RAKING RATIO ESTIMATION
AN APPLICATION TO THE CANADIAN MONTHLY RETAIL TRADE SURVEY

Z. Patak and M.A. Hidiroglou, Statistics Canada
Z. Patak, 11-R, R.H.Coats Bldg., Tunney's Pasture, Ottawa, Canada, K1A 0T6

Key Words: raking ratio, generalized regression, poststratification

1. Introduction

The Canadian Retail Sector is a very important segment of the Canadian economy both in terms of employment and revenue. Statistics that measure the amount of retail activity are used by all levels of government to develop national and regional economic programs and policies. In addition, the statistics are used by businesses, trade associations and others to assist in decision making, marketing efforts and to assess business conditions.

In this article we shall focus on the Canadian monthly Retail Trade Survey (MRTS). This survey collects **sales** and **locations** on a monthly basis for all sampled companies; **inventories** are collected monthly for a selected subset (companies whose annual sales exceed a preset threshold). In addition to the uses mentioned above, the collected information serves as an important indicator of personal expenditure in particular, and of Canadian economic performance in general.

The target population for MRTS consists of all statistical companies operating in Canada that have at least one statistical retail location within their structure. These units comprise the sampling list frame that resides on the Statistics Canada Business Register (BR). The BR is a database that stores information about business entities operating in Canada. It makes use of a number of different files from Revenue Canada Taxation and data collected by surveys to maintain and update existing information.

Auxiliary data are readily available for many surveys. For business surveys conducted by Statistics Canada many auxiliary variables are updated monthly. They represent the most current information related to the parameters of interest and their use in the estimation process may significantly improve the reliability of estimates without increasing the sample size.

One of the auxiliary variables available on the BR is known as the gross business income. It is well correlated with sales for many industry trade groups within MRTS. In this article, we will illustrate how this auxiliary variable can be used to improve the reliability of sales estimates.

A brief overview of the frame and the MRTS sample design are given in Section 2 and Section 3. The raking algorithm and estimation are described in Section 4 and Section 5. Numerical results are discussed in Section 6. Concluding remarks are summarized in Section 7.

2. MRTS Frame

To maintain the BR in a cost effective manner, the business entities on the frame are classified into three categories: IP (Integrated Portion), NIP (Non-Integrated Portion) and ZIP (out of scope for surveys). A business is part of IP when its gross business income (GBI) exceeds a predetermined upper threshold for its geographical and industrial classification. NIP/ZIP classification differs for annual and sub-annual surveys, but in general, businesses whose GBI is greater than $25,000 and less than the upper threshold belong to NIP.

GBI is an annualized size measure that is derived from the product of the following three terms: (a) a ratio of operating income to total wages and salaries, (b) a ratio of total wages and salaries to total remittances, and (c) the annualized sum of remittances received in the past M months, where M is at most 12. The use of GBI on the list frame is twofold. It is used to assess if a business entity shows IP tendencies (i.e. tends to be large) and thus should be profiled to determine if it belongs in the IP. It is also used by MRTS as a size stratification variable for births. Falardeau and Charron-Corbeil (1993) provide more details on the derivation of GBI.

3. MRTS Sample

The MRTS population is stratified by geography, trade group and size. The geography variable encompasses 12 provinces and territories as well as four census metropolitan areas; the trade group variable represents 18 groups of similar retail activities. The size variable splits the population into three size strata: (a) take-all, TA, (units are selected with certainty), (b) large take-some, LTS, (units are selected with a high probability), and (c) small take-some, STS, (units are selected with a low probability).

MRTS employs rotation group sampling. The population each stratum is randomly grouped into clusters, or rotation groups, and then a random sample of rotation groups is selected. The number of rotation groups in and out of sample depends on the sampling fraction in the stratum, as well as time-in and time-out constraints. The procedure for determining the number of rotation groups in the sample and population under time-in and time-out constraints is described by Hidiroglou, Choudhry and Lavallée (1991).

