THE CANADIAN RETAIL COMMODITY SURVEY

Marie Brodeur, Hélène Bérard and Martin St-Pierre, Statistics Canada

Marie Brodeur, Statistics Canada, Tunney's Pasture, R.H. Coats Bldg, Ottawa (Ontario), K1A OT6, Canada

Key Words: Two-Phase Design; Sales; Multivariate Allocation; Editing; Double Expansion Estimator.

1. Introduction

The Canadian Retail Commodity Survey (RCS) was launched by Statistics Canada in January 1997 to collect detailed information about retail commodity sales in Canada. Prior to the introduction of the RCS, the lack of detailed information on commodity was identified as a data gap by Statistics Canada. In particular, this information was required by the Systems of National Accounts. Since total retail sales are already being measured by the Monthly Retail Trade Survey (MRTS), a two-phase sample design methodology , where the first phase is the MRTS, was developed. The sample size of the RCS is composed of 10,000 companies in Canada. A total of 117 commodities which are assembled into nine major commodity groups such as food, clothing and accessories, and furniture and appliances, etc. are collected each month. The estimates are produced on a quarterly basis.

The two-phase design allowed RCS to take full advantage of the MRTS's information, timeliness and infrastructure. As a result, RCS was developed more quickly, more economically and, from a statistical perspective, more efficiently. For instance, the first-phase information was used to maximise efficiency in a number of areas, including the selection of the same respondents, the use of auxiliary information in sample allocation, edit, imputation and estimation, and the use of existing systems and staff. Similar surveys were conducted in 1974 and 1989, but in addition to being sporadic, they did not produce all the desired results. Nevertheless, the results of the 1989 survey helped in the development of the RCS's sampling plan.

The RCS design is also very innovative because the first-phase sample was completely restratified to make the second-phase sample design as efficient as possible. The methodology of the RCS is presented in greater detail in the following sections. More details are available in Brodeur *et al.* (1997). In order to better understand the methodology behind the RCS, a short description of the MRTS is provided.

2. Overview of the MRTS: First phase of the RCS

The Monthly Retail Trade Survey essentially measures retail sales by trade group (groups based on the three-or four-digit 1980 Standard Industrial Classification (SIC)), province and selected Census metropolitan areas (CMAs). It was last redesigned in 1988. The sample is selected from Statistics Canada's Business Register (BR). The target population consists of statistical companies with statistical locations identified on the BR as retailers. Some 20,000 companies are in sample each month.

The population is stratified by province, territory, selected CMA and trade group. Each combination of trade group and geographic area forms a stratum. Each stratum is divided into three substrata by size: one take-all stratum and two take-some strata, one composed of medium-sized firms and the other of small companies. The take-all strata include all companies with a complex structure, *i.e.*, companies that operate in more than one trade group or geographic area, as well as companies with a gross business income (GBI) above a certain limit. Other companies are assigned to one of the two take-some strata on the basis of their GBI. Sample allocation for the take-some strata is by the square root of GBI. The target coefficient of variation of sales is 1.2% at the national level, 2.5% at the provincial level and 3.5% at the trade group level.

The sample is partially rotated each month to lighten the response burden and keep the response rate high. The population within each take-some stratum is randomly divided into equal-size clusters or panels. The number of panels is determined by the sampling fraction computed at the time of allocation and by the number of months a unit must remain in the sample and outside the sample. A subset of the panels is selected for the sample. Monthly rotation involves removing systematically one panel of each take-some stratum from the sample and replacing it with a new panel. Each month, births are systematically added to the panels. More details are available in Trépanier *et al.* (1998).

3. Design of the RCS

3.1 Second- phase sampling plan

In the second phase of sampling, information from the first phase is used to restratify and allocate the second-phase sample. It is important to note that the second-phase sample is a subset of the first-phase sample. A unit that belongs to the second-phase sample must be part of the first-phase sample. This section deals primarily with stratification, the sample allocation method, and rotation of the RCS sample.

3.1.1. Stratification of the RCS

The frame for the RCS sample is the set of companies in the MRTS sample. As in the first phase, the sampling unit in the second phase is the statistical company. Using the latest information from the MRTS, the first-phase sample was restratified by trade group, province and company size. For the purposes of stratification, each company is assigned a dominant province and a trade group on the basis of its sales volume. This restratification was done only once, at the design stage of the RCS.

MRTS sales were used in determining the company size substrata. However for operational reasons, that variable was not available for the entire sample. Therefore, sales had to be modelled using GBI, which was available for all the units in the population. The model's parameter estimates were used to predict sales for companies whose sales were unavailable in the first phase sample.

