

A SOLUTION TO THE DESIGN AND IMPLEMENTATION OF A FAST-TRACK SURVEY: TWO-PHASE SAMPLING

M. Brodeur¹, W. Jocelyn and J. Trépanier

ABSTRACT

A new Retail Commodity Survey was launched by Statistics Canada in January 1997, to obtain more information on commodity by sales in Canada. The survey is based on a two-phase sample design, where the first phase is the Monthly Retail Trade Survey, in place since 1988. Information from the first phase is used in all steps of the commodity survey in order to maximize the efficiency of the design: Multivariate sample allocation, sample selection, maintenance and imputation. Some development has been done to provide explicit variance estimation methods for a two-phase stratified sampling design where the second phase sample is selected from the restratified first phase sample.

KEY WORDS: Sampling; Two-phase; Restratification; Multivariate; Estimation.

1. INTRODUCTION

Today, tight budgets and short deadlines are forcing statisticians to devise innovative sample designs. Survey methodologies must be tailored to achieve a certain degree of efficiency within the limits imposed by such constraints. The Retail Commodity Survey (RCS) presented just such a challenge. It was designed and implemented over a period of one year. Its purpose was to collect detailed information about retail commodity sales in Canada. Since total retail sales were already being measured by another survey, the Monthly Retail Trade Survey (MRTS), we decided to use a two-phase sample design for the RCS in order to take full advantage of the MRTS's information and infrastructure. By taking this approach, we were able to develop the RCS more quickly, more economically and, from a statistical perspective, more efficiently.

In this paper, we present a methodological solution to the problem of designing and developing a fast-track survey: two-phase sampling. A brief description of fast-track surveys is provided in section 2, while section 3 contains an overview of the MRTS. Section 4 provides a detailed picture of the RCS's methodology, including the sampling plan, the sample allocation method, the edit, imputation and estimation strategies.

2. FAST-TRACK SURVEYS

The RCS's main purpose is to measure the distribution of retail sales by commodity group. A quarterly survey, the RCS satisfies requirements stated by data users both within and outside Statistics Canada. 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.

A two-phase sample design is advantageous in many respects. First, it makes it possible to fast-track the introduction of a new survey. In this case, the time limit was one year. Second, the first phase information can be used to maximize 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.

Two-phase sampling was introduced by Neyman (1938) and was later treated at some length by Cochran (1963) under the name "double sampling". Since then, it has been studied by other researchers, including Särndal, Swensson and Wretman (1992) and Hidiroglou and Särndal (1995). However, none of them deals specifically with the case of a stratified design in which the first phase sample is completely restratified to make the second phase sample design as efficient as possible. That is what makes the RCS design innovative.

3. OVERVIEW OF THE MRTS: FIRST PHASE OF THE RCS

The Monthly Retail Trade Survey essentially measures retail sales by trade group (three- or four-digit groups based on the 1980 Standard Industrial Classification (SIC)) for each 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 in the BR as retailers. Some 16,000 companies are interviewed each month. The population is stratified by province, territory, selected CMAs 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

¹ Marie Brodeur, Chief, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

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 or have 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. The target coefficient of variation of sales is 1.5% at the national level, 2.5% at the provincial level and 3.5% at the trade group level. Sample allocation is by the square root of GBI.

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 remains in the sample and outside the sample. A subset of the panels is selected for the sample. Rotation involves systematically removing one panel from the sample each month and replacing it with a new panel. The sample is rotated only in the take-some strata. Each month, births are systematically added to the panels. The expected number of births in the sample is thus reached.

4. RETAIL COMMODITY SURVEY

4.1 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, therefore, be part of the first phase sample. This section deals primarily with stratification, the sample allocation method, and rotation of the RCS sample.

4.1.1. Stratification of the RCS

The frame from which the RCS sample is drawn 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 is 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 sales volume.

MRTS sales were used in determining the company size substrata. For operational reasons, however, that variable was not available for the entire sample. Therefore, sales had to be modelled using GBI, which was available for all units in the population. For consistency, the predicted sales provided by the model below were used to stratify the first phase sample. The following simple regression model was used:

$$V_i = \beta_0 + \beta_1 \text{GBI}_i + \epsilon_i$$

where V_i and GBI_i represent the sales and gross business income of company i . β_0 and β_1 are the model's usual parameters. The model's parameter estimates were used to

predict sales for companies whose sales were unavailable in the first phase sample.