The parameter of interest that we wish to estimate is total sales. Let it be denoted by y and let y_{hij} be its value for the *j-th* unit in rotation group i of stratum h. Let $\delta_{hij}\ (d)$ be an indicator variable defined as 1 if the *hij-th* unit belongs to domain d, and 0 otherwise. The population total $Y(d)$ is given by

$$Y(d) = \sum_{h=1}^{L} \sum_{i=1}^{P_h} \sum_{j=1}^{N_{hi}} y_{hij}(d) \ ,$$

where N_{hi} is the number of units in rotation group i in stratum h, and $y_{hij}(d) = \delta_{hij}\ (d)\ y_{hij}$. Let $y_{hi}\ (d)$ be the total response of the units belonging to domain d from the *i-th* sampled rotation group in stratum h, i.e.

$$y_{hi}(d) = \sum_{j=1}^{N_{hi}} y_{hij}(d), \ \mathrm{i} = 1,2,\dots,n_h \ .$$

We wish to estimate total sales at different levels of geographical, industrial and size aggregation. The corresponding statistic is a stratified π-expanded estimate of total sales adjusted for the realized sample size, i.e

$$\hat{Y}(d) = \sum_h \frac{N_h}{\hat{N}_h} \sum_{i \in s_h} \frac{P_h}{p_h} y_{hi}(d) = \sum_h \frac{N_h}{n_h} \sum_{i \in s_h} y_{hi}(d),$$

where P_h is the number of panels in the population in stratum h, p_h is the number of in-sample panels in stratum h, n_h is the number of in-sample companies in stratum h and $\hat{N}_h = n_h P_h \ / p_h$.

The corresponding variance estimate is

$$v(\hat{Y}(d)) = \sum_h \left(\frac{N_h}{\hat{N}_h} \right)^2 P_h^2 \left(1 - \frac{p_h}{P_h} \right) \frac{\sum_{i=1}^{p_h} \{ y_{hi}(d) - \bar{y}_h(d) \}^2}{p_h - 1},$$

where $\bar{y}_h(d) = \sum_{i=1}^{p_h} y_{hi}(d) / p_h$ is the mean per panel for domain d of the sampled panels in stratum h. We use domain estimation to compensate for the lack of most current stratification, or to partition multiple trade group companies into the appropriate estimation bins. Currently, MRTS yields levels of reliability for total sales that are close to the ones specified in the design. Can these be improved using auxiliary data such as the gross business income?

4. Raking

In the NIP portion of the frame GBI is updated monthly using remittance data made available to Statistics Canada by Revenue Canada Taxation. One way of incorporating GBI into MRTS estimates is through the use of the raking ratio estimator in which auxiliary information is used to adjust survey weights by raking on known marginal totals. This is the approach that we shall pursue in this article.

The MRTS sample is stratified by size into take-all, large take-some and small take-some strata. The biggest contribution to the variance of total sales can be traced back to the small take-some strata. This is where we shall incorporate auxiliary information into the estimation process. The small take-some strata include units from both IP and NIP, but GBI gets refreshed monthly in the NIP only. The IP GBI is updated sporadically which may cause its correlation with sales to degrade over time. This imposes a constraint on how the raking algorithm is applied.

The small take-some strata are poststratified into their IP and NIP components. The IP component is set aside and treated as poststratified at estimation. For the intersection of the sample with the NIP component we compute the correlation between sales and GBI. Cochran (1977, p.157) recommends the use of the ratio estimator over the expansion estimator whenever the correlation between sales and GBI exceeds $0.5 \left[\frac{CV(GBI)}{CV(Sales)} \right]$, where $CV(u) = \frac{SD(u)}{Mean(u)}$. We use the same test for the raking ratio estimator and add the condition that there be at least five observations within an estimation cell. If the two conditions are not satisfied, then we revert to the existing estimation scheme.

Denote strata as $\{U_h,\ h=1,2,\dots,L\}$. Strata where raking is applied to their NIP component will be decomposed into two subsets, (i) the NIP subset $U_{h,NIP}$ where raking is performed, and (ii) the IP subset $U_{h,IP}$ where poststratified estimation takes place. Note that $U_h = U_{h,IP} \cup U_{h,NIP}$ for these strata. Corresponding population sizes will be N_h, $N_{h,IP}$ and $N_{h,NIP}$. The samples will be denoted by s_h, $s_{h,IP}$ and $s_{h,NIP}$ with corresponding sample sizes n_h, $n_{h,IP}$ and $n_{h,NIP}$.

