ENHANCEMENTS TO THE CANADIAN MONTHLY WHOLESALE AND RETAIL TRADE SURVEY

Julie Trépanier, Colin Babyak, Isabelle Marchand, Joël Bissonnette and Martin St-Pierre, Statistics Canada
Julie Trépanier, Statistics Canada, Tunney's Pasture, R. H. Coats bldg, Ottawa (Ontario), K1A 0T6 Canada

Key Words: monthly survey, panel design, restratification, death removal, overlap, regression estimation.

1. Introduction

The Monthly Wholesale and Retail Trade Survey (MWRTS) is a major survey conducted by Statistics Canada. The design uses stratified simple random sampling without replacement (STSRSWOR), with stratification by industry, province and size. Estimation uses a form of the Horvitz-Thompson estimator for an STSRSWOR design.

MWRTS was last redesigned in 1988. At that time, Statistics Canada was conducting its important Business Surveys Redesign Project (BSRP) (Colledge and Armstrong 1989). Among others, one objective of the BSRP was that the redesigned Business Register (BR), a list of all businesses operating in Canada, was to be used as the frame for all business surveys. Moreover, all survey activities were to be hooked up, as much as possible, to the BR. This meant that the frame, stratification variables, sample maintenance, creation of collection entities and survey feedback would all be integrated to the BR environment.

MWRTS was the first survey to be hooked up to the BR in 1988. Since then, limited improvements have been made to MWRTS. Changes in industry or geography classification were increasingly dealt with by domain estimation resulting in a less efficient sample. Moreover, changes in units' size have occurred causing some design weights to be out of date and making the estimates less stable. The quality of the stratification variables on the BR has also improved over the past ten years. In order to improve the survey, two projects were undertaken, the first one to restratify MWRTS and the second one to explore regression estimation using auxiliary variables from the frame. Restratifying MWRTS is an exercise that should occur regularly. However, in order to reduce the impact of units that become misclassified between two restratification exercises, we have also initiated the second project to study more robust estimators than the Horvitz-Thompson estimator.

This paper describes the methodology of the restratification taking into account constraints of an on-going survey, presents results from the four month parallel run performed from December 1997 to March 1998 and gives an overview of the study on regression estimators.

2. Overview of MWRTS Design

The BR is a list of all known businesses operating in Canada. Its source of information is administrative data from Revenue Canada. It is divided into two main portions: the Integrated Portion (IP) and the Non Integrated Portion (NIP). In the IP portion, the employer payroll deduction (PD) accounts and the income tax files are linked and integrated, providing a unique and unduplicated list of large and complex businesses. The NIP contains all other businesses with at least $30,000 in revenue and is based on PD accounts only. The statistical structure of a business contains from top to bottom four levels of statistical entities: enterprise, company, establishment and location. MWRTS extracts its frame from the BR. The survey population for the retail component is any statistical company in the IP or the NIP having at least one statistical location coded to the retail trade sector. For wholesale, it is any statistical company with at least one statistical establishment in the wholesale trade sector. Once identified, these statistical companies become the sampling units.

The stratification of the two target populations is by geographic region, industry and size. The geographic breakdown is by province and territories and also for retail by four important Census Metropolitan Areas (CMAs). The industry breakdown is by groups of 1980 Standard Industrial Classification (SIC) codes at the three and four digit levels (called trade groups). Complex statistical companies (operating in more than one province or trade group) are assigned to their dominant trade group and province of activities. Stratification by size is carried out within each geographic region and trade group using a revenue variable present on the BR. Each of these combinations is divided into, at most, three substrata: one take-all (self-representing) for large or complex statistical companies, and up to two take-somes for medium and small companies. The take-all stratum is created because the distribution of revenue is highly asymmetrical. In 1988, the take-all thresholds were calculated using a method by Hidiroglou (1986). If necessary, two take-some strata were defined. The thresholds delimiting the two take-some strata were set

to equal those dividing the IP and the NIP. This was due to the unavailability of size measure for the NIP statistical companies at that time.

