Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 71–80

Raking Ratio Estimation: An Application to the Canadian Retail Trade Survey

Michael A. Hidiroglou[1] *and Zdenek Patak*[2]

National statistical agencies are increasingly using auxiliary information available from administrative sources. In this article, we will illustrate how auxiliary information can be used for monthly surveys collecting sales via ratio and raking ratio estimation. Although ratio estimation is widely used, raking ratio estimation has a number of desirable characteristics. It preserves auxiliary totals in all two or more dimensions. It also protects against outliers present in sparse cells, and ensures that the resulting calibrated weights are positive. We compare these two estimation procedures for the Canadian Monthly Wholesale and Retail Trade Survey.

Key words: Raking ratio; post-stratification; calibration.

1. Introduction

Several auxiliary variables related to size are available on Statistics Canada's Business Register (BR), a frame that maintains a list of Canadian businesses. Some of these auxiliary variables are available on a monthly basis, such as remittances, while others are available on an annual basis, such as operating income and total wages. They can be used to our advantage at either the sampling stage or subsequent estimation stage if they are well correlated with survey variables of interest. In terms of initial stratification or subsequent periodic restratification, the auxiliary data aid in determining stratum size boundaries, as well as the size of the sample and subsequent allocation to the resulting strata. Their use in the estimation process can result in improvements in the reliability of survey estimates.

In this article, we illustrate the use of one of the auxiliary variables available on the BR. This variable, known as Gross Business Income (GBI), is the gross income received from sales of products or services to customers. We will illustrate how this variable can be used to improve the reliability of the Canadian Monthly Retail Trade Survey (MRTS). MRTS collects *sales* and *locations* on a monthly basis for all sampled companies; *inventories* are collected monthly for a selected subset (companies whose expected annual sales exceed a preset threshold). The Canadian Retail Sector is an important segment of the Canadian economy, both in terms of employment and in terms of revenue. All levels of government

[1] Office for National Statistics, Newport U.K. Email: mike.hidiroglou@ons.gov.uk
[2] Statistics Canada, Ottawa, Canada. Email: zdenek.patak@statcan.ca
Acknowledgment: The authors gratefully acknowledge comments from an associate editor and two referees that improved this article.

use statistics measuring retail activity level to develop national and regional economic programs and policies. Businesses, trade associations and others use these statistics to assist in decision-making and marketing efforts, as well as to assess business conditions. The collected information also serves as an important indicator of personal expenditure, and of Canadian economic performance.

The article is structured as follows. An overview of the 1988 MRTS sample design is given in Sections 2 and 3. The current procedure and proposed estimation based on raking are described in Section 4. Results of a numerical study are given in Section 5, followed by some concluding remarks in Section 6.

2. Monthly Retail Trade Survey Frame

The BR is used as the principal frame for the economic statistics program of Statistics Canada. It is the central repository of information on businesses in Canada. The Business Register's role is to provide Statistics Canada with a comprehensive quality frame in terms of coverage and various stratification variables such as industrial classification, GBI, number of employees and total assets. The BR also stores statistical information regarding the composition of the population of businesses in Canada, in terms of their organization, industrial activity, size, geography, and dynamic evolution over time. Businesses are within the scope of subannual surveys if their annual GBI is larger than $25,000.

The Business Register is kept up-to-date using administrative sources as well as survey feedback. Survey feedback updates the frame with respect to industry, geography, true size, business structure, and contact information. Administrative sources include payroll deduction accounts (monthly basis), as well as tax records (annual basis).

The BR contains approximately one million businesses. It is not possible to instantly update all the businesses with their structure and classification information. It is for this reason that the BR is split into two main portions for maintenance purposes: large businesses and the remaining businesses. This split depends on maintenance thresholds that vary according to industry and province. The larger businesses, those above the maintenance thresholds, are continually updated with respect to their structure and classification information using administrative data, Business Register profiling activities, and survey feedback (on a subannual and annual basis). The remaining businesses are updated using subannual administrative sources in terms of births, deaths, and size measures (GBI or number of employees) and survey feedback. Their structures as well as contact information are updated using survey feedback.

The current MRTS was initially selected from the BR in October 1988. At the time of sample selection, auxiliary data were not available on the BR for the smaller businesses. Sample sizes were determined using a combination of the most recent Annual Retail Trade Survey and the previous Monthly Retail Trade Survey.