Before proceeding with raking, an *estimated annual sales* adjustment has to be applied to the auxiliary data to maximize their utility. This adjustment is used to make GBI that is an annual figure comparable with sales that is a monthly figure. The *estimated annual sales* adjustment involves the calibration of GBI to annualized monthly sales for in-sample units.

The contribution of each calendar month to the estimated annual total sales is expressed as a ratio at the stratum level (trade group by geography). The reported monthly sales are then multiplied by the inverse of the ratio to obtain estimated annual sales based on the pertinent month. For each stratum a factor $\sum_{k \in s_h} \textit{annualized sales}_k \,/\, \sum_{k \in s_h} \textit{GBI}_k$ is computed. This factor is used to inflate or deflate population GBI so that $\sum_s \textit{adjusted GBI} = \sum_s \textit{annualized sales}$. For a given stratum h, we denote the value of the adjusted GBI for unit k $(k \in U_{h,NIP})$ by x_{hk}.

After making the necessary adjustments to the auxiliary data, the raking procedure is invoked. Raking estimation is chosen for three reasons: (a) unlike typical ratio estimation that preserves auxiliary totals in only one dimension, raking ratio estimation preserves auxiliary totals in all dimensions; (b) in the presence of sparse cells, raking ratio estimation offers a degree of protection against outliers while reducing ratio bias; and most importantly, (c) the size stratification which is based on GBI may be incorrect as GBI is known to be highly variable during the first six months in the life of a business entity. Raking compensates for the lack of correct size stratification by partitioning the population into true NIP cells and *pseudo* IP cells.

For each geographical partition and small take-some strata a 2-dimensional data grid G is constructed with the corresponding marginals. The first dimension represents the trade group variable. To generate the second dimension, sampled units are divided into *small GBI* (true NIP) and *large GBI* (*pseudo* IP) cells. The split is based on NIP-IP revenue thresholds available at the provincial level on the BR.

Let the set of strata belonging to this grid be G_p for a given province p. Then $U_{h,NIP}$ is partitioned into large $U_{h,NIP,1}$ and small $U_{h,NIP,2}$ populations, where $NIP,1$ represents true NIP units while $NIP,2$ represents *pseudo* IP units. An iterative process then adjusts the sampling weights in each raking cell by cycling through the data grid until the sums of the weighted GBI in both dimensions are equal to the marginal totals defined by $X_{\bullet r}$ and $X_{h\bullet}$ below.

Let $X_{\bullet r} = \sum_{h \in G_p} \sum_{i \in U_{h,NIP,r}} x_{hri}$, $r = 1,2$;

$X_{h\bullet} = \sum_{r=1}^{2} \sum_{i \in U_{h,NIP,r}} x_{hri}$, $h \in G_p$ and $\hat{X}_{hr}^{(0)} = \sum_{i \in s_{h,NIP,r}} \frac{P_h}{p_h} x_{hri}$.

The raking step is then

$$\hat{X}_{hr}^{(t)} = \begin{cases} \hat{X}_{hr}^{(t-1)} \dfrac{X_{\bullet r}}{\hat{X}_{\bullet r}^{(t-1)}} & \text{for t odd,} \\[2ex] \hat{X}_{hr}^{(t-1)} \dfrac{X_{h\bullet}}{\hat{X}_{h\bullet}^{(t-1)}} & \text{for t even.} \end{cases}$$

Let the converged raked values of $\hat{X}_{hr}^{(t)}$ be denoted by $\hat{X}_{hr}^{(c)}$. The resulting sample weight adjustment is $g_{hr} = \dfrac{\hat{X}_{hr}^{(c)}}{\hat{X}_{hr}^{(0)}}$ for $h \in G_p$ and $r \in U_{h,NIP,r}$ .A different g_{hr} is produced for each raking cell.