For sample allocation, a target CV of 1.2% for retail and 1.7% for wholesale was first specified at the Canada level. The marginal CVs (trade groups, provinces, CMAs) were then obtained. Finally, CVs by trade group X geographic region were calculated via a raking ratio approach. An auxiliary variable X was used as a proxy for the calculations. For retail, the marginal CVs were 3.5% by trade group, 2.5% by province and 3.4% by CMA/rest of the province. For wholesale, the marginal CVs were 4.3% by trade group and 3.4% by provinces. For both retail and wholesale, the CVs by trade groups X geography were below 10%. Sample allocation was proportional to $\sqrt{X_k}$ in the take-some strata h. In addition, expected proportions p_h of live statistical companies for both take-all and take-some strata h were included in the calculations. More details on the procedure is given in Latouche (1988). As of November 1997, the retail sample contained around 15,000 live statistical companies out of a population of 137,000. The wholesale sample contained around 7,000 live units out of a population of 58,000.

When designing MWRTS in 1988, provisions were made to allow the rotation of the take-some sample. Rotation was to be achieved through panel sampling (Hidiroglou and Srinath 1993). This method consists of randomly allocating the entire set of statistical companies in each take-some stratum h to P_h panels of equal size. The first p_h panels are chosen for the initial sample such that p_h/P_h is approximately equal to the desired sampling fraction. The numbers P_h and p_h in each stratum are chosen in accordance with the sampling rate, and the maximum number of occasions that a unit must remain in the sample and the minimum number of occasions it must stay out of the sample (respectively 24 and 12 months in MWRTS). Rotation of the retail sample started in 1993. The first rotation was performed by dropping the first panel from the sample and adding the $(p_h + 1)^{th}$ panel to the sample.

Births are identified every month. They can represent either a new business or an existing business that changed its major activity to the retail or wholesale industries. 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.

Deaths also occur on a monthly basis. They are identified via the administrative sources updating the BR and survey feedback, including MWRTS'. They are coded as deaths on the MWRTS frame so that no questionnaire is sent. Their sales value is imputed to 0. Consequently, they do not contribute to the point estimates but do impact on the estimated variance. Once or twice a year, deaths are removed from the frame and sample in an unbiased manner (see 3.1).

Since p_h panels were selected among P_h for a given stratum h, the design weight is thus equal to P_h/p_h. By using a post-stratified estimator where the post-strata equal the strata so that the sum of the weights on the achieved sample size n_h equal the population size N_h, we obtain:

$$\hat{Y}(d) = \sum_h \frac{N_h}{\hat{N}_h} \frac{P_h}{p_h} \sum_{i=1}^{n_h} y_{hi}(d) = \sum_h \frac{N_h}{\frac{p_h}{p_h} n_h} \frac{P_h}{p_h} \sum_{i=1}^{n_h} y_{hi}(d) = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}(d) \quad (1)$$

where d is the domain of interest and $y_{hi}(d)=y_{hi}$ if company i is in d and $y_{hi}(d)=0$ if not. Note that a stratum h and a domain d may not always coincide since the domain information is more up to date than the stratification information. More general information on MWRTS can be found in Hidiroglou (1989).

3. Restratification Project

Over the past ten years, improvements have been made to MWRTS. In 1992 for example, the retail sample was increased by approximately 2,000 companies (Duggan 1992a, Hoyt and Duggan 1992). At the same time, some take-some companies were promoted to the take-all portion and vice-versa. This helped achieve better coefficients of variation (CVs). As well, in 1994, the stratification of two heterogeneous wholesale trade groups was each broken down into two more homogeneous trade groups. There have also been on occasion some removals of dead units from the frame and sample (Duggan 1992b). Some units were also promoted to the take-all portion or demoted to the take-some portion but the population was never completely restratified however. Some businesses were in the same major industry and geography but more notably, many statistical companies have grown since 1988 and were no longer stratified to the proper size strata. The major problem was with large companies that were stratified in the small take-some strata in 1988 and had large design weights. They caused variance estimates in some particular domains to increase over time and made estimates unstable.

