlo simulation from a synthetic pseudo-population based on data from Statistics Canada's Survey of Employment, Payroll and Hours (SEPH). The SEPH sample is currently stratified by 1980 three digit standard industrial classification (SIC3) within province and four size classes; however, under a proposed redesign of the survey the sample will no longer be controlled at the SIC3 level, but rather at some aggregation of SIC3s such as SIC2. An objective of the research reported in this paper is to investigate methods for estimation at the SIC3 by province level after the redesign. Because the sample will no longer be controlled at the SIC3 level this is a domain estimation problem. Larger establishments, and those with a complex structure, are subject to higher sampling rates so that direct estimates at the SIC3 level are satisfactory. However, for smaller establishments (size strata 1 and 2) of simple structure (in what is called the non-integrated portion of the frame, NIP) small domain estimation techniques could be necessary for production of SIC3 by province level estimates. A covariate which can be used for these units is PD7 data which records monthly income tax payroll deductions submitted to Revenue Canada.

To construct the pseudo-population used in our study, we took sample data from the province of Ontario for SIC1=3 (industrial manufacturing and products) and the NIP portion of size classes 1 and 2. Variables included were the SIC3 code, the number of employees, the 3 month average PD7 remittance, the size classification, and the survey weight. We used this data to fit the model

$$y_{ijk} = x_{ijk}(\beta + \nu_i + \xi_{ij} + \varepsilon_{ijk})$$

where y_{ijk} is the number of employees for the kth unit in the jth SIC3 in the ith SIC2, x is the 3 month average PD7 remittance plus 500, β is fixed, and ν, ξ, and ε are independent random components. Using the survey weights as replicate weights, we expanded the pseudo-population, which had 995 distinct units, to 24,074 units. The pseudo-population contained 42 SIC3s (small areas) in 9 SIC2s (*e.g.* fabricated metal products industries, non-metallic mineral products industries). The small area population sizes varied from 26 to 14,236 units. We generated new numbers of employees from the fitted model, except that the estimated variance components were scaled down to reduce the problem of zeros in the data. We simulated sampling from this pseudo-population using stratified simple random sampling by size class and SIC2. The sample size for each stratum was taken to match the total SIC2 by size class in the SEPH sample, though the sampling fractions at the SIC3 level would differ from the SEPH sample. The expected sample size within small areas varied from 1.10 to 142.16 and averaged 23.69.

3.1 Estimation methods used in the study

All of the general estimation methods described in Section 2 were included in the study, with some particular features as described here. Since SIC3s are entirely contained in the corresponding SIC2, each

SIC3 crossed at most two of the design strata corresponding to the two size strata within the SIC2.

The estimators **EXP**, **POST**, **SRAT**, **HAJEK** and **POST** are exactly as described in Section 2.

The remaining unbenchmarked estimators were applied separately within each size stratum and all further discussion of them in this subsection should be taken as being within size classes.

For the GREG estimator, the parameter β has two components, one corresponding to a constant term, and the second corresponding to x_i, the PD7 remittance plus 1000 (to avoid the problem of 0 remittances). All sample data within the SIC1 were used in the estimation of β and v_i was taken to be x_i.

Two sample size dependent estimators are considered, both with $d=2$ and with the synthetic part being $x_a^T\hat{\beta}$, where $\hat{\beta}$ is defined as in Section 2.4. The first, which we denote by **SSD**, has the estimator **POST** as the direct part; the second, denoted by **SSD***, has **GREG** as the direct part. The estimator **SSD*** was proposed by Särndal and Hidiroglou (1989).

There are four versions of the **EBLUP** estimator considered, based on two direct estimators, **POST** and **GREG**, and two different models. Both models take the matrix F as including a column of 1's and a column of x_a's, the small area totals of x_i, where x_i is as for the **GREG** estimator. They differ in how they model the small area effects, v_a. In the first we model them as $v_a = x_a^{\gamma/2}(v_k + \xi_a)$ where x_a is the domain a total of x_i, v_k is a random effect that is common to all SIC3s within the same SIC2 k, and ξ_a is a random effect for SIC3 a. It was assumed that $v_k \sim (0,\sigma_v^2)$, $\xi_a \sim (0,\sigma_\xi^2)$, and all random effects and the observation errors ε_a are independent. The standard variance estimator for simple random sampling without replacement was used for the entries of V (which is diagonal, estimation of β for **GREG** was ignored in estimation of V). When the observed sample size in an SIC3 was 1 a synthetic estimator of the design variance based on data from the corresponding SIC2 was used, and when the observed sample size was 0 the MSE was taken as infinity. Taking the estimated V as the true value, the variance components σ_v^2 and σ_ξ^2 were then estimated using Henderson's method. We will denote the estimator based on this model and **POST** by **EBLUP2** and the estimator based on **GREG** by **EBLUP2***. In the second model we assume the variance component σ_v^2 to be zero. The estimator based on **POST** and this second model will be denoted by **EBLUP1**, and that based on **GREG** will be denoted by **EBLUP1***. Note that the estimators **EBLUP2** and **EBLUP2*** are optimal, in the sense that they are based on a correctly specified model, while **EBLUP1** and **EBLUP1*** are pseudo-optimal.