5. Estimation

The MRTS sample contains approximately 1,000 large companies that operate in multiple geographies and/or trade groups. For the purpose of stratification, these units are assigned their dominant geography and trade group (i.e the origin of the largest contribution to their total GBI). At estimation, the reported sales are distributed into estimation domains to more accurately measure retail activity in individual geography by trade group cells. A similar approach is taken when sampled units are misclassified, i.e. reported sales are assigned to a domain that is different from the stratum.

Take-all, large take-some and small take-some strata, where gains cannot be made from using auxiliary data, will use the current domain estimation scheme. In the small take-some strata, to whose NIP portion raking has been applied, the sample weight will be adjusted by the corresponding g-weight. Thus the hybrid Horvitz-Thompson/generalized regression estimator of the domain total is $\hat{Y}(d) = \hat{Y}_{NR}(d) + \hat{Y}_{POST}(d) + \hat{Y}_{RAK}(d)$, where NR denotes nonraked strata, $POST$ is the IP portion of poststratified (into IP and NIP) strata, and RAK is the NIP portion of raked strata, i.e.

$$\hat{Y}_{NR}(d) = \sum_{h \in NR} \sum_{i \in s_h} \frac{N_h}{n_h} y_{hi}(d),$$

$$\hat{Y}_{POST}(d) = \sum_{h \in POST} \sum_{i \in s_{h,IP}} \frac{N_{h,IP}}{n_{h,IP}} y_{hi}(d) \quad \text{and}$$

$$\hat{Y}_{RAK}(d) = \sum_{h \in RAK} \sum_{i \in s_{h,NIP,r}} \frac{N_{h,NIP}}{n_{h,NIP}} g_{hr} y_{hi}(d) \ , \ r = 1,2 \ .$$

The estimated variance for $\hat{Y}(d)$ is made up of the same three components. The first two are

$$v(\hat{Y}_{NR}(d)) = \sum_{h \in NR} \left(\frac{N_h}{\hat{N}_h} \right)^2 P_h^2 \left(1 - \frac{p_h}{P_h} \right) \frac{\sum_{i \in s_h} (y_{hi}(d) - \bar{y}_h(d))^2}{p_h - 1},$$

$$v(\hat{Y}_{POST}(d) | IP) = \sum_{h \in POST} \left(\frac{N_{h,IP}}{\hat{N}_{h,IP}} \right)^2 P_{h,IP}^2 \left(1 - \frac{p_{h,IP}}{P_{h,IP}} \right) \frac{\sum_{i \in s_{h,IP}} (y_{hi}(d) - \tilde{y}_h(d))^2}{p_{h,IP} - 1},$$

where $\bar{y}_h(d) = \sum_{i \in s_h} y_{hi}(d) / p_h$ and $\tilde{y}_h(d) = \sum_{i \in s_{h,IP}} y_{hi}(d) / p_{h,IP}$.

The variance for $\hat{Y}_{RAK}(d)$ can be invoked from Deville and Särndal (1992). The regression model that corresponds to raking is a fixed-effect two-way analysis of variance with column effects ρ_h and row effects γ_r .It is given by $y_{hri}(d) = (\rho_h + \gamma_r) x_{hri} + \epsilon_{hri}$ for $i \in U_{h,NIP}$ and $h \in G_p$, where it is assumed that $E_\xi(\epsilon_{hri}) = 0$, $V_\xi(\epsilon_{hri}) = \sigma_h^2$ and $Cov_\xi(\epsilon_{hri}, \epsilon_{hrj}) = k_h \sigma_h^2$ for all $i \neq j \in U_{h,NIP}$, where k_h is a negative constant. The predicted value of $y_{hri}(d)$ is given by $\hat{y}_{hri}(d) = (\hat{\rho}_h + \hat{\gamma}_r) x_{hri}$. The last component of the estimated variance for $\hat{Y}_{RAK}(d)$ is

$$v(\hat{Y}_{RAK}(d) | NIP) = \sum_{h \in RAK} P_{h,NIP}^2 \left(1 - \frac{p_{h,NIP}}{P_{h,NIP}} \right) \frac{\sum_{i \in s_{h,NIP,r}} (g_{hri} e_{hri}(d) - \overline{g_{hri} e_{hri}(d)})^2}{p_{h,NIP} - 1},$$

where $r = 1,2$ and $e_{hri}(d) = y_{hri}(d) - \hat{y}_{hri}(d)$. The sampled IP component of the raked small take-some strata is generally quite small (<= 6 units). This would lead to unstable domain variances. To reduce the variability in the variance in the IP poststratum, the ordinary ratio variance $s_{h,NIP \cup IP}^2(d)$ across the corresponding NIP and IP poststrata is used as a good approximation to $s_{h,IP}^2(d)$.

