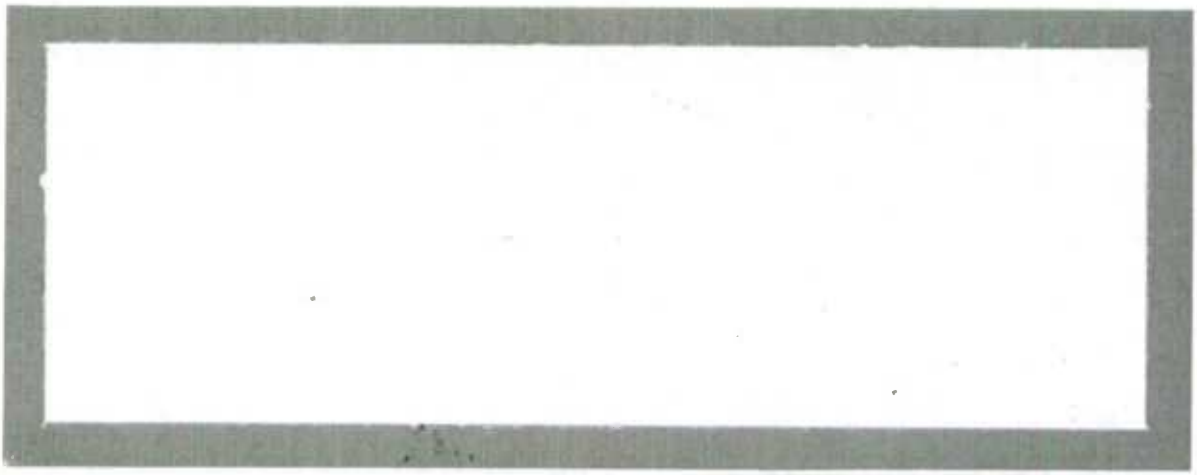
Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

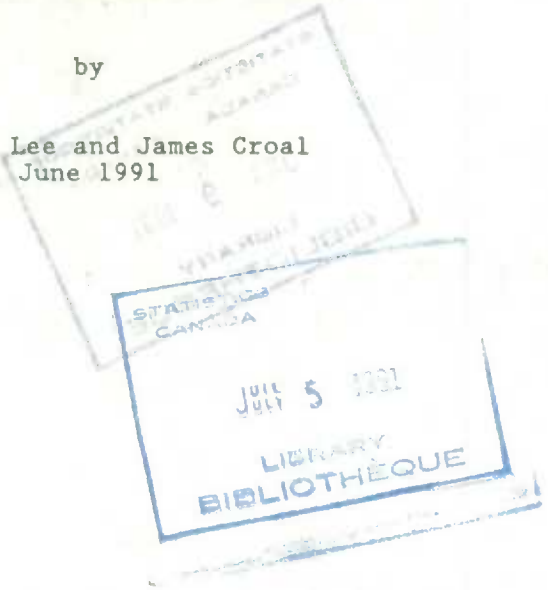
WORKING PAPER NO. BSMD-91-009E
METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-91-009E
DIRECTION DE LA MÉTHODOLOGIE

USE OF PD-7 DATA FOR ESTIMATION IN THE SURVEY OF EMPLOYMENT, PAYROLL AND HOURS

by

Hyunshik Lee and James Croal
June 1991




STATISTICS CANADA
STATISTIQUE CANADA
JUIL JULY 9 1991
LIBRARY
BIBLIOTHÈQUE

ABSTRACT

A simulation study of several ratio and regression type estimators which use data from payroll deduction as auxiliary variable for the estimation of the Survey of Employment, Payroll and Hours is reported. These estimators are compared on the basis of bias, mean square error, and efficiency relative to the current expansion estimator. The survey has a stratified design for which many of the basic strata are empty or have few establishments. Several small-area estimators, appropriate for estimation at the level of the basic strata, were examined and compared as above. Smoothing techniques were applied to the auxiliary data to improve the correlation with the employment variables. The impetus for this study is the policy of reducing the response burden for the small employers, and the possible reduction of the sample size.

The results suggest that the combined regression estimator performs better than the other estimators. For small-area estimation, the simple synthetic estimator is the best. A sample size reduction, of about 10,000 small units, is shown to be achievable, without impairing the reliability of the estimates. This study typifies the growing trend towards the multiple use of administrative data to effect savings in cost and time in the sphere of official statistics.

RÉSUMÉ

Cet article présente les résultats d'une simulation étudiant différents types d'estimateurs par le quotient et la régression en utilisant des données provenant de déduction à la source comme information auxiliaire, et ce, dans le cadre de l'Enquête sur l'emploi, la rémunération et les heures. La comparaison des différents estimateurs avec l'estimateur par expansion actuel est faite sur la base de leur biais, leur erreur quadratique moyenne ainsi que de leur efficacité relative. Le plan d'échantillonnage stratifié comporte plusieurs strates vides ou ne contenant que très peu d'établissements. Plusieurs estimateurs spécifiques à l'estimation de petits domaines, en particulier les strates de base, ont également été comparés selon les critères ci-haut mentionnés. Des techniques visant à stabiliser les données auxiliaires ont été utilisées afin d'améliorer leur corrélation avec les variables de l'emploi. Ce sont la politique ayant pour but de réduire le fardeau de réponse des petites entreprises et la possible réduction de la taille de l'échantillon qui ont mené à cette étude.