The MWRTS restratification project was initiated in 1997. The methodology included four main parts:

death removal; restratification by trade group, geography and size; review of sampling fractions; and selection of a new sample while maximising the overlap with the current sample. The restratification had to be accomplished under certain constraints: the total sample size had to remain about the same; the number of new statistical companies in the sample was limited to 5,000 for both retail and wholesale; CVs had to be less than or equal to the 1988 target CVs; and the restratification had to be performed without slowing down the regular production. The first two constraints were driven by the collection costs. Contacting a unit the first time is extremely costly compared to contacting other units (e.g., mailing of an introductory package, contact information not up to date). Because we knew it would be difficult to collect data for the new portion of the sample for the first few months and also because we could not alter the regular production and the scheduled data releases, it was decided that the restratified survey would be tested in parallel with the regular production for four months. In summary, two surveys, the old and the restratified, would be executed entirely in parallel from sampling to estimation including data collection for the December 1997 to March 1998 reference months. Official transition to the new restratified survey was to happen only in the April 1998 reference month. There were two additional advantages to the parallel run test. First, we would ensure that all production systems would "accept" the restratified survey. Second, we were expecting the level estimates to be more accurate but also different from the old survey, creating an artificial break in the time series. Linkage procedures needed to be put in place. This parallel run test would give time to analyse the data and get prepared for the transition in April 1998 when the old version would be dropped. A four month parallel run test is however short. Preliminary estimates are available two months after the reference month. By the time we see the preliminary February estimates, the April transition would have already occurred. Finally, to simplify the process, the rotation of the retail sample was stopped in November 1997.

3.1. Death Removal

There are two types of deaths. It can be a statistical company which has ceased its activities (called "out-of-business") or whose major activities are not retail or wholesale trade anymore (called "out-of-scope"). Deaths are identified via surveys and BR administrative sources. The surveys are much faster sources to identify deaths. With administrative sources there can be a time lag of a year. Deaths are first coded as such on the BR and then on the MWRTS frame. They are not automatically removed from the MWRTS frame and sample since the identification of deaths on the MWRTS frame is not independent from MWRTS (the collection for MWRTS serves in updating the BR, which in turn updates the MWRTS frame). We presently have no way to distinguish deaths from the administrative sources (which could be deemed independent) and deaths from our own survey. As mentioned earlier, the MWRTS sample is partially retained month after month. If we were to remove all deaths from the frame and sample, we would remove proportionally more units from the sample than from the out-of-sample portion. The resulting sample could be highly biased since it would show a very "live" picture, which would not represent the out of sample portion where a large number of deaths are unknown. The same situation holds with the new sample, which is not independent from the previous one since maximisation of overlap is performed. Consequently, a complete death removal is never performed in MWRTS. Rather, the following procedure is applied.

First, dead statistical companies belonging to take-all strata are all removed since their design weights are 1 and they are self-representing. For dead statistical companies in take-some strata, the proportions of deaths in sample and out of sample are first computed for each stratum, i.e., n_{hd}/n_h and $(N_{hd} - n_{hd})/ (N_h - n_h)$ where n_{hd} and N_{hd} are respectively the number of deaths in sample and the total number of deaths in the population of stratum h. For reasons mentioned in the previous paragraph, the proportion is usually higher in the in-sample portion of the strata. When this occurs, all dead statistical companies belonging to the out-of-sample portion are removed from the frame. The out-of-sample proportion of deaths $(N_{hd} - n_{hd})/ (N_h - n_h)$ is used to remove in-sample dead companies. This proportion is multiplied by n_h and rounded to get the expected number of deaths that should be removed from the in-sample portion. The removal is done by systematic sampling. Some deaths are left in the sample and will represent deaths in the out-of-sample portion that are not yet known. In rare strata where the out-of-sample proportion of deaths is higher than the in-sample one, all in-sample and out-of-sample deaths are removed. The hypothesis that the identification of deaths is dependent on the survey and that consequently more deaths are known in the sample does not seem to hold here, justifying the strategy. This is, however, different from the way the death removal is handled in normal production where, as above, some out-of-sample deaths are not removed. When such a death removal is performed in normal production, this has the advantage of keeping the weights constant, which is important when one is interested in keeping to a minimum the impact on the series. As a result of the death removal performed as part of the restratification

project, around 18% of units of both the retail and wholesale frames were dropped.

