

EVALUATION OF SMALL DOMAIN ESTIMATORS FOR THE SURVEY OF EMPLOYMENT PAYROLL AND HOURS

Susana Rubin Bleuer, Serge Godbout and Yves Morin¹

ABSTRACT

The Survey of Employment, Payroll and Hours provides monthly estimates of payroll, employment, paid hours and earnings. The proposed new design uses survey and payroll deduction (administrative) data to produce generalized regression (GREG) estimators for average weekly earnings, which are approximately unbiased and have a controlled cv for pre-determined strata. As in most surveys, there are many domains of interest with small sample size for which the GREG estimators might have unacceptable large measures of error. In this paper, we extend the Pseudo-EBLUP method of You and Rao (2002) to the case of unequal error variances, and we study the relative performance of several cross-sectional small domain estimators using real and synthetic populations.

KEY WORDS: Monte Carlo Mean Square Error, Pseudo EBLUP, Small Domain Estimation.

RÉSUMÉ

L'Enquête sur l'emploi, la rémunération et les heures de travail fournit des estimations mensuelles de la paie totale, de l'emploi, des heures travaillées et des gains. Le remaniement du plan de sondage proposé utilise des données d'enquête et un recensement de dossiers administratifs pour construire des estimateurs de régression généralisée (GREG) pour les gains hebdomadaires moyens. Ces estimateurs demeurent approximativement sans biais et ont un cv contrôlé pour des strates prédéterminées. Comme dans la plupart des enquêtes, il y a un grand nombre de domaines d'intérêt pour lesquels la faible taille d'échantillon peut entraîner une trop grande mesure d'erreur pour l'estimateur GREG. Dans cet article, nous appliquerons la méthode Pseudo-EBLUP de You et Rao (2002) dans le cas de variances inégales des erreurs et nous étudierons la performance relative de plusieurs estimateurs transversaux de petits domaines en utilisant des populations réelles et synthétiques.

MOTS CLÉS : Erreur quadratique moyenne par la méthode Monte Carlo; estimation pour les petits domaines; pseudo-EBLUP.

1. INTRODUCTION

The Canadian Survey of Employment, Payroll and Hours provides monthly estimates of s, employment, paid hours and earnings at detailed industrial & geography levels. The proposed new design uses survey and payroll deduction (administrative) data to produce generalized regression (GREG) estimators for average weekly earnings, which are approximately unbiased and have a controlled cv for pre-determined strata that are labelled as model groups. As in most surveys, there are many domains of interest with small sample size for which the GREG estimators will have unacceptably large coefficients of variation (cv). The domains of interest are the industry groups at the North American Industry Classification System (NAICS) level 4 and geography at the level of province, that is, the "NAICS 4 x province" domains, while the domains at which we establish the model (called model groups) are at the level of "NAICS 3 x Canada". The purpose of this study is to investigate the feasibility of producing reliable Small Area Estimators (SAE) for these domains.

The statistical unit is the establishment. The parameter of interest is AWE, the average weekly earnings in the establishment and is available for the sample respondents. The auxiliary variables are AME, the average monthly earnings and number of employees E in the establishment, are obtained from administrative data and are available for the whole

¹ Susana Rubin Bleuer, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, rubisus@statcan.ca

target population. Correlation between AWE and AME ranges between 0.2 and 0.8, because some industries have monthly earnings affected by extra pay periods and special payments.

2. SMALL AREA ESTIMATORS

2.1 Notation and Formulae

We denote by m the number of domains within a model group, with domain i containing N_i $i = 1, \dots, m$ establishments.

The Average Weekly Earnings for domain i is given by the weighted average $\bar{Y}_i = \sum_{k=1}^{N_i} E_{ik} AWE_{ik} / \sum_{k=1}^{N_i} E_{ik}$, where E_{ik} is the number of employees and $y_{ik} = AWE_{ik}$ is the average weekly earnings in establishment k of domain i , $i = 1, \dots, m$.

Though we are concerned with business data, the characteristic of interest for an establishment ik , $k = 1, \dots, N_i$, in area i , is itself a weighted mean, and hence the distribution of AWE in the population is not too skewed.

