SSC Annual Meeting, June 1997
Proceedings of the Survey Methods Section

REDUCTION OF BIAS IN PRELIMINARY ESTIMATES BASED ON INCOMPLETE POPULATION

Myra Andrews[1] and Lenka Mach[2]

ABSTRACT

The annual estimates of income and expenses for small businesses are produced by compiling data reported by these businesses on their tax returns to Revenue Canada. It takes two years before the universe of the small businesses for a given reference year is complete. In order to improve timeliness, preliminary estimates are produced at the time when we believe to have about 90% of the population. Thus the preliminary estimates are based on a stratified sample that represents the incomplete population and on an adjustment that brings the estimates to the level of the complete population. The adjustment is based on an assumption that the rate of the universe completion for year *y* is the same as in year *y-1* within each adjustment cell. In this paper we will show the bias of the preliminary estimates calculated using the current production system. We will also study the bias for estimates adjusted using different models in order to find the method that minimizes the bias.

KEY WORDS: Completion rate; weight adjustment; adjustment cell.

RÉSUMÉ

Les estimations annuelles des revenus et des dépenses pour les petites entreprises sont produites en compilant les données fournies par ces entreprises dans les rapports d'impôt à Revenu Canada. Deux ans s'écoulent avant que l'univers des petites entreprises pour une certaine année référentielle soit enfin complété. Afin d'améliorer les délais, des estimations préliminaires sont produites lorsque nous croyons disposer d'environ 90% de la population. Donc, les estimations préliminaires sont basées sur un échantillon stratifié qui représente la population incomplète et sur un ajustement qui amène les estimations au niveau de la population complète. L'ajustement est basé sur l'hypothèse que le taux de complétion de l'univers pour l'année *y* est le même à l'année *y-1* dans chaque cellule d'ajustement. Dans cet article, nous allons calculer les biais des estimations préliminaires en utilisant le système de production courante. Nous allons aussi étudier le biais pour les estimations ajustées en utilisant différents modèles pour trouver la méthode qui minimise le biais.

MOTS CLÉS: Taux de complétion; ajustement pondéré; cellule d'ajustement.

1. INTRODUCTION

The data from the income tax returns and attached income statements can be used to produce annual estimates for different financial variables. Presently, the Tax Estimates Program (TEP) at Statistics Canada (STC) uses this type of data, that is shortly referred to as *tax data*, to produce annual estimates of income and expenses for small businesses. A *small business* is currently defined as a business with annual Gross Business Income (GBI) ≥ $ 30 000 that does not belong to the list of large businesses that is maintained by STC. Large businesses are excluded from the TEP program because the annual financial data for them are collected *via* STC surveys. The obvious benefit of using tax data instead of survey data for the small businesses is a *lower cost* compared to a survey and, most importantly, *reduced respondent burden* imposed on the population of small businesses.

A description of the TEP program and its recent methodology can be found in Armstrong, Block and Srinath (1993). The sampling frame consists of two parts: i) a list of individuals that submitted a T1 form and are self-employed, that will be referred to as *T1 filers*, and

[1] Myra Andrews, School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6.

[2] Lenka Mach, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

ii) a list of corporations that submitted a T2 form and will be referred to as *T2 filers*. A filer is part of the population for year y if its business fiscal year end falls in the calendar year y. The frame for year y is being created at Revenue Canada/Taxation (RCT) during their processing of tax returns for year y: every time a tax return is assessed by RCT, a T1 or a T2 record is added to the frame. Bernoulli sampling is used to facilitate the selection during the processing cycle; every time a record is added to the frame, a decision is made whether or not the record is selected in the sample. Different sampling fractions are used for different sampling strata that are determined by a cross-section of the following variables: i) type of the filer (T1, T2, paper-filer or electronic filer), ii) standard industrial code, iii) province/ territory code, and iv) size (measured by GBI).

As mentioned above, sampling takes place during the processing cycle at RCT, well before the frame is complete. This enables STC to start its data processing in advance and thus improve the timeliness of the estimates. Even though the data is obtained from an administrative source, its processing requires considerable amount of time: the data must be transcribed from the income statements that were submitted in a free format, captured for all filers that submitted paper returns, also the industrial classification needs to be reviewed for all businesses in the sample and, finally, the data needs to be edited before the estimates can be produced. The pr
cycle of T1 records lasts over a year (February of year $y+1$ till April $y+2$) and the processing of T2 data takes two years (March of year y till April of $y+2$). Thus the final estimates based on the complete sample can only be produced in May of year $y+2$. Fortunately, a large proportion of the tax returns is usually processed by the fall of year $y+1$ and therefore it should be possible to produce reliable preliminary estimates before the population is complete.