Les résultats de l'étude suggèrent que l'estimateur par la régression combiné produit de meilleurs résultats que les autres estimateurs. Pour l'estimation des petits domaines, l'estimateur synthétique simple est le meilleur. Une réduction de la taille de l'échantillon d'environ 10 000 unités est possible, sans pour autant altérer la fiabilité des estimations. Cette étude se veut un reflet de la tendance grandissante à utiliser les données administratives de multiples façons à des fins d'économie de ressources et de temps dans la sphère des statistiques officielles.

1 Introduction

The survey of Employment, Payroll and Hours (SEPH) is a monthly survey of establishments. The primary objectives of the survey are to provide:

(1) monthly estimates of the total number of paid employees, average weekly earnings, average hourly earnings, average weekly hours and other related variables at the industry division (IND)× Province (PROV) level with a reliability level measured by a coefficient of variation (CV) of 3%;

(2) the above estimates for Canada at the 3-digit SIC (Standard Industrial Classification) level;

(3) standard errors of all the estimates produced.

In addition, the survey is supposed to provide estimates at 3-digit SIC× PROV level annually. It covers all industries except agriculture, fishing and trapping, private household services, religious organizations and military services.

This study was prompted by two major corporate concerns:

(1) to reduce the respondent burden for small employers;
(2) to reduce survey costs.

The goal was to address those concerns by using administrative data in the estimation of SEPH variables. The source of the administrative data was the PAYDAC (Payroll Deduction Accounts) file maintained by Revenue Canada-Taxation (RCT). The data are captured on the PD-7 form.

At least once a month each employer remits payroll deductions (called remittance) to RCT. These are Canada /Quebec Pension Plan premiums, Unemployment Insurance premiums and income tax. The agency captures these data and passes them to Statistics Canada every month for various uses such as business list maintenance.

A simulation study was conducted to identify the best sampling strategy and the best estimator among the ratio and regression types. Two subsidiary concerns guided the study:

(1) maintain the current reliability of the SEPH estimates for level;

(2) minimize any changes to the current SEPH design and processing systems.

This paper reports the results of the study and is structured as follows: section 2 gives a brief description of the SEPH design; section 3 discusses the key issues, objectives, and constraints of the study; section 4 describes the study procedure; section 5 details the results of the study; finally, section 6 states some conclusions and recommendations.

2 The SEPH Design

The SEPH universe consists of approximately 700,000 establishments which report paid employment and related data. These establishments are called employment reporting units (ERUs). A sample of about 70,000 ERUs is drawn each month. The current survey design uses a stratified simple random sample with rotation. SEPH uses the expansion estimator. There are four levels of stratification:

(i) geography: 10 provinces and 2 territories;
(ii) industry: SIC;
(iii) size-group: determined by the number of paid employees;
(iv) Take-all /Take-some grouping.

The size measure of a sampling unit is defined as the total number of paid employees in the ERU. There are 4 size groups:

- size-group 1: 0 - 19 employees;
- size-group 2: 20 - 49 employees;
- size-group 3: 50 - 199 employees;
- size-group 4: 200 or more employees.

The Take-all /Take-some classification is determined by the size and the organizational structure of companies. Currently all units in size group 4, and multi-ERU companies (Multis) with a total paid employment exceeding 199, are included in Take-all class; they are sampled with certainty. The other establishments belong to the Take-some class. At the time this study was done, all multis, regardless of size, were Take-all.

The lowest level for which estimates are required in SEPH is the cross-classification PROV × SIC3 × SIZE, called a cell. Cells are the building blocks of SEPH for producing estimates at higher levels of aggregation. There are 13,488 possible cells of which about one half are empty. Many of the nonempty cells have a small number of units.

The sample allocation for the Take-some portion is X-proportional, i.e., proportional to the total number of employees, at the PROV × IND × SIZE, and then N-proportional (i.e., proportional to the number of ERUs) at the cell level. Owing to the X-proportional allocation and the skewed nature of the population, the larger the size-group, the greater is its sampling rate.

The Take-some portion of the sample is rotated by replacing about 1/12 of the sampled ERUs each month. The portion rotated out stays out of the sample for at least the next 12 months.

For details of the SEPH methodology, readers are referred to Schiopu-Kratina and Srinath (1986).

3 Objectives, Constraints and Key Issues

In order to use administrative data as auxiliary information to improve a survey estimate, two basic requirements should be fulfilled:

(i) good linkage between administrative records and the survey units;

(ii) correlation between the auxiliary and the survey variables high enough to ensure some gains in efficiency.

The first requirement was satisfied. The Business Register Master File, on which the SEPH frame was based, provided the necessary linkage, not at the establishment level, but at the company level. Multi-ERU companies, the majority of the Take-all class, posed the problem of the possible lack of direct links from the administrative records to the individual ERUs. This problem does not exist for single-ERU companies, which constitute the bulk of the Take-some class. Therefore, estimation using PD-7 data was investigated for Take-some units only.

Preliminary studies showed that the second condition could be satisfied and substantial gains in efficiency could be realized.