We compare the GREG estimator with various small area estimators, including the projection estimator currently used. Some of these estimators are based on the general GREG model with fixed effects:

$$y_{ik} = x'_{ik}\beta + \varepsilon_{ik}, \quad \varepsilon_{ik} \square (0, \sigma_\varepsilon^2 / E_{ik}) \quad (2.1)$$

where $x_{ik} = (1, AME_{ik})'$, where AME_{ik} is the average monthly earnings in establishment k in domain i , the unit errors ε_{ik} are independent, $k = 1, \dots, N_i$, $i = 1, \dots, m$ and $\beta = (\beta_0, \beta_1)'$ is a vector of regression parameters.

We assume that samples are drawn independently across the domains according to a specified sampling design with overall sample size n . The sample size in domain i is n_i $i = 1, \dots, m$. The sample weight for unit k in domain i is denoted by w_{ik} . We also assume that the sample data obey the population model (2.1). For each of the different estimation methods we describe below, let the fitted values be denoted by $\hat{y}_{ik} = x'_{ik}\hat{\beta}$, where $\hat{\beta}$ varies with the corresponding method.

Under model (2.1) we look at \hat{Y}_{GREGi} , the GREG estimator:

$$\hat{Y}_{GREGi} = \sum_{k=1}^{N_i} E_{ik} \hat{y}_k / \sum_{k=1}^{N_i} E_{ik} + \sum_{k=1}^{n_i} w_{ik} E_{ik} (y_{ik} - \hat{y}_{ik}) / \sum_{k=1}^{N_i} E_{ik}, \quad \hat{\beta}_{GREG} = \left(\sum_{i=1}^m \sum_{k=1}^{n_i} w_{ik} E_{ik} \mathbf{x}_{ik} \mathbf{x}'_{ik} \right)^{-1} \sum_{i=1}^m \sum_{k=1}^{n_i} w_{ik} E_{ik} \mathbf{x}_{ik} y_{ik}$$

as well as at \hat{Y}_{GREG^*i} or GREG*, the domain specific GREG (calibrated by the estimated number of employees rather than the estimated population size):

$$\hat{Y}_{GREG^*i} = \sum_{k=1}^{N_i} E_{ik} \hat{y}_k / \sum_{k=1}^{N_i} E_{ik} + \sum_{k=1}^{n_i} w_{ik} E_{ik} (y_{ik} - \hat{y}_{ik}) / \sum_{k=1}^{n_i} w_{ik} E_{ik}, \quad \hat{\beta} = \hat{\beta}_{GREG}$$

the projection (synthetic) estimator \hat{Y}_{PROJi} :

$$\hat{Y}_{PROJi} = \sum_{k=1}^{N_i} E_{ik} \hat{y}_{ik} / \sum_{k=1}^{N_i} E_{ik}, \quad \hat{\beta} = \hat{\beta}_{GREG},$$

and \hat{Y}_{PREDi} , the linear predicted estimator PRED (average of observed and predicted values, which is a modified ‘‘Brewer estimator’’ with $\hat{\beta}$ estimated with weights $\propto (w_{ik} - 1)E_{ik}$, see Särndal and Wright, 1984):

$$\hat{Y}_{PREDi} = \sum_{k=1}^{n_i} E_{ik} y_{ik} / \sum_{k=1}^{N_i} E_{ik} + \sum_{k=n_i+1}^{N_i} E_{ik} \hat{y}_{ik} / \sum_{k=1}^{N_i} E_{ik},$$

$$\hat{\beta}_{PRED} = \left(\sum_{i=1}^m \sum_{k=1}^{n_i} (w_{ik} - 1) E_{ik} \mathbf{x}_{ik} \mathbf{x}'_{ik} \right)^{-1} \sum_{i=1}^m \sum_{k=1}^{n_i} (w_{ik} - 1) E_{ik} \mathbf{x}_{ik} y_{ik}$$

Two other small area estimators considered: the Prasad-Rao (PR) and You-Rao (YR) pseudo-EBLUPs based on nested linear regression models with random small area effects (domain effects):

$$y_{ik} = \mathbf{x}'_{ik} \beta + v_i + \varepsilon_{ik}, \quad v_i \stackrel{iid}{\square} (0, \sigma_v^2) \quad \varepsilon_{ik} \stackrel{iid}{\square} (0, \sigma_\varepsilon^2 / E_{ik}), \quad (2.2)$$

where the random small area effects v_i and ε_{ik} are independent of each other, $k = 1, \dots, N_i$, $i = 1, \dots, m$.