In the following section, we will describe the current methodology that is used at STC to produce the preliminary TEP estimates as well as the other methods that we had studied. In Section 3 we will present the results measuring the bias of the preliminary estimates for the different methods. Finally, we will make recommendations as to which method should be used to produce the estimates.

2. PRELIMINARY ESTIMATION METHODOLOGY

The preliminary estimates that are based on an incomplete sample, representing an incomplete population, are obtained by applying an adjustment to the estimate produced by a regular estimation technique. It is assumed that the regular estimation technique yields a nearly unbiased estimate for the incomplete population. Thus to obtain an estimate for the entire population of interest, the estimate must be adjusted based on some model assumptions since we have to *predict* for the part of the population that is missing at the time of the preliminary estimation.

2.1 Regular Estimator Based on the Complete Sample

The TEP program uses a post-stratified ratio estimator as the regular estimator of totals of different financial variables for the population that is represented by the sample. As mentioned in Section 1, Bernoulli sampling must be used in order to be able to select the sample during the process of building the frame at RCT. The sample size in this situation becomes a random quantity and the expansion estimator is then less efficient than in the case of simple random sampling. The post-stratified ratio adjustment improves the efficiency of the estimates based on Bernoulli sampling and also it calibrates the estimated totals with the known population totals.

Currently, a two-phase design is used by the TEP program and thus the formulae for the regular estimator of the total of variable y for a domain d can be written as

$$\hat{Y}_d = \sum_d W1_i\, W2_i\, y_i\,, \qquad (1)$$

where $\sum_d$ is a summation over all units in the second-phase sample that belong to the domain d, y_i is the value of the variable y observed for the business i and $W1_i$ and $W2_i$ are the post-stratified first-phase and second-phase weights, respectively, associated with the business i. Each of the two weights is a product of a design-based weight, $w_i = 1/\pi_i$, and a post-stratified ratio adjustment, g_i: $W1_i = w1_i g1_i$, and $W2_i = w2_i g2_i$. More details can be found for example in Armstrong and St-Jean (1994).

2.2 Current Method for Producing Preliminary Estimates

As mentioned in the Introduction, the TEP frame is complete only in April of year $y+2$ and thus the final estimates for year y can be available in May of year $y+2$ which is quite late. However, during the processing cycle at RCT, a high volume of tax returns is processed during the peak period and only a relatively low volume during the last few months. It has been observed over the years, that about 90% of the businesses are on the frame by September of year $y+1$ and, by January $y+2$, the frame is about 96% complete, and that these global completion rates are quite stable from year to year.

A method for producing preliminary estimates,

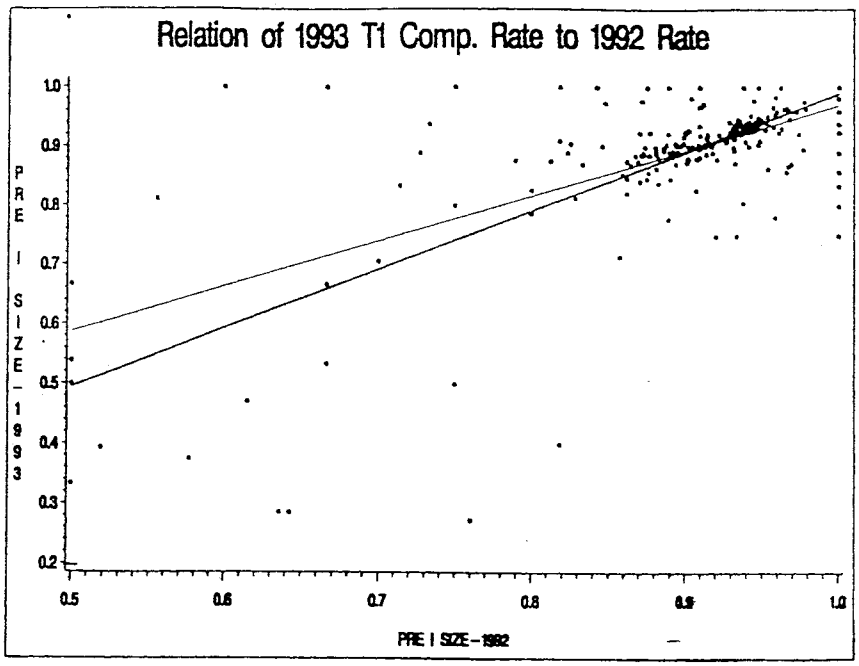
based on the assumption that the population completion rates remain stable from year to year, was proposed and studied by Beaucage (1992). The completion rates are calculated *within classes* of similar businesses however the classes must contain a sufficient number of units for the rates to be stable. Then the inverse of the completion rate is applied to the regular estimate in order to adjust it so it predicts the total for the complete population. Thus this method is based on the following *model* assumption:

$$NP_{AC}^{Y} / N_{AC}^{Y} = NP_{AC}^{Y-1} / N_{AC}^{Y-1} + \epsilon, \quad (2)$$

where NP_{AC}^{Y} is the population size of the adjustment cell AC at the time of the preliminary estimation for year y, N_{AC}^{Y} is the size of the complete population in the adjustment cell AC for year y and ϵ is a random error with mean zero. The preliminary estimate is then obtained by replacing $W1_i$ in (1) by an adjusted first-phase weight, $W1_i^{PRELIM} = (1 / CR_{AC}^{Y-1}) \; W1_i$, where $CR_{AC}^{Y-1} = (NP_{AC}^{Y-1} / N_{AC}^{Y-1})$ is the last year completion rate observed in the adjustment cell AC, where the unit I falls into.

Currently, the adjustment cells are defined by a cross-section of three variables: i) type of the filer, ii) a two-digit Standard Industrial Classification code (SIC2), and iii) province/ territory code. Only two types of filers are being distinguished - T1 and T2 filers. This cross-classification however yields a number of very small adjustment cells with quite unstable completion rates. Figure 1 below shows the relationship between the 1992 and 1993 completion rates used for Preliminary I Estimates (produced in September of $y+1$) for some selected industries. For the model assumption to hold the points would have to lie around the regression line that passes through the origin, however many outliers can be seen on Figure 1.

Figure 1



2.3 Studied Preliminary Estimation Methods

The analysis of the relationship between the completion rates for two consecutive years showed that most of the instability occurs in the small adjustment cells. For example, a large majority of the outliers on Figure 1 are completion rates in very small cells. Therefore a logical step to achieve an improvement without major changes to the production systems was to continue using the same method and model but with a redefined definition of the adjustment cell. The results of two alternations of the current method will be presented:

1) Regional Method: the adjustment cells are defined by a cross-section of three variables: i) type of the filer (T1 and T2), ii) a two-digit Standard Industrial Classification code (SIC2), and iii) a region code. Five regions were created : Atlantic provinces, Quebec, Ontario, Prairies and North-West Territories, British

Columbia and Yukon. In each region, there is one taxation centre where all the T2 tax returns are processed.

2) SIC2 Method: the adjustment cells are defined by a cross-section of two variables: i) type of the filer, and ii) a two-digit Standard Industrial Classification code (SIC2).

In addition, we also tried a method based on a different assumption, an assumption that a growth factor for each cell can be predicted by comparing the size of the population at the time of the preliminary estimation for two consecutive years:

$$NP_{AC}^{Y} \; / \; NP_{AC}^{Y-1} \; = \; N_{AC}^{Y} \; / \; N_{AC}^{Y-1} \; + \; \epsilon \; . \quad (3)$$

This method will be referred to as

3) Growth Factor Method : The size of the complete population for year y is predicted by applying the predicted net growth factor to the size of the complete population for year y-1: $N_d^Y = NG_{AC}^{\;Y} \; N_d^{Y-1}$. $NG_{AC}^{\;Y}$ is predicted by the left-hand side of (3) and the adjustment cells are defined by a cross-section of three variables: i) type of the filer (T1 and T2), ii) a two-digit Standard Industrial Classification code (SIC2), and iii) a region code.