An earlier study indicated that a significant reduction of the sample size would not be possible if the new estimator were applied at SIC3× PROV level. Hence, it was decided to collapse SIC3s into SIC2s in order to raise the application level. Assuming that the regression estimator was to be applied at the SIC2× PROV level, a rough estimate of 10,000 was obtained for the reduction of the sample size from the Take-some portion of size-groups 1 and 2.

If the new estimator is to be calculated at SIC2× PROV level, it is still necessary to calculate estimates at SIC3× PROV level by means other than the new estimator. Owing to the problem of small sample sizes at this level, a small area estimation technique might have to be used rather than an ordinary method. Then, an appropriate small area estimation method should be identified.

There are problems with the remittance data. The most serious is the irregular behavior, over time, of the remittance, even from the same employer. Cotton (1987) showed that the average remittance, over 2 or 3 consecutive months, has higher correlations with the survey variables than the raw remittance. However, averages may be influenced by the presence of outliers. Scatter plots, of survey variables versus remittance, showed that a few outliers significantly decreased the correlation between the survey and remittance variables, even when the average remittance was used. The effective use of the remittance as an auxiliary variable required some form of outlier treatment.

The focus of this study was the comparison of ratio and regression type estimators with the expansion estimator. Cotton (1987) showed that the separate regression estimator had a smaller estimated variance than the separate ratio estimator. It was possible that the combined ratio or combined regression estimator could perform better. The bias could be an important component of the mean square error, its importance being measured by the ratio of bias to standard deviation for the estimator. If so, it might be desirable to use an unbiased estimator such as Mickey's estimator (1959). These five estimators were included in the study.

The main questions to be addressed were the following:

(1) Is a reduction of the sample-size of 10,000 possible?

(2) What is the appropriate strategy to achieve this target?

The following related issues arose:

(1) Is it possible to achieve the objectives of the study using the ratio estimator? It is simpler and operationally easier to use than the regression estimator. However, it can be less efficient.

(2) Is it necessary to collapse SIC3s even up to the IND level?

(3) What should the minimum sample size be at the SIC2× PROV level?

(4) By how much can the sample size be reduced at the SIC2× PROV level?

4 Methodology of the Study

In order to address the issues raised in the previous section, we conducted a simulation study using SEPH sample data to generate population data sets. A number of simple random samples were drawn via a Monte Carlo technique, from each population following the SEPH design at the SIC2× PROV level and the estimators selected for study were calculated for each sample. Their bias, variance, mean square error and relative efficiency were then computed.

4.1 Description of Population Data Sets for Simulation

Four population data sets for simulation were created for two consecutive months. Three populations denoted SSP1, SSP2 and MSP, were generated directly from SEPH data. The fourth population, LSP, was generated artificially using parameters estimated from SEPH data. These populations differed in the average size (i.e., the number of ERUs) of their strata. Table 1 lists the the characteristics of these populations. At least two months' data were needed to estimate the month-to-month change.

4.2 Estimators Considered

Six estimators were studied, including the expansion estimator, which served as a basis of comparison for the other five estimators. These were the separate and combined ratio, Mickey's unbiased ratio, the separate and combined regression estimators (for definitions, see Cochran, 1977). The combining was done over the size-groups 1 and 2.

For small area estimation at the SIC3 × PROV level, 9 estimators were studied: the expansion, two synthetic, two composite estimators, and four empirical best linear unbiased predictors (EBLUP). These estimators are defined in the Appendix.

5 The Results of the Study

5.1 The Results for Level Estimates with Untreated Remittance

5.1.1 The Large Size Population

Table 2 shows the average relative efficiencies with respect to the expansion estimator of the ratio, regression and Mickey's estimators. Note that the table shows the results obtained from using the raw remittance. The relative efficiency (REFF) is defined in terms of the mean square error (MSE) as follows:

REFF =100 %× (MSE of expansion estimator/MSEof alternative estimator)

Estimates were calculated for three major SEPH variables: total employment (EMP), total gross payroll (GRP), and total hours (HRS). Stratification by size was not done for this population because of the difficulty in assigning realistic size measures. An attempt was made to assign size measures using each unit's number of employees, but characteristics of such defined size groups for the artificially generated population did not resemble those of the SEPH data which provided the population parameters. These parameters were calculated using a pooled sample of both size groups to get a large enough sample size. Therefore, the combined ratio and regression estimators could not be tried.

In the table, the number of cases out of 72 that show a gain over the expansion estimator (i.e., relative efficiency greater than 100) is presented.

Clearly the regression estimator is the best for all three variables with more than 80% of cases showing gains and with relative efficiencies of more than 200. The ratio estimator follows next and Mickey's estimator is the least efficient. It seems that for the large samples the bias of the ratio estimator is not serious and the premium paid to achieve unbiasedness by Mickey's estimator is somewhat larger in terms of loss of variance efficiency. It was not suprising to see that all three estimators performed better for GRP than for the other two variables because the correlation between GRP

and remittance (RMT) is the highest for the 3 variables. The cases which suffer loss of efficiency also show low correlations. Note also that the table shows the results with untreated remittance data.