Starting in 1990, auxiliary data in the form of "synthetic" GBI was added to the BR for the smaller businesses. This synthetic income variable is computed as the product of: (i) the ratio of operating income to total wages and salaries; (ii) the ratio of total wages and salaries to total remittance; and (iii) the annualised sum of remittances received in the last M months, where M is at most twelve. Falardeau and Charron-Corbeil (1993) provide

more details on the derivation of GBI. For the larger businesses, GBI is obtained either as part of a business profile update or feedback from annual surveys.

3. Monthly Retail Trade Survey Sample

The target population for MRTS consists of all in-scope businesses operating in Canada that have some activity related to retail trade. Businesses are in-scope to subannual surveys if their GBI is larger than $25,000. Businesses in-scope to Retail Trade were stratified by geographic region, industry, and size. The geographic breakdown is by province and territories. Four of the provinces, Quebec, Ontario, Manitoba, and British Colombia, were further split into major Census Metropolitan Areas and the remaining portion within the province. The major Census Metropolitan Areas for Quebec, Ontario, Manitoba, and British Colombia are respectively Montréal, Toronto, Winnipeg, and Vancouver. The industry breakdown is by groups of 1980 Standard Industrial Classification (SIC) codes at the three and four digit levels (called trade groups). Each geographic region and trade group is further stratified by size into, at most, three substrata: one take-all (self-representing) substratum for large or complex businesses and up to two take-some substrata for medium and small businesses. Since GBI was not available on the BR in 1988, the boundaries delimiting the two take-some strata were set equal to the corresponding maintenance thresholds dividing the larger and smaller businesses.

The take-all boundary, based on a cut-off rule developed by Hidiroglou (1986), accounted for the fixed maintenance thresholds. An overall sample size (constrained by an overall Canada coefficient of variation constraint at the Canada level) was allocated to the resulting strata (province by trade group by size) in three steps.

In the first step, given the overall Canada target coefficient of variation, target coefficients of variation were computed (via a raking algorithm) within each province by industry combination to achieve nearly equal coefficients of variation at the corresponding provincial and industrial levels. In the second step, the target coefficients of variation were further split into two parts for provinces requiring census metropolitan area information. In the third step, using the target coefficients of variation computed in the second step, sample sizes were obtained for the three size strata within each geography and industry combination. A square root allocation (Bankier 1986), based on the population business counts, was used to allocate the sample between the two take-some strata.

Sample rotation of businesses within the take-some strata was carried out using a procedure described in Hidiroglou, Choudhry, and Lavallée (1991). In that scheme, population businesses within each take-some stratum h (geography by industry by size) are first randomly allocated to P_h rotation groups. The number of in-sample rotation groups (p_h) and out-of-sample rotation groups ($P_h - p_h$) depends on the sampling fraction in the stratum, as well as time-in and time-out constraints. Businesses that belong to the take-some strata are rotated in and out of the sample on a monthly basis, and businesses that belong to the sampled rotation groups are surveyed on a monthly basis as well. The initial sample (first survey occasion) consists of all businesses that belong to the first p_h rotation groups. On succeeding survey occasions, businesses are rotated by dropping and acquiring rotation groups. For instance, on the second survey occasion, the first rotation group is rotated out of the sample and the $(p_h + 1)$th rotation group is rotated into the sample.

Businesses rotating into the sample are surveyed for a maximum of 24 months. Once businesses rotate out, they stay out of sample for at least twelve months. Births are stratified according to the same criteria as the initial population. They are assigned randomly to the panels. The panel to which the last birth is assigned is retained so that births appearing the month after are assigned to panels starting from the panel next to it. This prevents panel sizes from varying by more than one unit within a given stratum. Births that happen to be assigned to in-sample panels are in sample. More details of the sample design of MRTS are available in Hidiroglou (1989) and Trépanier et al. (1998).