3. BIAS OF THE PRELIMINARY ESTIMATES

In order to evaluate the bias, we compared the preliminary estimates, $\hat{Y}_d^{PRELIM}$, that were obtained for the current and the three studied methods, with the known population total, Y_d , for each domain d. Domains were defined by a cross-section of three variables: i) type of the filer (T1 and T2), ii) a two-digit Standard Industrial Classification code (SIC2), and iii) province/ territory code. Thus the domain coincides with the adjustment cell used by the current preliminary estimation method. The bias evaluation was done for two variables that are available for the population after it has been competed: i) size in terms of the number of units, and ii) size in terms of the total revenue. The bias was evaluated for different years.

Because we wanted to specifically evaluate the bias due to the adjustment used to bring the estimates for the incomplete population to the level of the complete population, the $\hat{Y}_d^{PRELIM}$ for the current, regional and SIC2 methods was calculated by applying the inverse of the completion rate directly to the known domain total of the incomplete population. Completion rates based on the number of units as well as on the total revenue were used. The $\hat{Y}_d^{PRELIM}$ estimate for the growth factor method was calculated as described in Section 2.3. Then the following measure, *absolute bias in %*, was calculated for every domain d and each preliminary estimation method :

$$AB\%_{\,d} = \{ \mid \hat{Y}_d^{PRELIM} - Y_d \mid / \; Y_d \} \; \text{x} \; 100. \quad (4)$$

To summarize the results, means and standard deviations of $AB\%_{\,d}$ were calculated. A sample of typical results will be presented in this Section; more results can be found in Andrews (1996).

In Table 1 below, the *mean absolute bias in %* is shown for the Preliminary I estimates of the population size in terms of the number of units. Note that the *Preliminary I estimates* are usually produced in September of year $y+1$ when it is assumed that the frame contains 90% of the units. The results pertain to the T2 part of the population and selected industries of interest (logging, transportation, wholesale trade, retail trade and business services) and are given for Canada as well as by province. Domains were defined by SIC2 and province which coincides with the adjustment cells used for the current method.

Table 1

1994 T2 Preliminary I Size Estimates: mean *AB%*

PROVINCE	N	MNS	CURRENT	REGIONAL	SIC2	GROWTH
CD	225	900	4.6	3.9	3.8	9.3
AB	20	1493	1.9	1.6	2.2	2.7
BC	20	1584	2.9	2.9	2.1	2.9
MB	20	278	3.6	3.1	3.9	4.2
NB	20	196	3.3	4.6	2.9	5.1
NF	20	134	4.4	3.6	4.5	8.1
NS	20	200	3.3	2.7	2.6	4.7
NT	14	12	16.1	8.8	9.2	32.6
ON	20	3424	1.3	1.3	1.5	1.3
PE	17	23	10.1	7.0	6.6	17.5
PQ	20	2545	2.4	2.4	1.0	2.4
SK	20	237	3.7	3.3	4.1	9.0
YT	14	11	7.6	9.6	8.3	37.8

N = the total number of domains over which the mean *AB%* is calculated.

MNS = the mean number of businesses in each domain at the time of the preliminary estimate.

Table 1 shows that the mean bias is less than 5% for Canada and most of the provinces for all three methods that use the inverse of the completion rate. However, for the three methods, the mean bias is high for the two territories and PEI - these are the three provinces with a very low mean number of businesses per domain (MNS $\leq$ 25). The regional and the SIC2 method decreased the mean bias for North-West Territories and PEI yet the bias still remains high. The two methods actually lead to an increase in the bias for Yukon. Thus even though the mean bias for Canada was slightly improved by the regional and SIC2 methods, the estimates for small provinces remained unacceptably biased. The estimates produced by the growth factor method are very biased except for the largest provinces. The results shown are for 1994, but very similar results were also obtained for the 1991-1993 years.

The growth factor method yielded very poor preliminary estimates and therefore it was excluded from further studies. Since none of the studied methods gave acceptable results for small domains, we introduced a minimum domain size: $\hat{Y}_d^{PRELIM}$ was calculated only if the number of units in the domain d was $\geq$ *MIN*. The minimum must have been attained for both the previous and current year, *i.e.* $NP_d^{Y-1} \geq MIN$ and $NP_d^{Y} \geq MIN$. Even though the estimates were not calculated for these small domains, the units included in them were still used for calculating the completion rates for the regional and SIC2 methods. In Table 2 below, mean and standard deviation of *AB%*, calculated over all domains of interest (the five industries and all provinces) are shown for the 1994 T2 population and a choice of minimum domain sizes.