5.1.2 The Medium Size Population

The medium size population has size groups, hence it was possible to use all five estimators. Table 3 shows the results. Again the regression estimator is the best, followed by the ratio, and Mickey's is the last. Even though the number of cases of gain is almost the same for both combined and separate estimators, the combined estimators are clearly more efficient than the separate estimators. Combining increases the efficiency, in a more noticeable way, for the ratio estimator than for the regression. The correlations are all fairly high (>0.5) and thus, more than 90% of cases showed gain. It is interesting to observe that, for this population, the gap between the performances of the ratio and Mickey's estimators is smaller than is the case for the large size population.

5.1.3 The Small Size Population 1

For this population, in terms of average REFF, the alternative estimators, in general, performed better than the expansion estimator. This was not true for efficiency. Some estimators even suffered a loss of efficiency. The combined regression estimator is the best here again. The next best performer is not clear among the separate regression and combined ratio. In terms of average REFF, they are quite close; in terms of the number of cases of gain, the former is clearly better. Mickey's estimator is again the worst even though its performance differs slightly from that of the ratio estimator. Even with the combined regression estimator, there are many cases (about 20-40%) which are less efficient than the expansion estimator. It seems that these alternative estimators do not perform well with a small sample size, unless the correlation is very high.

5.1.4 The Small Size Population 2

The characteristics of this population are quite different from the other populations, not only in terms of size, but also with respect to other features. Of course, the size of the population is the most important factor that distinguishes it from other populations. For the other populations, strata are defined by SIC2× PROV× SIZE (except for the large size population, which has no size group), but the strata in this population are defined by IND× PROV× SIZE. The data source of this population is the SEPH sample data for all industry divisions from the Yukon and the Northwest Territories. The average stratum size of 6 for size group 2 is very small (see Table 1 and compare with the average stratum size of 18 for size-group 2 of the small size population 1). Owing to this fact, the sample size in size-group 2 is often 1 or the same

as the stratum population size. Therefore, the ordinary separate ratio and regression estimators are often not defined. For the combined estimators, size-group 2 does not contribute to the estimate of the slope of the regression equation when the sampling fraction is equal to 1. Because of this problem, the ratio, regression and Mickey's estimators were calculated by using a pooled sample of size-groups 1 and 2, disregarding the sampling weights.

The results for this population were quite surprising in that the ratio and Mickey's estimators did very well under seemingly severe conditions and, in fact, were much better, in terms of average REFF, than the regression estimator. This is a marked deviation from the results for the other three populations. The most probable reason for this phenomenon is that collapsing of SIC2s at IND level makes the population data follow more closely the super-population model, $y_i = x_i + \epsilon_i$, $E(\epsilon_i) = 0$ and $V(\epsilon_i) = x_i \sigma^2$, for which the ratio estimator is the best.

5.2 Outlier Treatment

5.2.1 Outlier Treatment of Remittance Data

As mentioned in the introduction, the irregular behaviour of the remittance data over time poses problems when using the data as auxiliary information in a ratio or regression estimator.

The correlations of average remittance over 2 or 3 consecutive months, with the SEPH variables, are higher than those for the corresponding monthly remittance. However, large month-to-month fluctuations in remittance do occur, not infrequently. In this case averaging will dampen the effect of an unusually large remittance somewhat but not enough. A more serious problem with this method is that such a large value carries its influential effect over two or three months through the averaging process. This problem led to the study of simple robust techniques such as median and trimmed mean of remittances over 3, 5, 7, and 9 months. Other methods explored include the detect-and-replace method for unusual values. This was done by comparing the current value of the remittance with a fixed multiple of the median or trimmed mean for the remittances of previous months. When the current value is judged unusual, it is replaced by the median or trimmed mean. All these methods improve correlations even more than averaging and are quite similar to each other in terms average correlation. Among these, the median of the remittances for 3 consecutive months seems to be the best choice because it is less computationally intensive than the other alternatives.

Table 6 presents the average correlation coefficients at SIC2× PROV of treated and untreated remittances with EMP, GRP, and HRS for the two months, October and November 1987. The two- and three- months' averages are almost identical in terms of average correlation. The median of three months is consistently better than the other two. The correlations of the untreated remittance for November are much

smaller than those for October which indicates the problem of outliers in November data is more serious; all three treatment methods improve the correlations to almost the same level as the correlations of the treated data for October. This implies that the treatment methods not only improve, but also stabilize, the correlations over time.

In Table 7, average REFFs of the combined regression estimator with the untreated and the treated (median of three months) remittances are shown. It is clearly shown that the method is very effective in improving the performance of the estimator especially for the small size population 1, where correlations are generally lower than those for the medium size population.

5.2.2 Detection and Treatment for Relational Outliers

Scatterplots indicated that some outlier treatment was needed to use either the ratio or regression estimator more efficiently. Croal (1988) studied the ratio-range method of outlier detection. The *y*-value of a detected outlier was replaced by a predicted value. The predicted value was calculated using the mean ratio of non-outliers. This method was effective in improving the performance of the ratio estimator. It is, however, intuitively clear that this method is not good for the regression estimator with a non-zero intercept.

The regression quantile method was proposed for robust estimation of the linear regression model (Koenker and Bassett, 1978). The method gives an estimate of the regression equation using the least absolute deviation criterion rather than the usual least squares criterion. Portnoy (1987) studied the method for outlier detection for a linear regression model and recommended its use because of its very high break-down point as in the case of the median for the one-dimensional case.