The You-Rao (2002) pseudo-EBLUP estimator is obtained by estimating the variance components σ_v^2 and σ_ε^2 and the vector of fixed regression parameters β at the unit level and estimating the domain (area) means at the domain level:

$$\hat{Y}_{YRi} = \bar{X}'_i \hat{\beta}_{YR} + \hat{\gamma}_i (\bar{y}_{iw} - \bar{x}'_{iw} \hat{\beta}_{YR}), \quad \gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_\varepsilon^2 \delta_{iw}),$$

where $\bar{x}_{iw} = \frac{\sum_{k=1}^{n_i} w_{ik} E_{ik} x_{ik}}{\sum_{k=1}^{n_i} w_{ik} E_{ik}}$, $\bar{y}_{iw} = \frac{\sum_{k=1}^{n_i} w_{ik} E_{ik} y_{ik}}{\sum_{k=1}^{n_i} w_{ik} E_{ik}}$, $\delta_{iw} = \frac{\sum_{k=1}^{n_i} w_{ik}^2 E_{ik}}{\left(\sum_{k=1}^{n_i} w_{ik} E_{ik} \right)^2}$ and

$$\hat{\beta}_{YR} = \left(\sum_{i=1}^m \sum_{k=1}^{n_i} w_{ik} E_{ik} \mathbf{x}_{ik} (\mathbf{x}_{ik} - \hat{\gamma}_i \bar{x}_{iw})' \right)^{-1} \sum_{i=1}^m \sum_{k=1}^{n_i} w_{ik} E_{ik} (\mathbf{x}_{ik} - \hat{\gamma}_i \bar{x}_{iw}) y_{ik}.$$

Model 2.2 postulates unequal error variances, and variance components are estimated using the method of moments (Stukel and Rao (1992)). For the model (2.2) and the weighted mean AWE, we can show that the YR estimator possesses the automatic benchmarking property:

$$\text{if } \sum_{j=1}^{N_i} E_{ij} = \sum_{j=1}^{n_i} w_{ij} E_{ij} \quad \text{then} \quad \sum_{i=1}^m (E_i / E) \hat{Y}_{YRi} = \hat{Y}_{GREG},$$

where \hat{Y}_{GREG} is the GREG estimator of the overall mean \hat{Y} (Average Weekly Earnings at the model group level):

$$\bar{Y} = \sum_{i=1}^m \sum_{k=1}^{N_i} E_{ik} y_{ik} / \sum_{i=1}^m \sum_{k=1}^{N_i} E_{ik}.$$

The Prasad-Rao (1999) pseudo-EBLUP estimator is obtained by estimating the variance components at the unit level and the vector of fixed regression parameters and the domain means at the domain level:

$$\hat{Y}_{PRi} = \bar{X}'_i \hat{\beta}_{PR} + \hat{\gamma}_i (\bar{y}_{iw} - \bar{x}'_{iw} \hat{\beta}_{PR}), \quad \gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_\varepsilon^2 \delta_{iw})$$

and

$$\hat{\beta}_{PR} = \left(\sum_{i=1}^m \gamma_i \bar{x}_{iw} \bar{x}'_{iw} \right)^{-1} \sum_{i=1}^m \gamma_i \bar{x}_{iw} \bar{y}_{iw}.$$

The PR estimator does not have the benchmarking property.