Table 2

1994 T2 Preliminary I Size Estimates: mean *AB%* and standard deviation of *AB%*

			Mean *AB%*			Standard Deviation of *AB%*		
MIN	N	MNS	CURR	REG	SIC2	CURR	REG	SIC2
1	225	900	4.6	3.9	3.8	8.0	5.2	4.8
5	199	1017	3.5	3.0	2.8	5.0	3.7	2.8
10	187	1082	3.3	2.7	2.5	4.6	3.3	2.2
15	179	1130	3.0	2.6	2.3	4.2	3.2	2.1
25	167	1209	2.6	2.4	2.2	3.8	3.0	1.8

MIN = required minimum number of units to be included in a domain.

It can be observed in Table 2 that a considerable drop in the mean bias as well as in the standard deviation of *AB%* is achieved when the domains with less than five units are excluded. Further improvements are gained as the minimum domain size increases however that also significantly decreases the number of domains for which the preliminary estimates can be provided. For example, estimates can only be provided for 167 domains out of the existing 225 domains when we apply a minimum domain size of 25. Similar results were obtained for the years 1991-1993 for the T2 population and also for the 1994 T1 population.

So far we presented measures of bias for the preliminary estimates of the size in terms of the number of businesses. However preliminary estimates of the financial variables are also needed. Measures of bias were obtained for the estimates of the total revenue. Table 3 shows again the results for the 1994 T2 population and the five industries of interest. Table 3A contains results obtained using completion rates in terms of the number of units. The mean bias never drops below 5% and we can also see that there is a lot of variation in the bias between the different domains. A significant improvement was achieved when all businesses with GBI ≥ 7 million were excluded and revenue completion rate was used as one can see in Table 3B.

Table 3

1994 T2 Preliminary I Revenue Estimates: mean *AB%* and standard deviation of *AB%*

Table 3A: Size Completion Rates; All units.

			Mean *AB%*			Standard Deviation of *AB%*		
MIN	N	MNS	CURR	REG	SIC2	CURR	REG	SIC2
1	225	900	7.2	6.4	6.2	10.4	8.3	8.1
5	199	1017	6.6	5.9	5.7	8.8	8.0	7.7
10	187	1082	6.5	5.8	5.5	8.9	8.2	7.8
15	179	1130	6.3	5.8	5.5	8.8	8.3	7.9
25	167	1209	6.2	5.8	5.5	8.9	8.6	8.2

Table 3B: Revenue completion rate; Units with GBI < $ 7,000,000

5	199		3.4	3.0	2.9	5.8	3.6	3.0
10	187		3.2	2.9	2.7	5.8	3.4	2.6
15	179		3.0	2.7	2.5	5.5	3.2	2.2

To summarize, we found that the bias of the Preliminary I estimates, that are produced when the frame is about 90% complete, could be quite high. Only a slight improvement can be achieved by using the regional or SIC2 method. However a significant improvement in the size estimates can be realized when small domains are excluded.

To reduce the bias of the revenue estimates, the size of a business, measured for example by GBI, must be incorporated into the prediction model. By eliminating the small domains as well as the largest businesses a significant improvement in the prediction of the total revenue was obtained. That, on the other hand, has the negative impact of not being able to provide preliminary figures for all the domains of interest. In order to improve the quality and timeliness of the preliminary estimates, different prediction models should be studied

REFERENCES

Andrews, M. (1996). Study of the bias in the preliminary estimates. Report on Preliminary Estimates Project, Business Survey Methods Division, Statistics Canada.

Armstrong, J., Block, C., and Srinath, K.P. (1993). Two-phase sampling of tax records for business surveys. *Journal of Business and Economic Statistics*, 11, 407-416.

Armstrong, J., and St-Jean, H. (1994). Generalized regression estimation for a two-phase sample of tax records. *Survey Methodology*, 20, 97-105.

Beaucage, Y. (1992). Estimation pour une population incomplète basée sur une variable auxiliare. Paper presented at the ACFAS conference.