4. Estimation

We introduce some notation to describe the current and alternative estimation procedures for MRTS. The population of businesses is denoted as U and strata (province by trade group by size) of businesses as U_h, where $h = 1, \ldots, L$. A rotation group inherits the stratification of its business members. A rotation group $i\,(i = 1, \ldots, P_h)$ within a stratum h will be labeled as the hith rotation group, and businesses within the hith rotation group will be labeled as $j\,(j = 1, \ldots, N_{hi})$, where N_{hi} is the total number of businesses within the rotation group. A rotation group is effectively a cluster consisting of N_{hi} businesses. The total number of population businesses within stratum h is $N_h = \sum_{i=1}^{P_h} N_{hi}$. Define s_h as the set of in-sample rotation groups. The total number of sampled businesses within stratum h is $n_h = \sum_{i \in s_h} N_{hi}$.

The population total (monthly sales) is the parameter of interest to be estimated for a given domain U_d, where $U_d \subseteq U$. We focus on domain estimation for a number of reasons. Estimates can be obtained for arbitrary partitions of the sample. Furthermore, if businesses have changed their classification between time of sampling and time of estimation, such changes can be reflected in an unbiased fashion using domain estimation.

Let y be the variable of interest. Then y_{hij} will be the y-value for the jth business belonging to the hith rotation group. An indicator variable $\delta_{hij}(d)$ is defined as 1 if the hijth business belongs to domain U_d, and 0 otherwise. The population total for domain U_d is given by $Y(d) = \sum_{h=1}^{L} \sum_{i=1}^{P_h} \sum_{j=1}^{N_{hi}} y_{hij}(d)$, where $y_{hij}(d) = y_{hij}\delta_{hij}(d)$. The response of the businesses within the hith sampled rotation group belonging to domain U_d is $y_{hi}(d) = \sum_{j=1}^{N_{hi}} y_{hij}(d)$, where $i \in_{S_h}$.

The population total $Y(d)$ can be estimated in a number of different ways, depending on how auxiliary data are used.

4.1. Current Procedure

The Horvitz-Thompson estimator is given by $\hat{Y}_{HT}(d) = \sum_{h=1}^{L} \hat{Y}_h(d)$, where the individual strata components of $\hat{Y}_{HT}(d)$ are $\hat{Y}_h(d) = \sum_{i \in s_h} (P_h/p_h) y_{hi}(d)$. An advantage of this estimator is that it is unconditionally unbiased. A disadvantage is that it can be biased when conditioned on the realized number of businesses in the sample. Moreover, it may not be very efficient because it does not make use of the known auxiliary size information of the rotation groups. It may become more inefficient over time because death removal may increase the variation in the rotation group sizes. If the correlation between $y_{hi}(d)$ and the associated rotation group sizes N_{hi} is large, efficiency gains can be realised through the

separate ratio estimator:

$$\hat{Y}_{SRAT}(d) = \sum_{h=1}^{L} (N_h/\hat{N}_h) \sum_{i \in s_h} (P_h/p_h)\, y_{hi}(d) = \sum_{h=1}^{L} (N_h/\hat{N}_h)\, \hat{Y}_h(d) \tag{4.1}$$

where $\hat{N}_h = (P_h/p_h)n_h$ is the estimated number of population businesses within stratum h. It should be noted that (4.1) may also be written as the calibrated estimator

$$\hat{Y}_{SRAT}(d) = \sum_{h=1}^{L} \sum_{i \in s_h} \tilde{w}_{hi} y_{hi}(d)$$

where $\tilde{w}_{hi} = w_{hi}\, a_{hi}$, with $w_{hi} = P_h/p_h$ and $a_{hi} = N_h/\hat{N}_h$ for the hith sampled rotation group.

The corresponding variance estimator is:

$$v(\hat{Y}_{\mathrm{SRAT}}(d)) = \sum_{h=1}^{L} \frac{P_h^2}{p_h} \left(1 - \frac{p_h}{P_h}\right) \frac{1}{p_h - 1} \sum_{i \in s_h} a_{hi}^2 (\, y_{hi}(d) - N_{hi} \bar{\bar{y}}_h(d))^2 \tag{4.2}$$

where $\bar{\bar{y}}_h(d) = \sum_{i \in s_h} y_{hi}(d) / \sum_{i \in s_h} N_{hi}$ is the mean of businesses within the sampled rotation groups in stratum h that belong to Domain d. The estimator $\hat{Y}_{SRAT}(d)$ is the estimator currently used, and it yields levels of reliability for total sales that are close to the ones specified in the sampling design. Can estimation be improved using the GBI auxiliary data?