For the benchmarked estimators the benchmark was taken to be the estimator **EXP** at the SIC2 level. Ratio adjusted benchmarking was applied to all estimators. The two versions of MSE adjusted benchmarking were applied to the estimators **EBLUP2*** and **EBLUP1***, but not to any other estimators because of problems with estimated MSE matrices being singular. The MSE matrices of the **EBLUP** estimators were estimated by the "naive" estimator, *i.e.* $V - V(V+W)^{-1}V$ with V and W replaced by estimates.

3.2 Evaluation Measures

Suppose m simulations are performed in which m_1 sets of different vectors of realized sample sizes for SIC3s by strata are replicated m_2 times. The following measures can be used for comparing performance of different estimators. Let i vary from 1 to m_1 and j from 1 to m_2.

(i) Absolute Relative Bias.

$$\mathbf{ARB}_a = |m^{-1}\Sigma_i\Sigma_j(\mathrm{est}_{ija} - \mathrm{true}_a)/(\mathrm{true})_a| \quad (3.1)$$

The average of $\mathbf{ARB}_a$ over domains a will be denoted by **AARB**.

(ii) Root Mean Square Conditional Relative Bias.

$$\mathbf{RMSCRB}_a = \{m_1^{-1}\Sigma_i(m_2^{-1}\Sigma_j \mathrm{est}_{ija} - \mathrm{true}_a)^2/\mathrm{true}_a^2 - B\}^{1/2} \quad (3.2a)$$

$$B = m^{-1}(m_2-1)^{-1}\Sigma_i[\Sigma_j \mathrm{est}_{ija}^2 - (\Sigma_j \mathrm{est}_{ija})^2/m_2]/\mathrm{true}_a^2 \quad (3.2b)$$

The correction term B adjusts for bias in the first term due to m_2 being finite. **ARMSCRB** will denote the average of $\mathbf{RMSCRB}_a$ over areas a.

(iii) Mean Absolute Relative Error.

$$\mathbf{MARE}_a = m^{-1}\Sigma_i\Sigma_j|\mathrm{est}_{ija} - \mathrm{true}_a|/\mathrm{true}_a \quad (3.3)$$

and **AMARE** denotes the average of $\mathbf{MARE}_a$ over domains a.

(iv) Relative Root Mean Square Error.

$$\mathbf{RRMSE}_a = \{m^{-1}\Sigma_i\Sigma_j(\mathrm{est}_{ija} - \mathrm{true}_a)^2\}^{1/2}/\mathrm{true}_a \quad (3.4)$$

and **ARRMSE** as before denotes the average over domains.

The precision (i.e. the Monte Carlo standard error) of each measure depends on m_1, m_2. It can be seen that for all measures except (ii), the optimal choice of m_1, m_2 under the restriction that $m_2 > 1$ is $m_1 = m/2$, $m_2 = 2$, since this minimizes the Monte Carlo standard error. For the second measure, the appropriate choice of m_1, m_2 is less straightforward. For our simulation study we set $m_1 = 5000$, $m_2 = 2$.

3.3 Empirical Results

Figures 1 to 5 display the average evaluation measures from the Monte Carlo simulations for most of the estimators included in the study.

Figure 1 shows evaluation measures for unbenchmarked direct estimators. Clearly use of the covariate has a very beneficial effect in this example, as would be expected because of the model used to generate the data. The estimator **POST** is best among those which do not use the covariate, while **SRAT** and **GREG** are both best among those using the covariate.

Figure 2 shows the effect of combining the **POST** and **GREG** estimators with a regression synthetic estimator and compares the three methods of composite estimation. Generally, composite estimation shows some improvement in the evaluation measures **AMARE** and **ARRMSE** and some deterioration in the bias measures (**AARB** and **ARMSCRB**), with the **EBLUPs** showing a stronger effect than the **SSDs**. In this study there is very little difference between the two **EBLUPs**. The performance of the pseudo-optimal estimators, **EBLUP1** and **EBLUP1***, is the same as that of the optimal estimators, **EBLUP2** and **EBLUP2***, respectively; however, see also Figure 5 and the discussion below.

Comparing Figure 3 to Figure 2 we see the effect of benchmarking. Generally the effect of benchmarking here is a slight improvement in the overall bias (**AARB**) at the cost of some deterioration with respect to the other evaluation measures. The relatively poor performance of the benchmarked estimators is not surprising since the benchmark **EXP** performs relatively poorly; see Figure 4. Benchmarking would be expected to improve performance only in the case of serious model breakdown.

Figure 5 compares the three different methods of benchmarking. For the estimator **EBLUP1*** all three methods perform about the same. For **EBLUP2*** the ratio adjusted benchmarking method performs as well as for **EBLUP1***; however, the MSE adjusted methods perform more poorly. A possible explanation is that, with the extra variance component in the model underlying **EBLUP2***, the estimate of the MSE of **EBLUP2*** is of poor quality.

Acknowledgement

We are grateful to Dave Binder, Dave Dolson, and Jon Rao for helpful discussions. The second author's research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

References

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error-components model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistics Association*, **74**, 269-277.