4.1.2 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 plan, an existing method had to be adapted to suit our sampling plan. The method chosen in this case is a modification of the Bethel (1992) algorithm. It is summarized below. For more details, see Jocelyn and Brodeur (1996).

Let K be the commodity groups. We want to estimate T_k , the sales for a commodity group k . Consider the following double expansion estimator for a two-phase sampling plan:

$$\hat{T}_k = \sum_{i=1}^N \sum_{h=1}^H \sum_{g=1}^G \frac{N_h}{n_h} \frac{M_g}{m_g} z_i z_i^{(2)} a_{ih} a_{ig}^{(2)} y_{ik}.$$

The variance of \hat{T}_k is given by:

$$V[\hat{T}_k] = V_1 E[\hat{T}_k] + E_1 V[\hat{T}_k] = V_1 + V_2$$

with

$$V_1 = V_1 E[\hat{T}_k] = \sum_{h=1}^H \frac{N_h}{n_h} (N_h - n_h) S_{hk}^2$$

$$S_{hk}^2 = \left[\sum_{i=1}^N [a_{ih} y_{ik}]^2 - \frac{\left[\sum_{i=1}^N a_{ih} y_{ik} \right]^2}{N_h} \right],$$

$$V_2 = E_1 V[T_k] =$$

$$E \left[\sum_{g=1}^G \left(\frac{M_g(M_g - m_g)}{m_g(M_g - 1)} \right) \left\{ \sum_{i=1}^N \left(\sum_{h=1}^H \frac{N_h}{n_h} z_i a_{ih} a_{ig}^{(2)} y_{ik} \right)^2 - \frac{\left(\sum_{i=1}^N \sum_{h=1}^H \frac{N_h}{n_h} z_i a_{ih} a_{ig}^{(2)} y_{ik} \right)^2}{M_g} \right\} \right].$$

where y_i is the first phase value of the variable of interest for unit i . H is a first phase stratum. a_{ih} is an indicator variable whose value is 1 if unit i is in stratum h , and 0 otherwise. Hence we can write $N_h = \sum_{i=1}^N a_{ih}$. z_i is an indicator variable whose value is 1 if unit i is in the first phase sample, and 0 otherwise. The sample size for the h -th stratum is therefore $n_h = \sum_{i=1}^N z_i a_{ih}$. N is the size of the population. G is a second phase stratum. $a_{ig}^{(2)}$ is an indicator variable whose value is 1 if unit i is in second phase stratum

g , and 0 otherwise, and $z_i^{(2)}$ takes the value 1 if unit i is in the second phase sample, and 0 otherwise. We can therefore write $M_g = \sum_{i=1}^N z_i a_{ig}^{(2)}$ and $m_g = \sum_{i=1}^N z_i^{(2)} a_{ig}^{(2)}$.

Hence the sample allocation problem can be solved by minimizing:

$$C = E_1 \left[\sum_{h=1}^H c_h n_h + \sum_{h=1}^H \sum_{g=1}^G c_g m_g \right]$$

provided that $CV(\hat{T}_k) \leq \mu_k$, $k=1, 2, 3, \dots, K$, where c_h and c_g are the unit costs of first phase and two respectively, and μ_k is the desired coefficient of variation. The following is the procedure for adapting Bethel's algorithm to the problem of second phase allocation where both first phase and second phase sample sizes must be determined at the same time. We first apply the algorithm to the H first phase strata, using V_1 as the variance. We compute the value of V_2 by replacing the corresponding m_g . We then apply the algorithm to V with V_2 modified. Finally, we calculate the coefficient of variation (CV) using the optimal values. If we fail to achieve the desired CV, we start over. We can show that this procedure also preserves the convergence properties of Bethel's algorithm. In our case, since the first phase sample sizes were predetermined, we simply applied the algorithm to modified V_2 .

The CVs from the 1989 survey were used in applying Bethel's algorithm. The final sample size produced by the algorithm was about 10,000 units for a CV of sales of 7% at the Canada level for each major commodity group.