4.2. Alternative Procedure

As mentioned earlier, the estimated GBI has been generated for the BR's smaller businesses (those below the maintenance threshold) since 1990. There are several ways to incorporate these auxiliary data in the estimation process for the small take-some strata within each geography–industry combination. Raking estimation was chosen because it preserves auxiliary totals in two dimensions (size and industry) by region. Furthermore, the resulting adjustment is always positive. Raking ratio estimation is used only if the following three conditions are satisfied:

- A sufficient number of businesses (at least five observations) is available at the stratum level.
- A good correlation (based on test provided in Cochran 1977, p. 157) between the x-variable (estimated GBI) and the y-variable of interest (sales) exists.
- The auxiliary data are current.

If these conditions are not met, estimation defaults to the separate ratio estimator (4.1).

The estimated GBI's within the small take-some strata are either below the associated maintenance threshold or above it. The businesses within each take-some stratum U_h eligible for raking can be further split into two size post-strata. As post-stratum membership is decided at the business level, the members of a rotation group will either belong to a single post-stratum or to two post-strata. The number of businesses that belong to the hith rotation group and to the rth post-stratum ($r = 1, 2$) will be denoted as N_{hir}.

Raking is carried out separately within each region (province or metropolitan and nonmetropolitan region within each province) for eligible small take-some strata. The two

dimensions where raking is applied are the trade groups and the two post-strata. The set of strata eligible for raking within a geographical region m are denoted as R_m, where $m = 1, \ldots, 16$. The number of unique trade groups within a geographical region m will be denoted as T_m. The set of strata where raking does not take place are denoted as $\bar{R}_m$ where $m = 1, \ldots, 16$. The set of population businesses that belongs to the strata sets R_m and $\bar{R}_m$ are nonoverlapping calibration groups whose union is U. The set of strata for geographical region m in R_m that belongs to trade group t will be denoted as $R_{m,t}$, where $R_{m,t} \subset R_m$.

An iterative process adjusts the original sampling weights $w_{hi} = P_h/p_h$ within each raking cell (trade group by size post-stratum) until the sums of the weighted GBI's are equal to the marginal totals. Raking cells are obtained by post-stratifying into size groups the units in the strata $h \in R_{m,t}$. The weighted GBI totals within each raking cell (trade group t by size post-stratum r) within R_m are raked until their sum is equal is close to the known GBI marginal totals. Let the auxiliary data associated with each hijth business be denoted as x_{hij}. The marginal population totals for the rth ($r = 1, 2$) size post-stratum and tth ($t = 1, \ldots, T_m$) trade group within geographical region m and set R_m are respectively:

$$\text{(i)}\ X_{m,r,+} = \sum_{t=1}^{T_m} \sum_{h \in R_{m,t}} \sum_{i \in s_h} \sum_{j=1}^{N_{hir}} x_{hij} \quad \text{for the } r\text{th size post-stratum and}$$

$$\text{(ii)}\ X_{m,+,t} = \sum_{r=1}^{2} \sum_{h \in R_{m,t}} \sum_{i \in s_h} \sum_{j=1}^{N_{hir}} x_{hij} \quad \text{for the } t\text{th trade group} \tag{4.3}$$

The starting point for raking is $\hat{X}_{mrt}^{(0)} = \hat{X}_{mrt} = \sum_{h \in R_{m,t}} \sum_{i \in s_h} w_{hi} \sum_{j=1}^{N_{hir}} x_{hij}$. Proceeding as in Deming and Stephan (1940), the raking steps ($c = 1, \ldots, C$) for $r = 1, 2$ and $t = 1, \ldots, T_m$, are:

$$\hat{X}_{mrt}^{(c)} = \begin{cases} \hat{X}_{mrt}^{(c-1)} \dfrac{X_{m,r,+}}{\hat{X}_{m,r,+}^{(c-1)}} & \text{for } p \text{ odd} \\ \hat{X}_{mrt}^{(c-1)} \dfrac{X_{m,+,t}}{\hat{X}_{m,+,t}^{(c-1)}} & \text{for } p \text{ even} \end{cases} \tag{4.4}$$