6. Numerical Results

The numerical results that follow in this section have been produced using December, 1993, MRTS data. The retail population comprised 159,072 businesses of which 18,631 belonged to the IP and 140,440 to the NIP. A sample of 18,436 units was drawn for an overall sampling rate of approximately 11.6%.

Businesses in take-all and large take-some size strata were set aside as these were not to be raked. The small take-some size stratum was poststratified into IP and NIP to identify businesses eligible for raking. There were 4,650 units that satisfied the correlation and minimum sample size selection criteria. These units were assigned to 284 raking cells. Convergence in all raking grids was reached after five iterations of the raking algorithm. The resulting sampling weight adjustments, g-weights, ranged from 0.5 to 3.5.

The number of iterations to convergence and the range of the g-weights are indicative of a well behaved system. This means that the GREG variance estimator is likely to provide a good approximation to the raking ratio variance estimator even for small sample sizes. Although the methodology in this article provides results for cluster sampling, the numerical study was done at the element level. A large percentage of the clusters contain only a single business entity and the resulting design effect has little or no impact on the estimates.

The linear regression part of the numerical study does not account for the covariance between rotation groups within strata. This is conducive to producing mildly conservative variance estimates. The covariance factor k_h was of the order of $-\frac{1}{P_h(P_h - P_h)}$ and was deemed negligible for all practical purposes. Note that k_h varied between -0.0032 and -0.000035. Using sampling weights and the corresponding g- weights, estimates of total sales, variances and coefficients of variation for the domains of interest were produced. At the Canada level the strictly expansion MRTS estimate of total sales was $19.72 billion compared to $20.15 billion as obtained with raking weight adjustments. The change represents an increase of 2%. The standard errors were $175.9 million and $154.3 million for MRTS and raked MRTS; the resulting CV's were 0.892% and 0.766%. These results translate into a 16.4% improvement over existing methodology. This means that we may reduce the sample size by approximately 4,000 units and achieve the existing CV's for the Horvitz-Thompson estimator.

Figure 1 shows that similar improvements can be achieved at the provincial level, as well. The estimates of total sales have increased slightly as a result of the raking weight adjustment. Figure 2 illustrates the size of the change; it ranges between 0% and 3.13%.

Figure 1. Comparison of Raked vs MRTS CV's at the Province/CMA Level.