4.1.3. 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 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. That is what we did for the RCS.

To take the MRTS rotation into account, every month we select a subsample of units from the new MRTS panel. Since we assume 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.

4.2. Data collection

As previously mentioned, the RCS sample is a subsample of the MRTS sample. 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 in 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 group. 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 100 commodity groups, which are themselves assembled into major commodity groups such as food, clothing and accessories, and furniture and appliances. The total sales for all major groups equals total retail sales. Respondents can report their sales by commodity group as a dollar amount or as a percentage of their 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 can be input to the collection system and subsequently 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 its industry and commodity groups, as reflected in previous responses. The tailored questionnaire eases the response burden and helps boost the response rate. 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.

4.3. Editing and imputation

As stated above, the RCS uses a complex questionnaire that includes many different commodity groups. These groups in turn form other groups, and so on. In summary, the questionnaire is filled with totals and subtotals and, as a result, it was difficult to develop an editing and imputation strategy. The strategy we devised was implemented within a tight schedule and was not tested since there were no usable historical data. Consequently, we tried to keep it simple, robust and flexible. The editing and imputation system consists of three main modules: pre-editing, automated editing, and imputation. Each module is described in detail below.

4.3.1. Pre-editing

The purpose of pre-editing is to perform a series of checks 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 checked 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.

4.3.2. 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. It allows the substitution of an imputed value for a reported value only if the reported value is involved in sums of parts that do not equal the totals. This is the only rule in the automated edit stage because without historical data we would have had great difficulty in setting the boundaries between the acceptance region and the rejection region for any other type of rule. Of course, this strategy may be reviewed after a few years of production. Checking sums of parts and totals may appear simple, but it is actually very complex because the subtotals are added together to form other totals. When it encounters non-response, the automated editing system tries to determine 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 and eventually the same month of the previous year) and even the unit's industrial classification are used in this process.

4.3.3. Imputation

Prior to actual imputation, a final series of checks is performed on the records that might be used to calculate values imputed to other records. The checks ensure that no outliers will 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 regions may be different. This step is designed to prevent situations such as the following: if a women's clothing store that also sells cosmetics were used, it might generate cosmetics 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 size.

Ratio imputation and adjusted historical imputation are the methods currently used in the RCS. Let $y_{i,t}$ be unit i 's sales of commodity Y at time t . Let $y_{i,t-1}$ and $y_{i,t-12}$ be unit i 's sales of commodity Y the previous month and the previous year respectively. Finally, let $x_{i,t}$ be unit i 's total sales at time t . If unit i 's sales of commodity Y have to be imputed, we will use $y_{i,t}^*$, which can be calculated in one of the following ways.

Ratio imputation:

$$y_{i,t}^* = \frac{\sum_{j \in G} y_{j,t}}{\sum_{j \in G} y_{j,t-12}} y_{i,t-12}, \quad (1)$$

$$y_{i,t}^* = \frac{\sum_{j \in G} y_{j,t}}{\sum_{j \in G} y_{j,t-1}} y_{i,t-1}, \quad (2)$$

$$y_{i,t}^* = \frac{\sum_{j \in G} y_{j,t}}{\sum_{j \in G} x_{j,t-1}} x_{i,t} \quad (3)$$

where G is the set of all units that responded to commodity Y in unit i 's imputation group and passed the above-mentioned checks. Methods 1 and 2 ensure that imputed commodity Y sales have the same relationship to the previous year's or previous month's sales as those of respondent units in the same imputation group. Method 3 ensures that the percentage of imputed commodity Y sales in relation to total sales (X) is the mean percentage for responding units in the same imputation group.

Adjusted historical imputation:

$$y_{i,t}^* = \frac{x_{i,t}}{x_{i,t-12}} y_{i,t-12}, \quad (4)$$

$$y_{i,t}^* = \frac{x_{i,t}}{x_{i,t-1}} y_{i,t-1}. \quad (5)$$

Methods 4 and 5 ensure that the percentage of imputed commodity Y sales in relation to total sales (X) equals the percentage of commodity Y sales the previous year or the previous month.