3.2. Restratification

The stratification by geographic regions and trade groups was first updated using the most recent industrial and geographic information present on the BR. Complex statistical companies were assigned to their most current dominant geographic region and trade group. Deaths left on the frame by the death removal described in 3.1 were stratified to trade groups and geographic regions according to the last information available for them on the BR. The updates made to the BR regarding the industry and geographic classification which serve for restratification could be deemed dependent on the sample due to the survey feedback. Changes in industry and geography stratification from 1988 to 1997 were quantified and no major differences were observed between the in-sample and out-of-sample portions. In fact, less than 2% of the companies had changed industry classifications and less than 4% had changed geography classifications and the changes were evenly spread between in-sample and out-of-sample units. Larger differences between the out-of-sample and in-sample portions could have been observed if there had been a smaller time lag between the two stratification periods. Finally, due to operational constraints with the MWRTS systems, the stratification could not be changed (i.e. based on different groupings, such as NAICS) although this would have been a desirable enhancement.

For stratification by size, the method from Lavallée and Hidiroglou (1988) was first tested to compute new size thresholds since it computes optimal thresholds in the case of a take-all stratum and a certain number of take-some strata. For MWRTS, the Lavallée and Hidiroglou method gave quite different thresholds from the 1988 ones. This had the undesirable impact of making the maximisation of overlap difficult. The number of new companies in the sample went beyond the maximum allowed. Time being limited, it was decided with the subject matter economists that the 1988 thresholds that were not anymore relevant in 1997 would simply be increased by a rate approximately equal to the retail and wholesale economy growth over the past nine years. Using these updated thresholds, statistical companies were assigned to their new size stratum using their most recent revenue value present on the BR. Since the take-all thresholds were increased, the number of take-alls went from 6,500 to 5,500 companies for retail. For wholesale, the number of take-all remained around 4,000 due to the increase in the number of complex statistical companies. The revenue variable present on the BR is obtained or derived via the administrative sources and is, in that sense, independent from MWRTS. Complex statistical companies were again assigned to the take-all strata. Around 11% of the retail companies and 13% of wholesale companies changed size strata.

3.3. Review of Sampling Fractions

The plan was to increase the sampling fractions in areas where the target CVs (see section 2) were no longer satisfied by the existing sample. On the other hand, we did not decrease any of the existing sampling fractions in the take-some strata; we did not want to introduce problems in areas where there were none. We first started with the existing sampling fractions and computed the total sample size. In retail, it happened to be lower than the current total sample size due to the decrease in take-all companies. In trade group and geography combinations where there was room to increase the total sample size, a sample allocation proportional to $\sqrt{\text{revenue}}$ from the BR was performed. This enabled us to achieve the desired total sample size. For wholesale, applying the existing sampling fractions already gave the desired total sample size. Once the sample selection was completed (see 3.4), we ensured using the revenue variable from the BR that the new sample achieved the target CVs. For retail, this led us to an increase of some sample sizes. The final number of live companies in sample was 15,100 and 7,100 for retail and wholesale respectively.

3.4. Selection of a New Sample (Maximisation of Overlap)

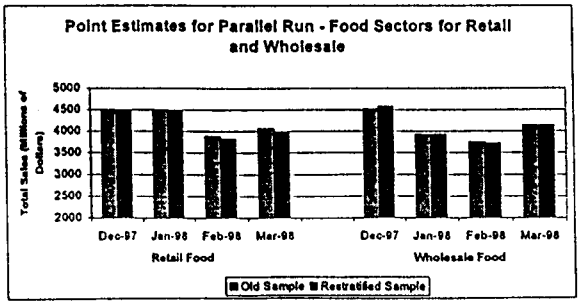
The new sample had to overlap as much as possible with the old sample. This was done using the Kish and Scott method (1971) adapted for panel design (Hidiroglou, Choudhry and Lavallée 1991). The procedure is based on the fact that each panel is a simple random sample of the population units and it is summarised as follows.