3.1.2 Sample allocation

The RCS is intended to provide sales estimates for many commodity groups. Consequently, the sample allocation has to be multivariate. Since there is no conventional solution to the problem of optimal multivariate allocation when the survey is using a stratified two-phase sampling design, an existing method had to be adapted to suit our needs. The method chosen in this case is a modification of the Bethel's algorithm (1989). For more details, see Jocelyn and Brodeur (1996). The CVs from the 1989 Retail Commodity Survey were used in applying Bethel's algorithm. The final sample size produced by the algorithm was about 10,000 companies for a CV of sales of 7% at the Canada level for each major commodity group.

3.1.3. Sample selection and rotation

As mentioned earlier, the MRTS sample is made up of a subset of panels. The sample is partially rotated every month in the take-some strata by replacing one of the panels. For the initial selection of the RCS sample, we ignored the panel structure of the first-phase sample; this approach streamlined the process considerably. Since each first-phase panel can be regarded as a simple random sample of the MRTS sample, the set of all first-phase panels is also a simple random sample. Consequently, if we wish to draw a simple random subsample from that sample, we can do so without regard for the panel structure. We took this approach for RCS.

To take the MRTS rotation and sample updates into account, every month we select a subsample of units from the new MRTS panels as well as from the new births. This procedure maximises the overlap from month to month in the RCS sample. It also ensures that the RCS sample reflects any changes made to the MRTS sample. Since we assumed in our estimation process that simple random sampling is used, we examined the effect that deviating slightly from that assumption would have on the estimates. The results showed that the effect was virtually non-existent.

3.2. Data collection

MRTS data are mostly collected by telephone. The RCS's unit of collection is the same as the MRTS's. For these reasons, it was advantageous to combine data collection for the two surveys. Although the surveys have different questionnaires, collection and follow-up are done for both surveys through one telephone call. A further justification for this approach is the fact that the RCS can be regarded as a supplement to the MRTS. The latter gathers total monthly retail sales, while the RCS asks respondents for a breakdown of sales by commodity groups. Since we were able to use the MRTS's infrastructure to collect RCS data, development of the RCS's collection system focussed on development of the questionnaire, data capture system, edit rules specific to the RCS, and data transmission.

The RCS questionnaire lists over 117 commodity groups, These commodities are regrouped into 27 combinations of parts and totals. For example, footwear is one of the major groups on the questionnaire. It is divided into two parts: athletic and non-athletic footwear. Non-athletic footwear is further subdivided into three parts: women's, men's and children's footwear. The respondent is asked to report its sales for all these levels. The total sales for all the major groups equals the total retail sales from MRTS. The respondent can report its sales by commodity group as a dollar amount or as a percentage of its total

retail sales. If the respondent is unable to provide the data in either form, the interviewer will attempt to at least find out what types of commodities the respondent sells. This information is used at the edit and imputation stage to determine what fields need to be imputed.

Since the majority of companies are in the survey sample month after month, we try to tailor each company's questionnaire to the commodities reported in previous responses. The first time a unit is contacted for the RCS, the interviewer creates a profile containing a list of the commodities usually sold by the unit. The profile is used initially in preparing the tailored questionnaire and later in edit and imputation. It is updated regularly. The tailored questionnaire eases the response burden and helps boost the response rate. The respondent also has the choice to report its data if he wishes on a monthly, quarterly, semi-annual or annual basis. Respondents can report annually when the distribution of their sales does not vary through the year. For the other months that they are not reporting, their data are imputed. Monthly respondents represent 61% of the sample, annual respondents represent 33 % of the sample, and other types of respondents represent the remaining 6 % of the sample.

3.3. Editing and imputation

It was complex to develop an editing and imputation strategy for the RCS. The strategy chosen had to consider the many totals and subtotals as well as the many different commodity groups. The list of commodities varies from company to company even in a given trade group. It was also difficult to borrow results from similar surveys since there are almost no commodity surveys in Canada and abroad. The methods proposed are simple, robust and flexible. The editing and imputation system consists of three main modules: pre-editing, automated editing, and imputation. A customised system was developed for this survey.

3.3.1 Pre-editing

The purpose of the pre-editing module is to perform a series of verifications on the data supplied by the units that contribute the most to the estimate of total sales in each retail trade sector. Those units may be either large companies or small companies that have a high sampling weight. The data they provide are verified to ensure that sums of parts and totals add up, that reported commodities match the type of business, and that there are no sudden changes in sales from month to month or year to year. Data that fail pre-editing are examined by subject-matter experts, who either correct the most obvious errors or contact the respondent for clarification.