In this study, we applied these two methods for outlier detection and treatment.

In the case of the regression quantile, we used the FORTRAN program written by Koenker (1987) which computes the regression quantile for a given theta between 0 and 1 ($0 \leq \theta \leq 1$). We obtained the median regression equation from the program with $\theta = 0.5$. We then identified outliers from an examination of the residuals from the regression. A fixed percentage of the identified outliers were simply deleted from the calculation for the estimators.

The results of the study are presented in Table 8. The regression quantile method in its simplest form seems to be quite effective for improving the performance of the regression estimator.

5.3 Estimating Month-to-Month Change by Regression Estimators

The change estimate is also important for SEPH. The estimates of change by regression estimators are compared with the expansion estimator in Table 9. The average REFFs,

although still exceeding 100, are much smaller than those for the level estimates. This means that the regression estimators are not as efficient for estimating change as for estimating level. Thus, a regression estimator which maintains the current reliability for level with a reduced sample size will not guarantee the same reliability for change as the current expansion estimator with full sample size.

5.4 Comparison of Aggregated Estimates

The ratio and regression estimators are biased. The bias could pose a serious problem when it is cumulative at higher levels of aggregation. The biases of the ratio and regression estimators were positive in about 70 % of the cases. Therefore, aggregated estimates are positively biased and the average REFFs are decreased substantially. Tables 10-12 show performances of the aggregated estimates of the expansion, ratio, regression and Mickey's estimators.

5.5 Small Area Estimation for SIC3s

Eight small area estimators were studied. They are based on 4 basic linear models which are described in the Appendix. A short description of these estimators is given in the table below.

No	Name	Description
1	Expansion	Blow-up estimate or survey estimate
2	Synthetic 1	Based on Model 1 with constant B
3	Synthetic 2	Based on Model 2 with constant B
4	Composite 1	Linear combination of Expansion and Synthetic 1
5	Composite 2	Linear combination of Expansion and Synthetic 2
6-9	EBLUP 1-4	Based on Models 1-4

EBLUP stands for the Empirical Best Linear Unbiased Predictor and the definitions of EBLUP 1-4 can be found in Choudhry and Rao (1988).

In general, the sum of these small area estimates for SIC3s at a SIC2 is not the same as the estimate obtained by an alternative estimator, say the combined regression estimator, at the SIC2. In order to avoid this non-additivity, a benchmarking procedure was applied as follows.

Let Y be the combined regression estimate, at the SIC2× PROV level, and $Y_{s1}, \ldots, Y_{sk}$ be the small area estimates for the SIC3s in the SIC2. Then the benchmarked small area estimate, Y'_{si} is defined as:

$$Y'_{si} = \frac{Y_{si}}{Y_{s+}} Y,$$

$$Y_{s+} = \sum_{j=1}^{k} Y_{sj}.$$

For Synthetic 1 and 2, the formula reduces to:

$$Y'_{si} = \frac{X_{si}}{X_{s+}} Y,$$

where $X_{s1}, \ldots, X_{sk}$ are the total remittances of the 3-digit SICs and $X_{s+} = \sum_{j=1}^{k} X_{sj}$.

Tables 13-14 show the results of the 9 estimators, unbenchmarked and benchmarked. Undoubtedly an appropriate model selection is very important as can be seen from the tables. Benchmarking has a tremendous effect on the performances of these estimators, especially the bad ones. Regardless of benchmarking, Synthetic 1 seems to be the best estimator. With benchmarking, Synthetic 2 is almost as good as Synthetic 1. EBLUPs performed well in other business surveys (Choudhry and Rao, 1988). The results of this study indicate otherwise.

6 Concluding Remarks

Concluding remarks are given in point form as follows:

(1) In general, the ratio, regression and Mickey's estimators are superior to the expansion estimator. But it is risky to use these estimators when sample sizes are small, unless the correlation is very high.

(2) The regression estimator is generally better than the ratio which is better than the Mickey's. The difference, however, gets smaller as stratum population and sample sizes get smaller.

(3) Collapsing at the industry division level favours the ratio estimator.

(4) The combined estimators are generally better than the separate ones.

(5) Biases of the ratio and regression estimators are generally positive and thus cumulative at higher levels of aggregation. The bias of the ratio estimator is larger than that of the regression.

(6) The average REFFs of the ratio and regression estimators at the higher level aggregation tend to be reduced by accumulated bias.

(7) Using the median of three months' remittances improves and stabilizes correlation over time.

(8) The regression quantile method for relational outlier treatment was found to be quite effective for improving the efficiency of the regression estimator.

(9) The performances of the ratio and regression estimators for estimating change are not as good as for estimating level even though they are generally still better than the expansion estimator.

(10) Selecting a proper model is very important for small domain estimation. The model with error variance proportional to x seems to be appropriate.

(11) The simple synthetic estimator is the best among the small area estimators studied.

On the basis of these conclusions, we recommend the combined regression estimator at the SIC2× PROV level. However, it should be applied selectively to those SIC2× PROV substrata which have a high correlation and large sample size. The magnitude of the reduction of the sample size should depend, not only on the REFF, but also on the correlation and the stratum population and sample sizes. The estimate for level generated by the alternative estimator and the reduced sample should achieve the same reliability as the expansion estimator with the full sample. However, in the light of Point (9) above, the estimate for change by the alternative estimators with the reduced sample will be slightly less reliable than the current estimate.