2. 2 Steps for the evaluation

The estimators were evaluated using a simulated population based on SEPH sample data. The population was created by imputing AWE for non-sampled units by the nearest neighbour method, preserving the relationship between $y=AWE$ and

x=AME. In order to simulate the longitudinal nature of SEPH, 12 monthly populations were created independently of each other. One thousand stratified simple random samples (mimicking the SEPH design) were drawn and for each sample, the estimators described above were calculated. Since the simulation was done independently each month, the longitudinal relationships were not taken into account. There is some inherent positive correlation between the consecutive y-values, due to the preservation of the y-x relationships and the longitudinal correlation of the auxiliary variables. But this is usually not enough to use this population to evaluate small domain estimators that borrow strength across time.

The estimators were evaluated in terms of the Average Absolute Relative Bias, where the average is over the m small domains, and the Average Root Relative Mean Square Error:

$$ARB = \frac{1}{m} \sum_{i=1}^m \frac{\left| \frac{1}{1000} \sum_{b=1}^{1000} (\hat{Y}_{ib} - \bar{Y}_i) \right|}{\bar{Y}_i}, \quad RRMSE = \frac{1}{m} \sum_{i=1}^m \sqrt{\frac{1}{1000} \sum_{b=1}^{1000} (\hat{Y}_{ib} - \bar{Y}_i)^2}.$$

3. ANALYSIS

Figure 1. Average Absolute Relative Bias by Reference Month

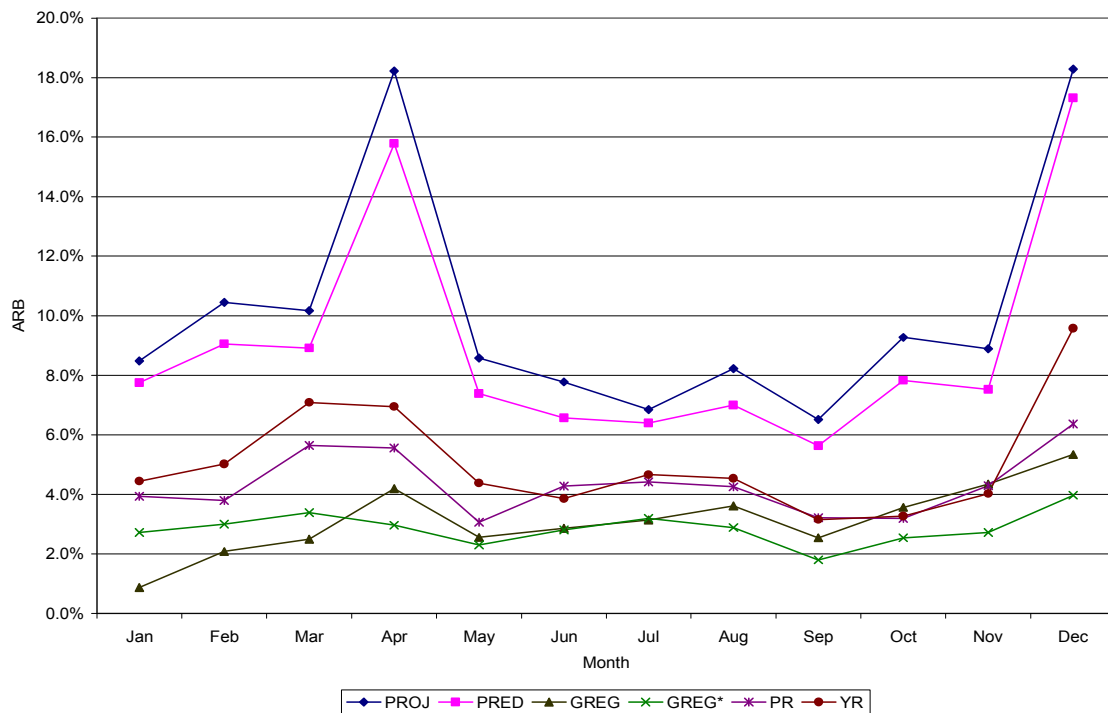
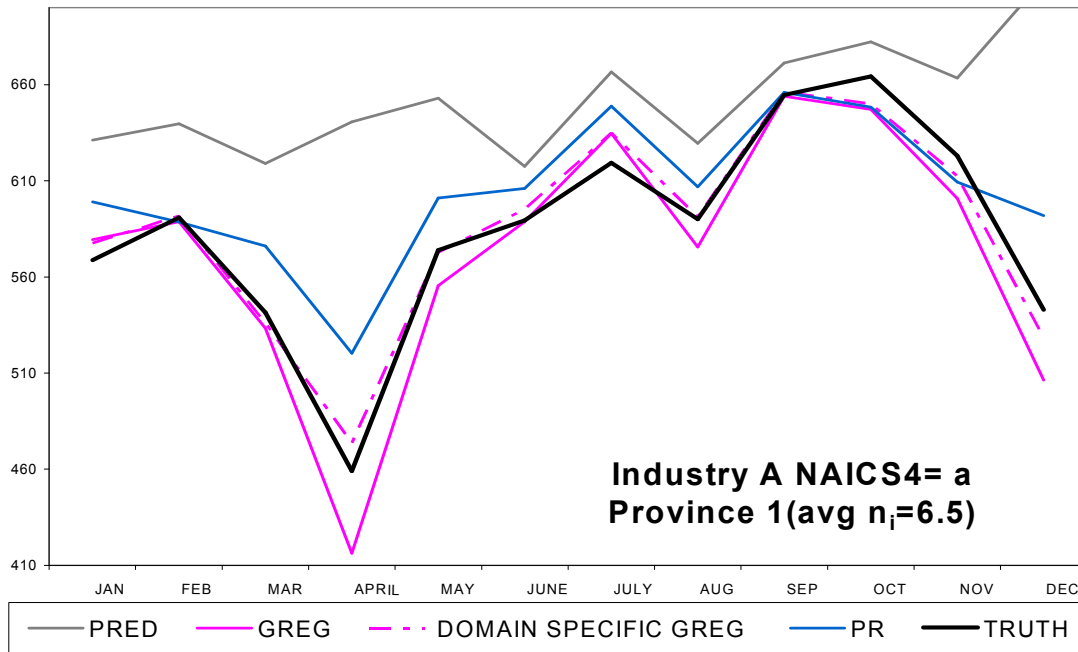