where $\hat{X}_{m,r,+}^{(c-1)} = \sum_{t=1}^{T_m} \hat{X}_{mrt}^{(c-1)}, \hat{X}_{m,+,t}^{(c-1)} = \sum_{r=1}^{2} \hat{X}_{mrt}^{(c-1)}$. The recursive process is terminated at C, where convergence (or near convergence) is reached. For businesses hij belonging to the set of strata $R_{m,t}$, the resulting adjustment to the sample weight is:

$$a_{hij} = \frac{\hat{X}_{mrt}^{(C)}}{\hat{X}_{mrt}^{(0)}} \tag{4.5}$$

It should be noted that if all of the businesses within a rotation group belong to a single post-stratum, this adjustment will be unique within that rotation group. Otherwise, the adjustment for a business will depend on the size post-stratum ($r = 1, 2$) to which it belongs.

Hence, the adjustment factor is $a_{hij} = N_h/\hat{N}_h$ for businesses within rotation groups that did not participate in the raking (i.e., $h \in \bar{R}_m$), while it is $a_{hij} = \hat{X}_{mrt}^{(C)}/\hat{X}_{mrt}^{(0)}$ for businesses within rotation groups that participated in the raking (i.e., $h \in R_m$).

The estimator of the total for a given domain $U\,(d)$ is

$$\hat{Y}_{MIX}(d) = \sum_{h=1}^{L} \sum_{i \in s_h} \sum_{j=1}^{N_{hi}} w_{hi}\, a_{hij}\, y_{hij}\,(d) \tag{4.6}$$

The building blocks for the estimator of variance for (4.6) are the calibration groups $\bar{R}_m$ and R_m, depending on how the adjustment factor has been computed. That is,

$$v(\hat{Y}_{MIX}(d)) = v\left(\sum_{m=1}^{16} \sum_{h \in \bar{R}_m} \sum_{i \in s_h} \sum_{j=1}^{N_{hi}} w_{hi} a_{hij} y_{hij}(d)\right) + v\left(\sum_{m=1}^{16} \sum_{h \in R_m} \sum_{i \in s_h} \sum_{j=1}^{N_{hi}} w_{hi} a_{hij} y_{hij}(d)\right) \tag{4.7}$$

Since $a_{hij} = a_{hi}$ for $h \in \bar{R}_m$, the first part of (4.7) is estimated as in (4.2). That is

$$v\left(\sum_{m=1}^{16} \sum_{h \in \bar{R}_m} \sum_{i \in s_h}^{16} \sum_{j=1}^{N_{hi}} w_{hl} a_{hij} y_{hij}(d)\right) = \sum_{m=1}^{16} \sum_{h \in \bar{R}_m} \frac{P_h^2}{P_h}\left(1 - \frac{P_h}{P_h}\right) \frac{1}{P_h - 1} \sum_{i \in s_h} a_{hi}^2 (y_{hi}(d) - N_{hi} \bar{\bar{y}}_h(d))^2$$

The second part of the estimated variance in (4.5) is associated with the set of raked strata R_m, $m = 1, \ldots, 16$. Brackstone and Rao (1979) developed expressions for the standard error of the raking ratio estimators of population total up to four iterations that were derived under simple random sampling. However, the associated computations can be greatly simplified by approximating the raking procedure as a two-way analysis of variance model that includes column effects ρ_{mr} (for size post-strata), row effects γ_{mt} (for trade groups), as well as the auxiliary data x_{hij}. This differs from Deville and Särndal (1992) who approximated raking for the case of known marginal counts. The variable of interest $y_{hij}\,(d)$ is modeled as $y_{hij}(d) = (\rho_{mr} + \gamma_{mt}) x_{hij} + \varepsilon_{hij}$, with $E_\xi(\varepsilon_{hij}) = 0$, $V_\xi(\varepsilon_{hij}) = \sigma_t^2$ for all $h \in R_m$.

The raking procedure can also be approximated as a calibration problem with known marginal totals that are the sum of known x continuous variables. Hence, the form of the resulting estimated variance needs to be developed in the context of continuous variables. To this end, for a given unit "*hij*" with $h \in R_m$, let $\boldsymbol{\delta}'_{hij} = \left(\delta_{1+,hij}, \delta_{2+,hij}, \delta_{+1,hij}, \ldots, \delta_{+T_{m,hij}}\right)$ denote a $(2 + T_m)$ dimensional vector of indicator variables $\delta_{r+,hij}$, where $\delta_{r+,hij} = 1$ if the *hij*th business is in post-stratum r and 0 otherwise; also $\delta_{+t,hij} = 1$ if the *hij*th business is in trade-group t and 0 otherwise.