Analyzing the simulation results, we developed a sample-size reduction scheme given in the appendix. Using the scheme, we obtained an estimate of sample-size reduction of about 10,000. Collapsing SIC3s to SIC2 was sufficient to achieve this reduction. However, it is possible that accumulated bias in regression estimation at higher levels of aggregation may adversely affect the reliability. It is not clear how much this cumulative bias would affect the design objective of 3% CV for IND× PROV estimates of employment.

For small area estimation at the SIC3× PROV level, Synthetic 1 is recommended.

7 Acknowledgement

The authors are grateful to two reviewers, David Dolson and Jack Gambino, for their very helpful comments.

8 Appendix

8.1 Models for the Small Area Estimators

The models used to derive the small area estimators studied in this paper are given in the following.

Model 1:

$$y_{ij} = \beta_i x_{ij} + \epsilon_{ij},\ \beta_i = \beta + \alpha_i,\ V(\epsilon_{ij}) = x_{ij}\sigma^2.$$

Model 2:

$$y_{ij} = \beta_i x_{ij} + \epsilon_{ij},\ \beta_i = \beta + \alpha_i,\ V(\epsilon_{ij}) = x_{ij}^2\sigma^2.$$

Model 3:

$$y_{ij} = \beta x_{ij} + \alpha_i + \epsilon_{ij},\ V(\epsilon_{ij}) = x_{ij}\sigma^2.$$

Model 4:

$$y_{ij} = \beta x_{ij} + \alpha_i + \epsilon_{ij},\ V(\epsilon_{ij}) = x_{ij}^2\sigma^2.$$

For all these models, it is assumed that $E(\alpha_i) = 0$, $E(\epsilon_{ij}) = 0$. The subscript i stands for the i-th small area and the subscript j for the j-th unit in the i-th small area.

8.2 Sample Reduction Scheme

The formula used in the scheme is a large sample one. When the original sample size is less than 100, a reduced correlation is used to compensate for the inaccuracy of the formula. The amount of the compensatory reduction of the correlation is determined by a regression analysis of the results of the simulation study. The formula is shown below:

$$n_2 = \frac{n_1 N(1 - r'^2)}{(N - n_1 r'^2)}$$

where

n_1: current sample size at SIC2× PROV,
n_2: reduced sample size at SIC2× PROV,
N: population size at SIC2× PROV,
r' : reduced correlation coefficient between EMP and RMT (see the table below).

However, the reduced sample size was not allowed to be less than 10. The sample reduction scheme is given as follows:

Conditions to Use Regression Estimator		Determination of	Lower
SEPH Sample Size[1] (n_1)	Corr. with RMT (r)	Reduced Correlation r'	Bound of n_2
10 - 15	> 0.85	$r' = 1.6r - 0.8$	10
16 - 20	> 0.8	$r' = 2.5r - 1.5$	10
21 - 30	> 0.7	$r' = r - 0.15$	$0.5\,n_1$
31 - 50	> 0.6	$r' = r - 0.1$	$0.4\,n_1$
51 - 100	> 0.5	$r' = r - 0.1$	$0.3\,n_1$
>100	> 0.4	$r' = r$	$0.25\,n_1$

Note 1: the sample size is of both size groups 1 and 2 at SIC2× PROV.

If the conditions in the above table are not satisfied in a SIC2 × PROV, the current expansion estimator is used with the full SEPH sample.

9 References

Choudhry, H., and Rao, J.N.K., (1988). Evaluation of Small Area Estimators: An Empirical Study. Presented at the International Symposium on Small Area Statistics, New Orleans.

Cochran, W.G., (1977). Sampling Techniques; 3rd Edition. New York: John Wiley & Sons.

Cotton, C., (1987). PD-7 Remittances and SEPH. Technical Report, Business Survey Methods Division, Statistics Canada.

Croal, J., (1988). Stability of the Ratio, Gross Pay to Remittance, and the Regression Coefficient (Slope) of Gross Pay on Remittance. Technical Report, Business Survey Methods Division, Statistics Canada.

Koenker, R.W., and Bassett, G.W., (1978). Regression Quantiles. Econometrica, Vol. 46, pp. 33-50.

Koenker, R., and D'Orey, V., (1987). Computing Regression Quantiles. Applied Statistics, Vol. 36, pp. 383-393.

Mickey, M.R. (1959). Some Finite Population Unbiased Ratio and Regression Estimators. Journal of American Statistical Association, Vol. 54, pp. 594-612.

Portnoy, S., (1987). Using Regression Fractiles to Identify Outliers. Statistical Data Analysis Based on the L1-Norm and Related Methods; Y. Dodge, (Ed.). Amsterdam: North-Holland.

Schiopu-Kratina, I., and Srinath, K.P., (1986). The Methodology of the Survey of Employment, Payroll and Hours. Technical Report, Business Survey Methods Division, Statistics Canada.