3.3.2 Automated editing

All data, even those which have undergone pre-editing, must go through the automated editing stage. The object of automated editing is to identify fields requiring imputation, while altering the data reported by respondents as little as possible. Automated editing finds erroneous data that must be replaced with imputed values when there is inconsistency between the parts and the totals. When the automated editing was developed, other type of rules, such as relationship edits, could not be defined due to the lack of historical information. Verifying sums of parts and totals may appear simple, but it is actually very complex because subtotals are added together to form other totals. For example, adjusting one of the subtotals to resolve an inconsistency with the sum of its parts may induce an inconsistency when considering the sum of the subtotals with the grand total. For total nonresponse, the automated editing system determines which of the fields involved should be zeroed and which should be imputed. The profile created during data collection, historical data (for the previous month or the same month of the previous year) and even the unit's industrial classification are used in this process.

3.3.1 Imputation

Prior to the actual imputation, other verifications are performed on the records that might be used to calculate values imputed to other records. The verifications ensure that outliers will not be employed in those calculations. To that end we apply rules of the same type as those used in pre-editing, though the rejection and acceptance criteria may be different. For example, a women's clothing store that also sells furniture will be considered as an outlier. If this unit was used to calculate imputed values, it might generate furniture sales for all non-respondent women's clothing stores.

The first step in the imputation process involves defining imputation groups. An imputation group consists of a set of homogeneous units. A value imputed to a unit will usually be derived from the values of respondents belonging to the same imputation group. In other words, we want to use units with similar profiles in the imputation process. The RCS's imputation groups are defined on the basis of the latest information about industrial classification, geographic area and unit's size. Seven imputation groupings were defined, each successive imputation grouping is defined at a less detailed level. When the imputation is

not feasible at a more detailed level, for example if there is not a sufficient number of respondents to calculate trends, the imputation is performed at the next level.

Ratio imputation and adjusted historical imputation are the methods currently used in the RCS. Please refer to Bérard *et al.* (1999) for more details. Although edit and imputation are applied to the dollar values of commodity sales, the survey is much more concerned with the distribution of commodities, *i.e.,* the proportion of sales of each commodity in relation to total retail sales. Consequently, ratio imputation methods were preferred for the RCS rather than adjusted historical methods, since the latter methods tend to conceal changes in the distribution. But this strategy is now under revision. In addition, wherever possible, imputation is performed within the imputation groups defined earlier. Only one commodity is imputed at a time. Finally, since imputation does not ensure that the parts will add up to the totals, it is followed by a prorating step.

3.4 Estimation

The goal of RCS is to produce estimates for the distribution of the total retail sales among various commodities. MRTS is the source for the level of the sales (total sales). We chose an estimator that would enable us to use the total sales information from the first phase while maintaining a degree of simplicity in the estimation of variance without sacrificing precision. We had to explicitly develop a variance estimation formula for a two-phase design in which the second-phase sample is selected from a restratified first-phase sample. Two estimators were studied: the Double-expansion estimator and the Reweighted expansion estimator of Kott and Stukel (1997) in the context of stratified sampling at both phases. We also considered a combined ratio version for each of the estimators. We use the linearization method (Taylor type arguments) presented in Binder (1996), to obtain their estimated variances.

The performance of the different variance estimators was examined using a Monte-Carlo simulation study. Actual MTRS data was used to create the simulation population. Various statistics were computed from the simulation. The properties of the estimators were studied conditionally and unconditionally. The relative bias was close to zero for all the estimators. The mean square error (MSE) of each estimator divided by the MSE of the full first-phase estimator were similar amongst the estimators. Generally speaking, the unconditional results for all the estimators were fairly similar. No single estimator stands out.

When comparing the conditional relative bias for the variance estimates and the point estimates, we noticed that the variance estimators of the two ratio estimators were closer and more stable around zero. Other results suggested that the ratio estimators were doing better than the non ratio estimators. These findings support and show the improvements available through the use of auxiliary information for the RCS.

The Double-expansion ratio estimator and the Reweighted ratio expansion estimator have similar results. However, the Double-expansion ratio estimator was selected because the variance estimator was very simple. Further details about simulation results and the different estimators and variance formulas are available in Binder *et al.* (1999) and (1997).

Note that the data are being collected monthly, and the results are published quarterly. Because the samples are not independent from month to month, the covariance was also calculated. This covariance component was added to the total variance estimator of the quarterly estimates.