The vector of known totals within R_m is given by $\boldsymbol{X}_m = \sum_{h \in \bar{R}_m} \sum_{i=1}^{P_h} \sum_{j=1}^{N_{hi}} \boldsymbol{x}_{hij}$, $m = 1, \ldots, 16$, with components $\boldsymbol{x}_{hij} = \boldsymbol{\delta}_{hij} x_{hij}$. Note that $\boldsymbol{X}_m$ can also be written as $\left(X_{m,1,+}, X_{m,2,+}, X_{m,+,1}, \cdots X_{m,+,T_m}\right)$ where $X_{m,r,+} (r = 1, 2)$ and $X_{m,+,t} (t = 1, \ldots, T_m)$ are respectively the population post-stratum and trade-group marginal controls given by (4.1). Letting $\boldsymbol{\lambda}'_m = \left(\rho_{m1}, \rho_{m2}, \gamma_{m1}, \ldots, \gamma_{m,T_m}\right)$, we have that $\boldsymbol{\delta}'_{hij} \boldsymbol{\lambda}_m = (\rho_{mr} + \gamma_{mt})$ for $h \in R_m$, whenever the unit *hij* belongs to post-stratum r and trade group t. We seek generalized raking weights $\tilde{w}_{hij} = w_{hi} F(\boldsymbol{\delta}'_{hij} \hat{\boldsymbol{\lambda}}_m)$ with $F(a) = \exp(a)$, where the vector $\hat{\boldsymbol{\lambda}}_m$ is determined by solving the following calibration equation:

$$\boldsymbol{X}_m = \sum_{h \in R_m} \sum_{i \in s_h} \sum_{j=1}^{N_{hi}} w_{hi} F(\boldsymbol{\delta}'_{hij} \boldsymbol{\lambda}_m) \boldsymbol{x}_{hij} \tag{4.8}$$

Noting that $F(\boldsymbol{\delta}'_{hij}\boldsymbol{\lambda}_m)$ depends on the raking cell but not on the label within the cell, calibration equation (4.5) simplifies to:

$$\sum_{t=1}^{T_m} \hat{X}_{mrt} F(\rho_{mr} + \gamma_{mt}) = X_{m,r,+} \quad \text{for } r = 1, 2$$

and

$$\sum_{r=1}^{2} \hat{X}_{mrt} F(\rho_{mr} + \gamma_{mt}) = X_{m,+,t} \quad \text{for } t = 1, \ldots, T_m \tag{4.9}$$

where $\hat{X}_{mrt} = \sum_{h \in R_{m,t}} \sum_{i \in s_h} w_{hi} \sum_{j=1}^{N_{hir}} x_{hij}$ $(m = 1, \ldots, 16)$

Solving (4.9) yields $F(\boldsymbol{\delta}'_{hij}\hat{\boldsymbol{\lambda}}_m) = F(\hat{\rho}_{mr} + \hat{\gamma}_{mt}) = \exp(\hat{\rho}_{mr} + \hat{\gamma}_{mt})$. Hence the calibrated cell estimates are $\hat{X}^{(W)}_{mrt} = \hat{X}_{mrt} F(\hat{\rho}_{mr} + \hat{\gamma}_{mt})$. $\hat{X}^{(W)}_{mrt}$ should be numerically quite close to $\hat{X}^{(C)}_{mrt}$ obtained via raking in (4.2). The calibrated weights can also be expressed as $\tilde{w}_{hij} = w_{hi} a_{hij}$, where $a_{hij} = (\hat{X}^{(W)}_{mrt} / \hat{X}_{mrt})$, and the resulting generalized calibrated weights $\tilde{w}_{hij} = w_{hi} F(\boldsymbol{\delta}'_{hij}\hat{\boldsymbol{\lambda}}_m)$ are always positive. The generalized calibrated weights $\tilde{w}_{hij} = w_{hi} F(\boldsymbol{\delta}'_{hij}\hat{\boldsymbol{\lambda}}_m)$ are a nonlinear extension to the instrumental variable weights given in Estevao and Särndal (2000). In our notation, the instrumental weights given by Estevao and Särndal (2000) would have been $\tilde{w}_{hij} = w_{hi} + \boldsymbol{\delta}'_{hij}\hat{\boldsymbol{\lambda}}_m$. Kott (2004) also discusses such an extension for calibrated weights in his discussion of Demnati and Rao (2004).