Figure 1 is a plot of the Average Relative Bias for industry A. This industry consists of 26 sub-domains at the NAICS 4 x province level, and each point in Figure 1 represents the average over the 26 domains in a period. The graph shows that the bias is lowest for GREGs, moderate for the pseudo EBLUPs and highest for PROJ and PRED. In the months of April and December, the bias of these two estimators is the worst. The month of December is always bad for most estimators: there is low correlation between AWE and AME because of bonus payments are included in AME while most of the time are not included in the reference week when AWE is collected.

Figures 2 and 3 show the Monte Carlo expectation of each one of four estimators and the true mean value in individual areas of varying sample sizes. Figure 2 examines the design bias for a particular domain (or area) from Industry A. When 1000 samples were simulated under the SEPH design, the average sample size for this domain was 6.5. Here the domain specific GREG is the least biased estimator and the PR pseudo EBLUP follows. In the months of April and December the linear predicted estimator performs the worst due to a strong area effect.

Figure 2. True domain mean and Monte Carlo expectation of PRED, GREG, GREG* and PR estimators for a domain with average sample size equal to 6.5



In Figure 3 the domain under study has an average sample size of 3.9. Here GREG and GREG* is in some periods more biased than PR. Next in Figure 4 we have a domain in the same industry but with an average sample size of 70. As in most domains of this industry, in April, AWE has a peak and PRED estimates over the other domains are influenced by the 70 units in this province and in other provinces with larger sample sizes. Even though this domain contains a large number of units, PRED exhibits a large bias in this domain as well. Domain Specific GREG and PR are nearly unbiased. We observe that when the sample size in the domain is large, the pseudo EBLUP is approximately unbiased, which is a very desirable property. We are concerned with small domains, but there are always some larger domains in the model group, and for those, it is desirable to produce estimates that are approximately equal to the design-based estimates.