Pfeffermann, D., and Barnard, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economics Statistics*, **9**, 73-84.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, **16**, 217-237.

Platek, R., Rao, J.N.K., Särndal, C.E., and Singh, M.P. eds (1987). *Small Area Statistics: An International Symposium*; New York; John Wiley & Sons.

Rao, J.N.K. (1986). Synthetic estimators, SPREE and best model-based predictors of small area means. Technical Report, Laboratory for Statistics and Probability, Carleton University, Ottawa.

Rao, J.N.K., and Choudhry, G.H. (1993). Small area estimation: overview and empirical study. *Monograph proceedings of the International Conference on Establishment Surveys, Buffalo, June 1993*, to appear.

Särndal, C.E., and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, **84**, 266-275.

Singh, A.C., and Mantel, H.J. (1991). State space composite estimation for small areas. Proceedings of *Symposium 91: Spatial Issues in Statistics*, Statistics Canada, Ottawa, November 1991, 17-25.

Symbols used in figures

△ AARB – average absolute relative bias

○ ARMSCRB – average root mean squared conditional relative bias

▽ AMARE – average mean absolute relative error

□ ARRMSE – average relative root mean squared error

Figure 1: unbenchmarked direct estimators

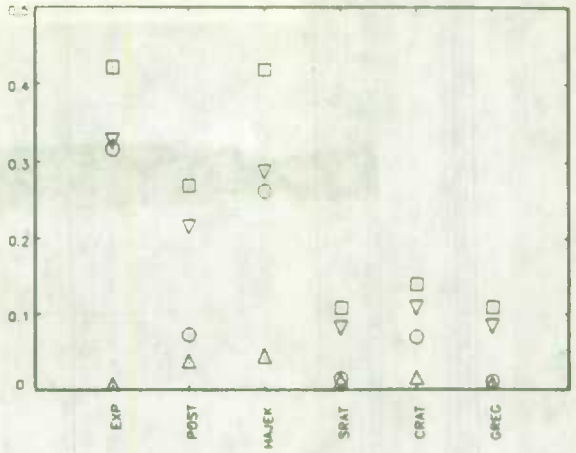

Figure 2: comparison of direct and composite estimators

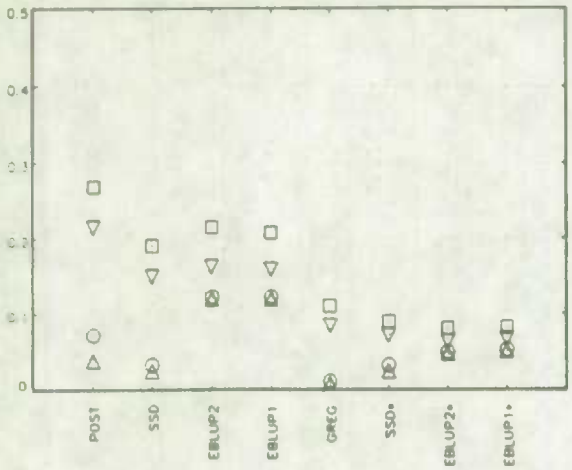

Figure 3: benchmarked direct and composite estimators

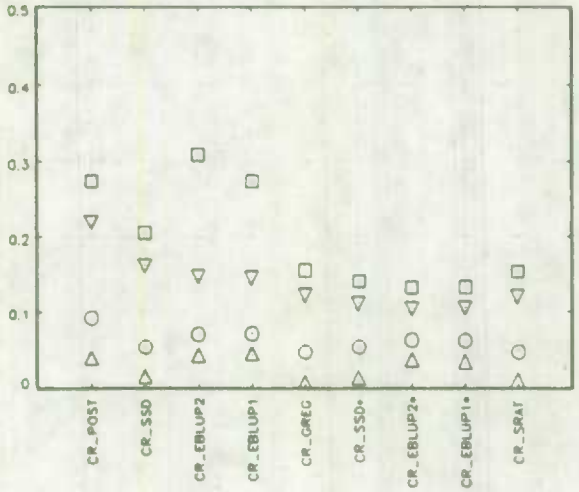

Figure 4: unbenchmarked estimators aggregated to large areas (SIC2 level)

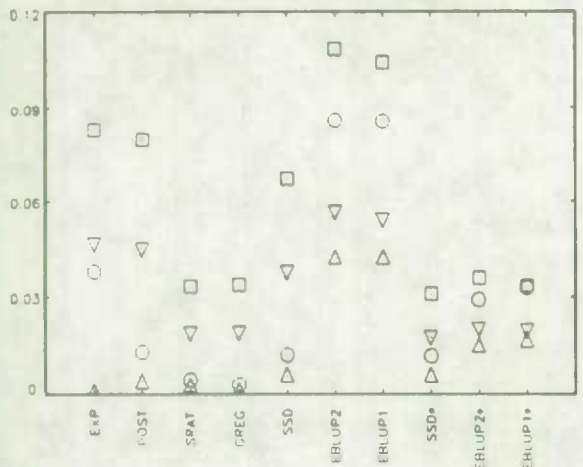

Figure 5: comparison of three benchmarking methods