4. Survey Results

The survey results were first released in December 1998. The RCS brought a new perspective to the retail data in Canada. For instance, Canadians's retail spendings were release by type of commodities. Furthermore, the RCS allows analysis of the market shares of various types of retail stores with respect to certain commodities. The data show the types of retail outlets where consumers prefer to buy these commodities, and shifts in consumer preferences. More details are available in the The Daily [1]. The survey results were above expectations and either achieved or even surpassed the desired precision in some trade groups. Table 1 shows the coefficients of variation by the major commodity groups. For certain major commodities like food and clothing and accessories, the coefficient of variation was as low as 2% even if the design CV was around 7%, at the Canada level. It shows that major gains were achieved with the restratification of the first-phase sample with more recent information, and the use of the multivariate allocation with the sales.

[1] *The Daily, official release of statistical data and publications produced by Statistics Canada. The electronic version of The Daily can be find at http://www.statcan.ca .*

Table 1. Coefficients of variation by major commodity group
Second quarter of 1998

A. Major commodity group	CV
Food (excl. pet food & meals & lunches)	2.2%
Drugs (prescription & OTC), vitamins & other health supplements	3.8%
Personal care products/Health & beauty aids (non-electric)	2.2%
Clothing & accessories	1.8%
Footwear	3.7%
Furniture (indoor), household appliances & electronics (incl. Cameras)	4.6%
Hardware & home renovation products	8.6%
Used automotive vehicles	2.6%
Automotive fuels, oils & additives	4.4%

5. Future Developments

The different systems designed for the RCS are performing very well. Some minor modifications were done to the edit and imputation system to enhance the quality of the data throughout the production cycle. Now that we have two years of data available, we have suggested improvements to better impute commodity data. We have started a review of the edit and imputation strategy. In particular, we are considering including the weights in the ratios used for imputing. With weighted ratios, inferences remain valid both under the design and model-based framework. We are also considering a different approach for general trend ratio imputation. In the current method, the ratio adjustment corresponds to the sum of the sales, for a given commodity, over the sum of all the sales. This ratio has a tendency to reflect more the sales distribution of large companies. An alternative method would be to calculate the distribution of the sales for each company, for a given commodity and then to calculate the average distribution. In this manner, the ratio adjustment represents an average sales distribution that is independent of the magnitude of the sales. One of the greatest challenges remains to define better imputation groups. For some imputation groups, the validity of the model used was found to be questionable. We are also studying the feasibility of using nearest neighbour imputation when other imputation methods fail.

6. Acknowledgements

The authors would like to thank Janet Sear, Julie Trépanier and Elaine Wilson for reviewing this paper and Nick Budko for his assistance.

7. References

Bethel, J.W. (1989), "Sample Allocation in Multivariate Surveys". Survey Methodology, Vol. 15, No 1, 47-57.

Bérard, H., M. Brodeur, M. St-Pierre (1999), "Retail Commodity Survey". Proceedings of the Survey Methods Section, Statistical Society of Canada, to be published.

Binder, D.A. ,C. Babyak, M. Brodeur, M. Hidiroglou, W. Jocelyn (1999), "Variance Estimation for Two-Phase Stratified Sampling". Canadian Journal of Statistics, to be published.

Binder, D.A. , C. Babyak, M. Brodeur, M. Hidiroglou, W. Jocelyn (1997), "Variance Estimation for Two-Phase Stratified Sampling". Proceedings of the Section on Survey Research Methods, American Statistical Association, 267-272.

Binder, D.A. (1996), "Frequency Valid Multiple Imputation for Surveys with a Complex Design". Proceedings of the Section on Survey Research Methods, American Statistical Association, Vol. 1, 281-286.

Brodeur, M., W. Jocelyn, J. Trépanier (1997), "A solution to the design and implementation of a fast-track survey: Two-Phase Sampling. Proceedings of Symposium 97, Statistics Canada, 31-35.

Jocelyn, W., M. Brodeur (1996), " Méthodes de répartition multivariées pour l'échantillonnage à deux phases: Application à l'enquête trimestrielle sur les marchandises''. Recueil des communications des XXVIIIe Journées de Statistiques de l'ASU, 433-436.

Jocelyn, W., M. Brodeur and C. Babyak (1997), "Comparaisons de différents estimateurs de variance à deux phases: étude Monte-Carlo basée sur l'Enquête des marchandises au détail. . Proceedings of the Survey Methods Section, Statistical Society of Canada, 133-137.

Kott, P.S. and D.M. Stukel (1997), "Can the Jackknife be used with a Two-Phase Sample?. Survey Methodology, Vol. 23, 81-89.

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