Figure 3. True domain mean and Monte Carlo expectation of PRED, GREG, GREG* and PR estimators for a domain with average sample size equal to 3.9

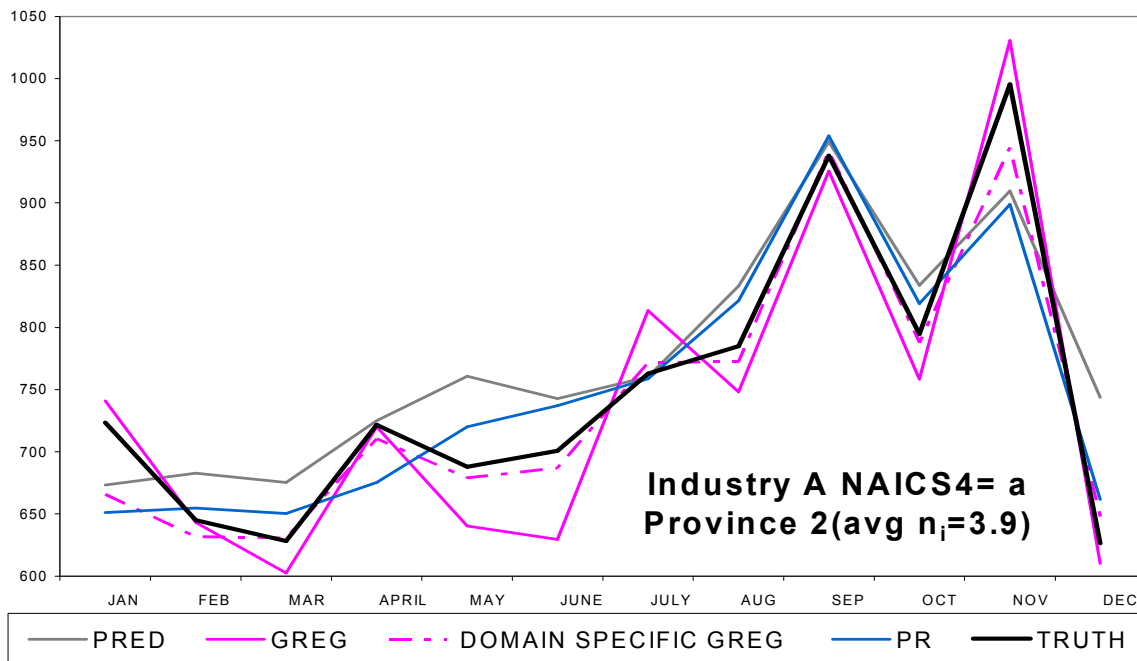


Figure 4. True domain mean and Monte Carlo expectation of PRED, GREG, GREG* and PR estimators for a domain with average sample size equal to 70