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, imputation methods 1 and 2 are better for the RCS than methods 4 and 5, and even method 3, since the latter methods tend to conceal changes in the distribution. In addition, wherever possible, imputation is performed within imputation groups defined by industrial classification, geographic area and company size. However, some groups do not contain enough respondent units, and we have to broaden the definition by removing geographic area, for example. Finally, since imputation does not ensure that the parts will add up to the totals, it is followed by a prorating step.

4.4. Estimation

First phase information has been heavily used in every stage of the RCS so far, and the next stage, estimation, will be no different. Since total sales are known in phase one, we chose an estimator that would enable us to use that information 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). For further details, see Binder *et al.* (1997). According to Jocelyn, Brodeur and Babyak (1997), a simulation of a double-expansion estimator produced results comparable to those of the reweighted expansion estimator. However, the double-expansion ratio estimator was selected because the variance estimator is very simple. The rest of this section deals exclusively with the double-expansion ratio estimator. Let \hat{T}_k be the estimator used to measure the total sales of commodity k . It can be written as follows:

$$\hat{T}_k = \frac{\hat{Y}_k}{\hat{X}_k} \hat{X}_k^{(1)}$$

$$\hat{X}_k^{(1)} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{j=1}^{n_h} x_{hjk}$$

where $\hat{X}_k^{(1)}$ represents the estimated sales based on the first phase sample (MRTS) \hat{Y}_k represents the estimated total sales of commodity k , and \hat{X}_k represents the estimated sales from the second phase sample.

The variance estimator of this estimator is given by:

$$\hat{V}(\hat{T}_k) = \hat{V}_1(\hat{T}_k) + \hat{V}_2(\hat{T}_k)$$

$$\text{with } \hat{V}_1(\hat{T}_k) = \sum_{g=1}^G M_g^2 (1 - f_g^{(2)}) \frac{s_{2g}^2}{m_g}$$

and

$$\hat{V}_2(\hat{T}_k) =$$

$$\sum_{h=1}^H \sum_{g=1}^G \frac{N_h^2 (1 - f_h) M_g^2 (1 - f_g^{(2)}) \frac{s_{1(h)g}^2}{m_g}}{n_h^2 (n_h - 1)} +$$

$$\sum_{h=1}^H N_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

where s_{2g}^2 , $s_{1(h)i}^2$, s_h^2 are defined in Binder *et al.* (1997), and $f_g^{(2)} = m_g / M_g f_h = n_h / N_h$.

This variance estimator was developed with the Taylor linearization method (Binder (1996)).

5. CONCLUSION

The RCS was developed very quickly. Choosing a restratified sampling plan for the second phase helped us

keep to a very tight schedule; it also improved the design's efficiency and reduced costs by using auxiliary information. At present, the data are being collected monthly, but the results are published quarterly. Hence it is essential to estimate the covariance since the samples are not independent from month to month.

A few challenges remain. For example, it will be beneficial to track commodity imputation over time and adjust the imputation strategy as required. Also, the commodity data for many companies remain quite stable from quarter to quarter. We are considering switching to annual collection in those cases.

6. ACKNOWLEDGMENTS

The authors wish to thank all those who worked on the methodology of the RCS, including Colin Babyak, Hélène Bérard, Sarah Franklin, Julie Girard, Leslie Kiss, Martin St-Pierre and Janet Sear. The authors would also like to thank François Maranda and Don Royce for their invaluable comments.

REFERENCES

- Binder, D.A. (1996). Linearization methods for single phase and two phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.
- Binder, D.A., Babyak, C., Brodeur, M., Hidioglou, M., and Jocelyn, W. (1997). Variance estimation for two-phase stratified sampling. *Proceedings of the Section on Survey Research Methods*, Annual American Statistical Association. To be published.
- Cochran, W.G. (1963). *Sampling Techniques*. John Wiley.
- Hidioglou, M.A., and Särndal, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods*, Annual American Statistical Association, 873-878.
- Jocelyn, W., and Brodeur, M. (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., Brodeur, M., and Babyak, C. (1997). Comparaisons de différents estimateurs de variance à deux phases: étude Monte-Carlo basée sur l'Enquête des marchandises au détail. *Recueil des de la section des méthodes d'enquête*, SSC, juin 1997.
- Kott, P.S., and Stukel, D.M. (1997). Can the Jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-89.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.