As suggested by Kott (2004), a possible variance estimator for this type of weighting is:

$$\sum_{m=1}^{16} \sum_{h \in R_m} \frac{P_h^2}{p_h} \left(1 - \frac{p_h}{P_h}\right) \frac{\sum_{i \in s_h} (\tilde{e}_{hi}(d) - \bar{\tilde{e}}_h(d))^2}{p_h - 1}$$

where $\tilde{e}_{hi}(d) = \sum_{j=1}^{N_{hi}} F(\boldsymbol{\delta}'_{hij}\hat{\boldsymbol{\lambda}}_m)\, e_{hij}(d)$ and $\bar{\tilde{e}}_h(d) = \sum_{i \in s_h} \tilde{e}_{hi}(d)/p_h$, $e_{hij}(d) = (y_{hij}(d) - \boldsymbol{x}'_k \hat{\boldsymbol{B}}_{m,\lambda})$ with

$$\hat{\boldsymbol{B}}_{m,\lambda} = \left(\sum_{h \in R_m} \sum_{i \in s_h} \sum_{j=1}^{N_{hi}} w_{hi}\, F(\boldsymbol{\delta}'_{hij}\hat{\boldsymbol{\lambda}}_m) \boldsymbol{\delta}'_{hij} \boldsymbol{x}_{hij} \right)^{-1} \sum_{h \in R_m} \sum_{i \in s_h} \sum_{j=1}^{N_{hi}} w_{hi}\, F(\boldsymbol{\delta}'_{hij}\hat{\boldsymbol{\lambda}}_m) \boldsymbol{\delta}'_{hij} y_{hij}(d),$$

the estimated regression vector within region m. (In general, $F'(X)$ rather than $F(X)$ is used in constructing $\hat{\boldsymbol{B}}_{m,\lambda}$, but the two are identical in this context.)

An alternative estimator of the variance can be based on the prediction model:

$$y_{hij}(d) = (\rho_{mr} + \gamma_{mt})\, x_{hij} + \varepsilon_{hij}, \quad \text{with } E_\xi(\varepsilon_{hij}) = 0, \quad V_\xi(\varepsilon_{hij}) = \sigma_t^2 \quad \text{for all } h \in R_m$$

The sample residuals obtained from this regression fit are $e_{hij}(d) = y_{hij}(d) - \hat{y}_{hij}(d)$ where $\hat{y}_{hij}(d) = (\hat{\rho}_{mr} + \hat{\gamma}_{mt})\, x_{hij}$.

5. Numerical Results

The numerical results were produced using December 1993, MRTS data. The retail population was made up of 159,072 businesses of which 140,440 had a current GBI.

A sample of 18,436 businesses was drawn, resulting in an overall sampling rate of approximately 11.6%. The small take-some stratum was post-stratified into large and small post-strata to identify businesses eligible for raking. This resulted in 284 raking cells with 4,650 businesses. Convergence was reached after five iterations of the raking algorithm. The resulting calibration adjustments ranged from 0.5 to 3.5. Note that businesses in the take-all and large take-some strata were not subject to raking.

The separate ratio estimate was 19.72 billion dollars at the Canada level. The corresponding estimate using the raking procedure within the small take-some strata resulted in a total estimate of 20.15 billion dollars, representing an increase of 2%. The corresponding standard errors were 175.9 million and 154.3 million dollars, translating into a 16.4% improvement for the raking procedure. Figure 1 shows that similar improvements can be achieved at the provincial level as well.

The estimates of total sales have increased slightly because of the raking weight adjustment.

Figure 2 illustrates the size of the change in billions of dollars; in relative terms, it ranges from 0% to 3.13%.

The positive effect of incorporating auxiliary data in the estimation of total sales is also evident at the cell level (province by trade group). In several cells, the improvement in efficiency due to raking is quite good (Hidiroglou, Choudhry, and Lavallée 1991).

6. Summary

Several estimation procedures methods use auxiliary data. The use of auxiliary data provides more complete information about the population of interest and often reduces the sampling variance of the estimated parameters. This also holds true for subdomains covering only smaller businesses.