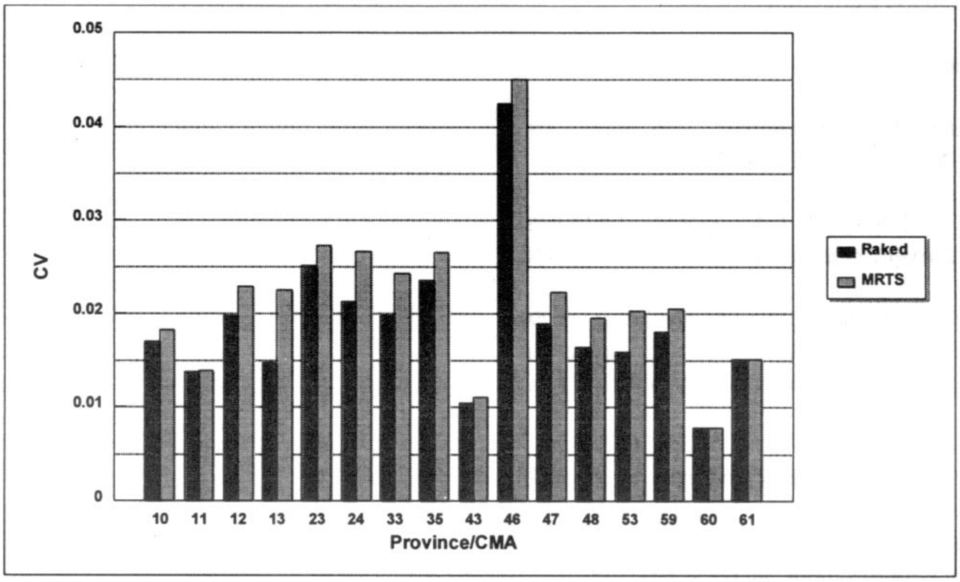
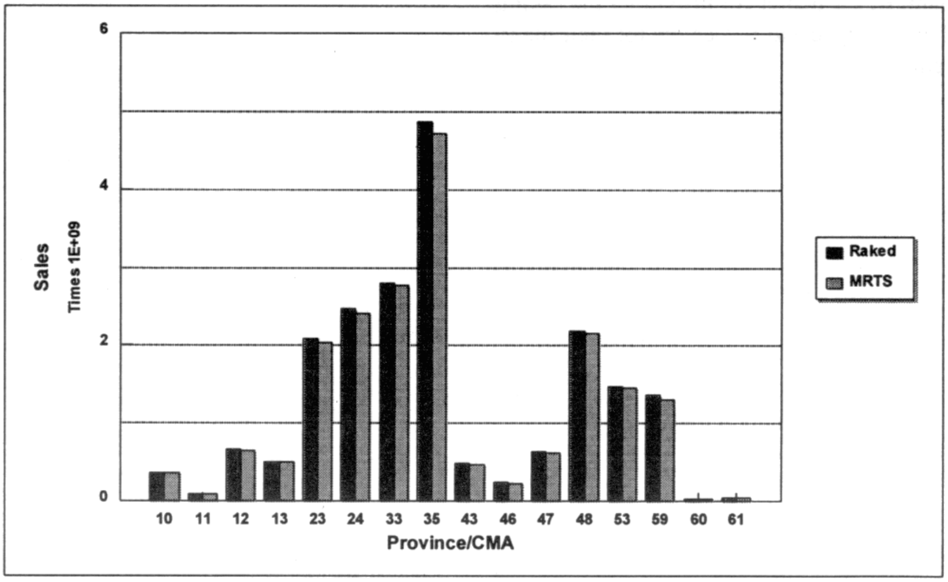


Figure 2. Comparison of Raked vs MRTS Estimated Total Sales at the Province/CMA Level.



The positive impact of incorporating auxiliary data in the estimation of total sales is also evident at the cell level (province by trade group). In many cells the improvement in efficiency due to raking is quite dramatic.

7. Summary

There are many estimation methods that take advantage of auxiliary data; we have chosen raking for its smoothing properties and a degree of resistance to outlying observations. The presence of auxiliary data provides more complete information about the population of interest and their usage often reduces the sampling variance of the estimated parameters. For MRTS this hypothesis has been corroborated by a numerical study.

The benefits of using auxiliary information are enhanced reliability of the estimates and/or reduced processing costs. To maintain the current target CV's, the take-some sample may be reduced by 4,000 units. If the size of the sample remains unchanged, more efficient estimates of total sales may be produced by incorporating auxiliary information into the estimation process.

8. References

Cochran, W.G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley & Sons.

Deville, J.C. and Särndal, C.E., (1993), *Calibration Estimators in Survey Sampling,* JASA, Vol. 87, 376-382.

Hidiroglou, M.A., Choudhry, H. and Lavallée, P., (1991), *A Sampling and Estimation Methodology for Sub-annual Business Surveys.* Survey Methodology, Vol. 17, No. 2, 195-210.

Falardeau, N. and Charron-Corbeil, M., (1993), *Improvements to Frame Data Quality in the Small Business Sector,* Proceedings of the International Conference on Establishment Surveys, 904-909.

9. Acknowledgements

The authors thank R. Carpenter for providing the raking algorithm, and to D. Royce and S. Demers for their many helpful comments.