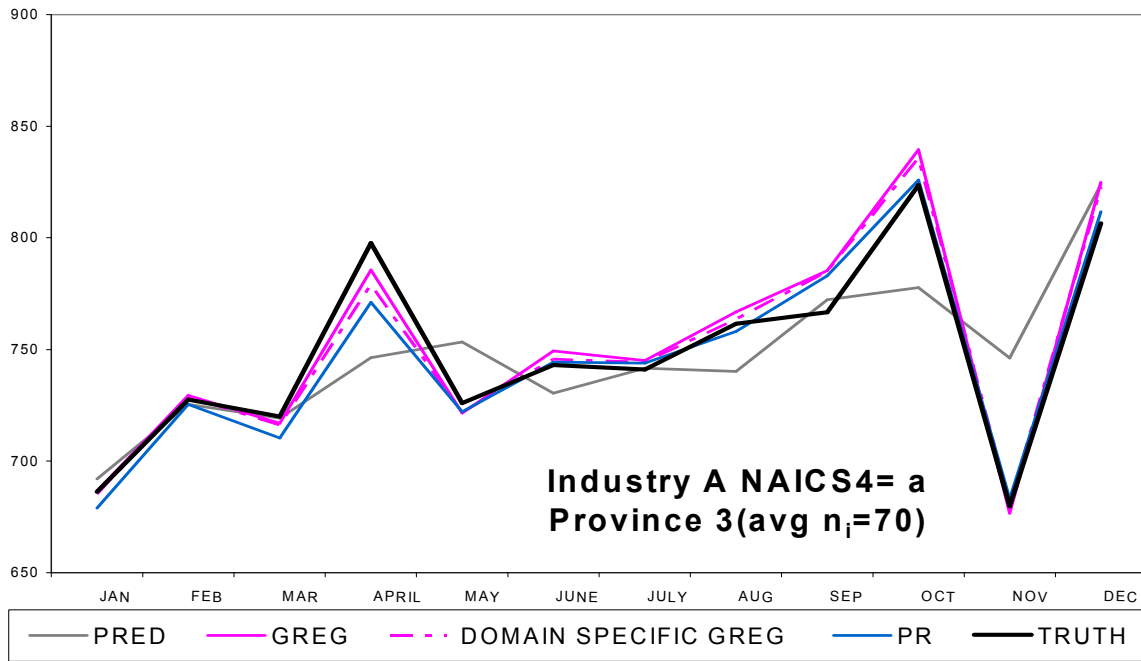


Figure 5. Average Relative Root Mean Square Error by Reference Month in Industry A

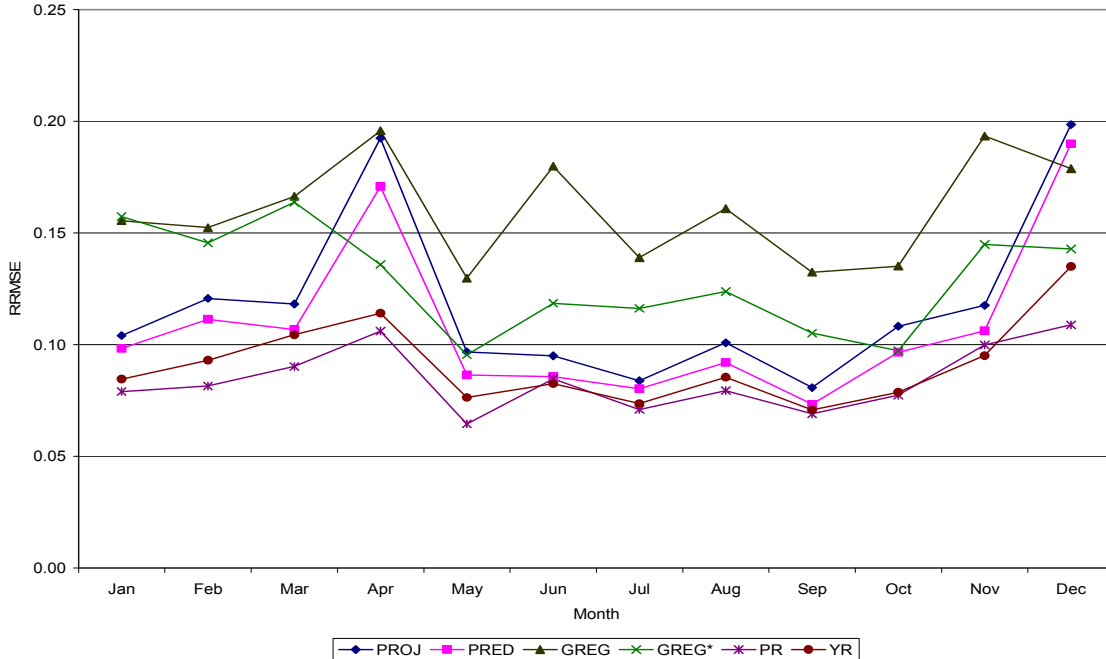


Figure 5 displays the average root relative Mean Square error for the same industry. The graph shows that the GREG and Domain specific GREG have the largest relative MSEs. This is obviously due to the large variability of these estimators for small areas.