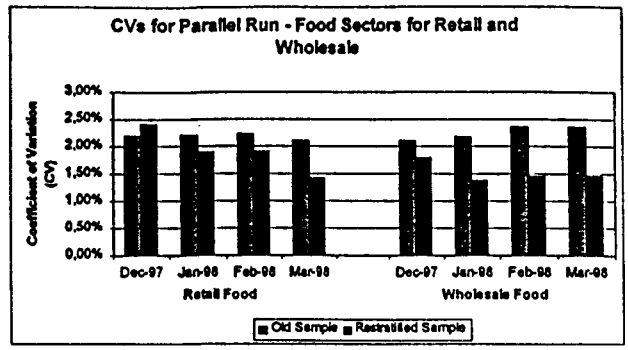
Let h represent the old stratum, h' the new stratum, p_h and P_h the number of panels in sample and the total number of panels in stratum h. In addition, let I_h be the current sampling interval for the panels before restratification. I_h can be of two types: 1) $I_h=[i,i+p_h]$ meaning that panel i to $i+p_h$ are in sample with $i+p_h \leq P_h$; 2) $I_h=[i,P_h]U[1,p_h-(P_h-i+1)]$ with $i+p_h>P_h$. For the second type, let us assume p_h, P_h and i are 5, 10 and 8. This means panels 8, 9, 10, 1 and 2 are in sample. The moving of the sampling interval occurs because of the rotation. Since the wholesale sample has never had any rotation, the sampling interval has always been $[1,p_h]$. The first step is to rewind the sampling interval to the interval $[1,p_h]$ in all retail strata h. This can be done

using a modulo operation based on the number of rotations that have occurred and the total number of panels P_h. Then, companies within each panel are sorted randomly. Ranked values r_{hi} that respect the order of the panels and the random order are assigned to the companies within each old stratum h. Each old stratum h is then divided into smaller sets of companies that belong to the same new stratum h'. Let us call $U_{hh'}$ those sets of companies belonging to the old stratum h and new stratum h'. Combine the sets $U_{hh'}$ that refer to the same new stratum h'. According to the new sampling fractions obtained for strata h', determine how many panels $p_{h'}$ must be selected in sample. Given the complexity of the 1988 MWRTS sampling system and time constraints, the total number of panels within a given stratum was not changed. Over a given new stratum h', the companies are sorted based on the rank variable r_{hi} previously defined. Note that units part of old panels #1 are still first, units in old panels #2 are second etc. Within each stratum h', companies are assigned to new panel numbers by observing which interval they belong to, thus completing the selection procedure. Estimated CVs were then calculated based on the revenue variable from the BR to ensure that target CVs seemed to be achieved. Estimates of that revenue based on the new sample were also produced and compared to the true revenue based on all population units to ensure that no systematic bias was present. Finally, the total number of new units in both retail and wholesale samples was 4,865, meeting the maximum of 5,000 fixed as a constraint at the beginning of the restratification project.

3.5. Results

From December 1997 to March 1998 reference months, data for both the new sample and the old sample were collected and all steps of the survey were performed in parallel. This allowed us to compare point estimates and estimated CVs from the two sources. At very aggregated levels, the level of the point estimates was not much affected. However, a significant decrease in the CVs was observed. This was appreciated by the users. Below is an example for two trade groups, the food sectors for both retail and wholesale across Canada. It represents what was observed in most trade groups and geographic regions.



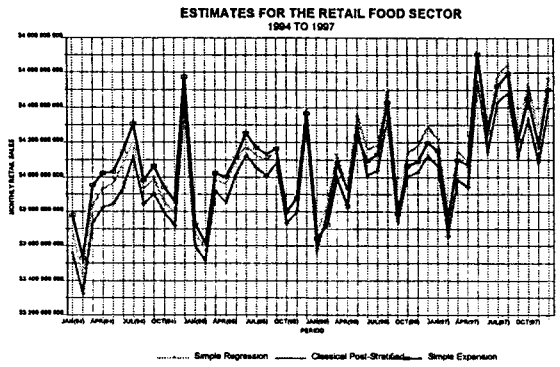


4. Regression Estimators

The restratification project will improve the series for the future. This new series of estimates needed to be linked to the old series. We knew however that the quality of some past estimates in the old series was not very good due to the deterioration of the stratification (large units with large weights). As mentioned earlier, a strict expansion estimator was used. A solution to improve the quality of the past estimates was to recompute them using a regression type estimator. If this estimator had been successful, we could have also used it to produce estimates in the future.