The benefits of using auxiliary information are enhanced reliability of the estimates and/or reduced processing costs. The gains can also be translated into sample size reductions. For instance, to maintain the current target coefficients of variation, the take-some sample could be reduced by 4,000 businesses. If the sample size remains unchanged, incorporating auxiliary information into the estimation process results in more efficient estimates of total sales.

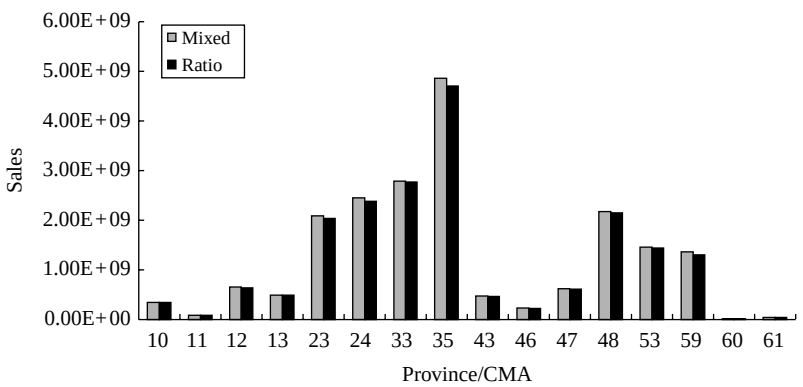

Fig. 1. CV comparison of mixed and ratio estimators at province/CMA level

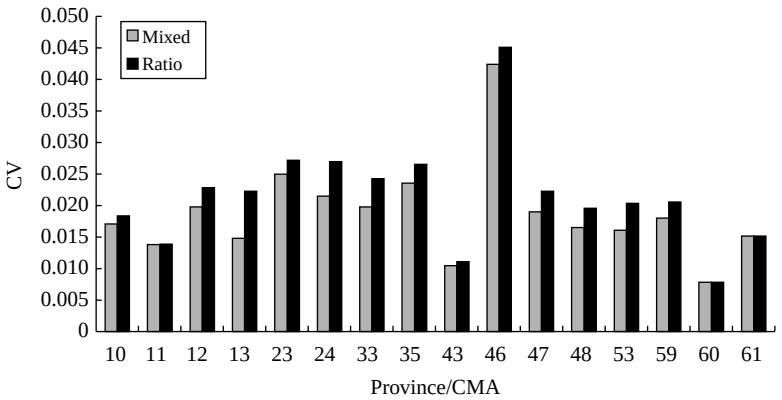

Fig. 2. Comparison of ratio vs. mixed estimated total sales at province/CMA level

7. References

Brackstone, G.J. and Rao, J.N.K. (1979). An Investigation of Raking Ratio Estimators. Sankhya, Series C, 41, 97–114.

Cochran, W.G. (1977). Sampling Techniques (third edition). New York: John Wiley and Sons.

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of Sampled Frequency Table when the Expected Marginal Totals are Known. Annals of Mathematical Statistics, 11, 427–444.

Demnati, A. and Rao J.N.K. (2004). Linearization Variance Estimators for Survey Data. Survey Methodology, 30, 17–26.

Deville, J.C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Estevao, V.A. and Särndal, C.-E. (2000). A Functional Form Approach to Calibration. Journal of Official Statistics, 16, 379–399.

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.

Hidiroglou, M.A. (1986). On the Construction of a Self-Representing Stratum of Large Units in Survey Design. The American Statistician, 40, 27–31.

Hidiroglou, M.A. (1989). Methodology for Monthly Wholesale and Retail Trade Surveys. Methodology Branch Working Paper, BSMD-89-002E/F, Statistics Canada.

Hidiroglou, M.A., Choudhry, H., and Lavallée, P. (1991). A Sampling and Estimation Methodology for Sub-annual Business Surveys. Survey Methodology, 17, 195–210.

Kott, P.S. (2004). Comment on Demnati and Rao. Survey Methodology, 30, 27–28.

Trépanier J., Babyak C., Marchand I., Bissonnette J., and St-Pierre M. (1998). Enhancements to the Canadian Monthly Wholesale and Retail Trade Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, 487–492.

Received February 2003
Revised June 2005