Table 1. Description of the Population Data Sets Used in the Simulation

		SSP1[a]	SSP2	MSP	LSP
Data Source	Industry[b]	61	All	41	87
	Province[c]	10-59	60,61	10-59	24-59
No. of Cases[d]		60	50	40	72
Average Size of Population	Size Group 1	38	40	182	-
	Size Group 2	18	6	48	-
	Size Grp 1 & 2	56	46	230	5004
Average Correlation Coefficient	EMP	0.72	0.69	0.85	0.63
	GRP	0.83	0.77	0.89	0.78
	HRS	0.74	0.67	0.84	0.66
Method of Generation		Sample Data	Sample Data	Sample Data	Artificially
Level of Sampling		2-Digit SIC	Industry	2-Digit SIC	2-Digit SIC
No. of Random Samples		200	200	200	100

Note: a. SSP1 = Small Size Population 1
SSP2 = Small Size Population 2
MSP = Medium Size Population
LSP = Large Size Population

b. Industry Division 61 = Wholesale Trade
Industry Division 41 = Building Construction
Industry Division 87 = Commercial Service

c. Province Code Definition

10 = Newfoundland	46 = Manitoba
11 = Prince Edward Island	47 = Saskatchwan
12 = Nova Scotia	48 = Alberta
13 = New Brunswick	59 = British Columbia
24 = Quebec	60 = Yukon
35 = Ontario	61 = North Western Territory

d. This is defined as 2 times number of strata where independent sampling takes place.

Table 2. The Average Relative Efficiency of the Alternative Estimators w.r.t. the Expansion Estimator for the Large Size Population with Untreated Data (sampling rate same as SEPH)

Variable		Ratio	Regression	Mickey's
EMP	No. of Cases of Gain (%)	40 (56%)	59 (82%)	29 (40%)
	Ave REFF	147	202	116
GRP	No. of Cases of Gain (%)	53 (74%)	66 (92%)	50 (69%)
	Ave REFF	394	488	319
HRS	No. of Cases of Gain (%)	42 (58%)	61 (85%)	39 (54%)
	Ave REFF	191	241	148

Table 3. The Average Relative Efficiency of the Alternative Estimators w.r.t. the Expansion Estimator for the Medium Size Population with Untreated Data (Sampling rate = 0.2)

Variable		Ratio		Regression		Mickey's
		Separate	Combined	Separate	Combined	
EMP	No. of Cases of Gain (%)	30 (75%)	31 (78%)	39 (98%)	39 (98%)	27 (68%)
	Ave REFF	141	176	221	240	132
GRP	No. of Cases of Gain (%)	38 (95%)	38 (95%)	40 (100%)	40 (100%)	34 (85%)
	Ave REFF	300	350	377	391	274
HRS	No. of Cases of Gain (%)	33 (83%)	33 (83%)	37 (93%)	37 (93%)	31 (78%)
	Ave REFF	167	196	217	233	157

Table 4. The Average Relative Efficiency of the Alternative Estimators w.r.t. the Expansion Estimator for the Small Size Population 1 with Untreated Data (Sampling rate = 0.3)

Variable		Ratio		Regression		Mickey's
		Separate	Combined	Separate	Combined	
EMP	No. of Cases of Gain (%)	18 (30%)	26 (43%)	31 (52%)	36 (60%)	24 (40%)
	Ave REFF	74	113	115	133	66
GRP	No. of Cases of Gain (%)	35 (58%)	38 (63%)	42 (70%)	49 (82%)	33 (55%)
	Ave REFF	201	287	278	313	170
HRS	No. of Cases of Gain (%)	23 (38%)	29 (48%)	38 (63%)	40 (67%)	22 (37%)
	Ave REFF	107	129	141	163	92

Table 5. The Average Relative Efficiency of the Alternative Estimators w.r.t. the Expansion Estimator for the Small Size Population 2 with Untreated Data (Sampling rate = 0.2)

Variable		Ratio Pooled[a]	Regression Pooled	Mickey's Pooled
EMP	No. of Cases of Gain (%)	21 (42%)	19 (38%)	17 (34%)
	Ave REFF	246	159	232
CRP	No. of Cases of Gain (%)	38 (76%)	36 (72%)	36 (72%)
	Ave REFF	585	343	582
HRS	No. of Cases of Gain (%)	27 (54%)	29 (58%)	22 (44%)
	Ave REFF	332	209	262

Note: a. Samples from size groups 1 and 2 were pooled to calculate the estimators because ordinary ratio and regression estimators were often not defined due to small population size of size group 2 (see Table 1).

Table 6. Average Correlation Coefficients at 2-Digit X Province of Treated and Untreated Remittances for October and November 1987

Treatment Methods	EMP		GRP		HRS	
	Oct	Nov	Oct	Nov	Oct	Nov
Untreated	0.75	0.68	0.82	0.75	0.74	0.68
2 Month Average	0.77	0.75	0.84	0.83	0.75	0.75
3 Month Average	0.77	0.76	0.84	0.83	0.76	0.75
Median of 3 Months	0.79	0.78	0.86	0.85	0.78	0.77

Table 7. Relative Efficiency of the Combined Regression Estimator with Untreated and Treated (Median of 3 Months) Remittances

	EMP		GRP		HRS	
	U[a]	T[b]	U	T	U	T
	Small Size Population 1					
Cases of Gain (%)	60%	77%	82%	92%	67%	82%
Average REFF	115	177	313	527	163	230
	Medium Size Population					
Cases of Gain (%)	98%	98%	100%	98%	93%	93%
Average REFF	240	260	391	474	233	253