The PROJ and PRED estimators are in the middle except in months where the domain effect is strong. YR and YR have the lowest RRMSE. And even though the differences between YR and PR are very small, PR is consistently better. This may be due to a better fit of the model at the domain level. An examination of the unit level and area level scatter plots

corresponding to an industry with relatively high correlation (Industry D, correlation from 0.6 to 0.7 across 12 months) indicates that some outliers at the unit level may be more influential than outliers at the area level.

In Table 1 we compared the reduction in RRMSE for the various industries in terms of the correlation between y and x and the strength of the domain effect. For the month of December, which is the month where the correlation is weakest, Table 1 gives:

Table 1. RRMSE – December

<i>Industry</i>	<i>Correlation</i>	<i>GREG</i>	<i>GREG*</i>	<i>PR</i>	$\sigma_v^2 / \sigma_\varepsilon^2$
A	0.27	18%	14.3%	10.9%	0.0136
B	0.30	22%	18.5%	15.5%	0.099
C	0.34	38%	23%	17%	0.0013
D	0.67	27%	18%	8.5%	0.0039

4. SUMMARY

The target parameter for this study is the Average Weekly Earnings, which is a mean weighted by employment shares. We compared the GREG and domain specific GREG estimators (which are approximately unbiased as the overall sample size increases to infinite) with several types of small area estimators. Two are based on the fixed effects GREG model, the currently used projection (synthetic) estimator and the linear predicted estimator. The PR-EBLUP and YR-EBLUP small area estimators are based on a mixed effects model. The original model used for the YR-EBLUP was extended to a model with varying error variances for estimating a weighted mean. We have proved that the benchmark property of YR extends to the case of the weighted mean.

As expected, the GREG estimators are nearly unbiased but with large variability for domains with small sample size. If the area effects ($\sigma_v^2 > 0$) are strong, PROJ and PRED show large bias. The synthetic estimator exhibit the largest bias, and what it gains in efficiency, it loses in bias. Indeed, the MSE of the projection estimator is approximately equal to the MSE of the GREG. The linear predicted estimator is efficient (in terms of RRMSE) when there are no strong individual effects with respect to the regression coefficient.

The YR and PR estimators display the lowest bias among all Small Area Estimators. They also yield the lowest RRMSE among all estimators tested. The efficiency gains or reduction in RRMSE is considerable even when the correlation between AWE and AME is low and the area effects are weak (see Table 1). The differences between YR and PR are small but for these data, PR yields consistently the lowest MSE.

In theory, the mse or estimate of the model-based MSE, should estimate the model expectation of the design-based MSE and thus, this mse should track the Monte Carlo design based MSE that we calculated in this study. If that were so, we could gauge the quality of the PR and YR estimates during production. The problem is that the estimates of the model-based mean square errors do not follow the empirical (Monte Carlo) design-based MSEs. There is some indication that perhaps parametric bootstrap MSE estimates would track them better and more research is needed in this area.

Another outstanding issue concerns built-in benchmarking. The pseudo EBLUP estimator of You and Rao (2002), has built-in benchmarking for one set of control totals, whereas the need is often to benchmark for different sets of control totals. It would be nice if we could extend the pseudo-EBLUP method to yield built-in benchmarking for more than one set of control totals.

REFERENCES

- Beaucage, Y., Godbout, S. and Morin, Y.(2005). Survey of Employment, Payroll and Hours: New Modelling Perspectives, *Internal Document, Business Survey Methods Division*, Statistics Canada.
- Prasad, NGN and Rao, JNK (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25, 67-72.
- Rao, JNK and Choudhry, H. (1995). Small Area Estimation: Overview and Empirical study. *Business Survey Methods*, Edited by Cox, Binder, Chinnappa, Christianson, Colledge, Kott, Chapter 27.
- Särndal, C.E. and Wright, R. L. (1984). Cosmetic Form of Estimators in Survey Sampling. *Scandinavian Journal of Statistics* 11: 146-156, 1984.
- Stukel, D. and Rao, JNK (1997). Estimation of Regression Models with Nested error Structure and Unequal Error variances under two and three stage Cluster Sampling. *Statistics and Probability Letters*, 35, 401-407.
- You, Y. and Rao, JNK (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, Vol.30, No3, 2002, pages 431-439.