We studied two cases of the regression estimator: the classical post-stratified (uses counts only) and the simple regression estimators (uses counts and an auxiliary variable X). The study was performed on every month from 1994 to 1997 for the retail component of MWRTS. The two estimators were not applied to take-all companies since these units did not cause any problem (their weights were 1). A file created monthly by the BR and representing the most current retail universe was used to define post-strata (also called model groups or calibration groups) and to obtain the auxiliary information, i.e., counts and revenue (X). The post-strata were defined as trade groups X geographic region X size.

The classical post-stratified and simple regression estimators were applied in post-strata where there were at least 15 companies. If possible, some geographic regions and size categories were aggregated. For the simple regression estimator, we also ensure that the correlation between the revenue variable (X) and sales were greater than 0.5 times the ratio of CV(revenue)/CV(sales). If the size criteria or the correlation criteria for the simple regression estimator was not met, the simple expansion estimator was used. Computations were performed using the Generalized Estimation System of Statistics Canada. More details on that study are available in Bissonnette et al. (1998). Results for the trade group representing the retail food sector across Canada are presented next.



The graph above demonstrates the difficulty in analysing the performance of the two new estimators. So we looked at some sectors that were problematic in the past. For example, the expansion estimator showed a questionable decrease of 5.0% in sales from 1995 to 1996 for the food sector in Québec. The classical post-stratified and simple regression estimators showed respectively more acceptable decreases of 0.8% and 0.9%. Unfortunately, some non-problematic sectors were also affected by the two new estimators, making their use difficult to justify. Neither new estimators were used to review the 1994 to 1997 estimates.

4. Conclusion

Restratification provided a significant reduction in the CVs and solved some of the problems noticed with the old sample. This was, however, a very complex process to perform. All systems were put in parallel so that the current survey would be executed with no impact from the restratified survey. Problems encountered were solved during the parallel run, minimising problems when the restratified survey was brought in production in April 1998. This project should be repeated more regularly.

As far as regression estimation is concerned, we would like to spend more time analysing the results and fine-tuning the estimators such as using as auxiliary variables the Goods and Services Tax (GST) data that have been made available recently. We are also looking at ways other ways of dealing with recurring misclassified units other than restratification and calibration.

Bibliography

Bissonnette, J., I. Marchand, M. St-Pierre and J. Trépanier (1998), "Amélioration de la série d'estimations mensuelles des ventes au détail au moyen d'un estimateur par calage", 1998 *Proceedings of the Survey Methods Section,* Statistical Society of Canada, to be published.

Colledge, M., and G. Armstrong (1988), "Statistical Units, Births and Deaths at Statistics Canada after the Business Survey Redesign", *Third International Roundtable on Business Survey Frames*, Auckland, New Zealand.

Duggan, J. (1992a), "Documentation on the Reselection Process", Internal Memorandum, Statistics Canada.

Duggan, J. (1992b), "Current Death Removal Methodology for MWRTS", Internal Memorandum, Statistics Canada.

Hidiroglou, M. A. (1986), "The Construction of a Self-Representing Stratum of Large Units in Survey Design", *The American Statistician*, 40, pp. 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., G. H. Choudhry and P. Lavallée (1991), "A Sampling and Estimation Methodology for Sub-Annual Business Surveys", *Survey Methodology*, 17, 195-210.

Hidiroglou, M. A., and K. P. Srinath (1993), "Problems Associated with Designing Subannual Business Surveys", *Journal of Business & Economic Statistics*, 11, 397-405.

Hoyt, P., and J. Duggan (1992), "Recomputing Sample Sizes for an Ongoing Survey", *1992 Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 203-208.

Kish, L., and A. Scott (1971), "Retaining Units after Changing Strata and Probabilities", *Journal of the American Statistical Association*, 66, 461-470.

Latouche, M. (1988), "Sample Size Determination, Allocation and Selection", Methodology Branch Working Paper, BSMD-88-021E/F, Statistics Canada.

Lavallée, P., and M. A. Hidiroglou (1988), "On the Stratification of Skewed Populations", *Survey Methodology*, 14, 33-43.