Note: a. U = Untreated
b. T = Treated

Table 8. Average Relative Efficiency of the Alternative Estimators with the Ratio-Range and Regression Quantile Outlier Treatment Methods and with the Untreated Remittance

	Ratio			Regression			Mickey's		
	U[a]	RR[b]	RQ[c]	U	RR	RQ	U	RR	RQ
	Medium Size Population								
Cases of Gain (%)	95%	95%	90%	100%	95%	95%	85%	95%	83%
Average REFF	300	432	431	377	448	503	274	428	357
	Large Size Population								
Cases of Gain (%)	74%	83%	71%	92%	85%	89%	69%	78%	63%
Average REFF	394	384	420	488	495	572	319	526	365

Note: a. U = Relational outliers untreated
b. RR = Relational outliers treated by the ratio-range method
c. RQ = Relational outliers treated by the regression quantile method.

Table 9. The Average Relative Efficiency of Estimate of Change by the Regression Estimator with Treated Remittance

	EMP	GRP	HRS
		Small Size Population 1	
Cases of Gain (%)	60%	63%	50%
Average REFF	105	181	116
		Medium Size Population	
Cases of Gain (%)	90%	95%	85%
Average REFF	172	218	172
		Large Size Population	
Cases of Gain (%)	47%	64%	50%
Avearge REFF	117	239	126

Table 10. Comparison of Aggregated[a] Estimates
for the Large Size Population with Untreated Remittance

Summary Statistics	Variables	Expansion	Ratio	Regression	Mickey's
Relative Bias (%)	EMP	0.0	3.2	2.0	-0.3
	GRP	0.0	2.7	1.9	-0.5
	HRS	0.2	3.2	2.0	-0.2
Relative Root Mean Square Error	EMP	3.6	5.3	3.5	5.7
	GRP	3.9	4.9	3.1	5.8
	HRS	4.2	5.6	3.7	6.1
Average Relative Efficiency	EMP	100	47	109	41
	GRP	100	65	155	45
	HRS	100	57	132	48

Note: a. Aggregated over six 2-digit SICs and six provinces.

Table 11. Comparison of Aggregated[a] Estimates
for the Medium Size Population with treated Remittance

Summary Statistics	Variables	Expansion	Comb. Ratio	Comb. Regression
Relative Bias (%)	EMP	0.6	1.4	1.0
	GRP	0.7	0.8	0.8
	HRS	0.7	1.4	1.1
Relative Root Mean Square Error	EMP	3.6	2.9	2.6
	GRP	4.5	2.4	2.2
	HRS	3.9	3.2	2.7
Average Relative Efficiency	EMP	100	125	200
	GRP	100	315	377
	HRS	100	156	215

Note: a. Aggregated over two 2-digit SICs and ten provinces.

Table 12. Comparison of Aggregated[a] Estimates
for the Small Size Population with treated Remittance

Summary Statistics	Variables	Expansion	Comb. Ratio	Comb. Regression
Relative Bias (%)	EMP	0.0	1.8	1.1
	GRP	0.0	1.2	0.5
	HRS	-0.1	1.7	1.1
Relative Root Mean Square Error	EMP	3.5	3.5	2.9
	GRP	3.6	2.5	2.2
	HRS	3.8	3.6	3.1
Average Relative Efficiency	EMP	100	101	150
	GRP	100	210	295
	HRS	100	108	147

Note: a. Aggregated over three 2-digit SICs and ten provinces.

Table 13. The Performances of the Small Area Estimators for EMP
(averaged over 179 small areas)

Estimator	Unbenchmarked			Benchmarked		
	ARB[a]	ARE[b]	REFF	ARB	ARE	REFF
Expansion	5	68	100	6	70	100
Synthetic 1	26	28	248	26	28	260
Synthetic 2	***[c]	***	0	26	28	263
Composite 1	19	34	188	18	38	168
Composite 2	***	***	2	133	176	26
EBLUP 1	171	180	6	17	27	201
EBLUP 2	***	***	1	41	66	73
EBLUP 3	17	25	248	17	26	229
EBLUP 4	***	***	2	25	41	67

Note: a. ARB = 100 x Absolute Bias / Population Total.
b. ARE = 100 x Absolute Error / Population Total.
c. This indicates the number is greater than or equal to 1000.

Table 14. The Performances of the Aggregated[a]
Small Area Estimators for EMP
(averaged over 19 3-digit SICs)

Estimator	Unbenchmarked			Benchmarked		
	ARB	RRMSE[b]	REFF	ARB	RRMSE	REFF
Expansion	2	118	100	4	124	100
Synthetic 1	13	76	356	13	76	428
Synthetic 2	***[c]	***	0	13	76	423
Composite 1	11	79	218	96	354	49
Composite 2	***	***	6	96	353	50
EBLUP 1	245	***	126	9	90	217
EBLUP 2	***	***	0	25	158	93
EBLUP 3	7	82	225	6	97	208
EBLUP 4	***	***	0	21	631	54

Note: a. Aggregated over 10 provinces for each 3-digit SIC.
b. RRMSE = 100 x Root Mean Square Error / Population Total.
c. This indicates the number is greater than or equal to 1000.