

# THE REDESIGN OF THE CANADIAN ANNUAL WHOLESALE AND RETAIL TRADE SURVEY

Zdenek Patak, Keith Davis, Serge Grenier

Zdenek Patak, 11-N, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

**Key Words:** Survey Redesign, Record Linkage, Calendarization

## 1. Introduction

The primary reason for the redesign of the Canadian Annual Wholesale and Retail Trade Survey (AWRTS) was twofold: (a) to produce timely operating financial statistics with little reliance on data coming from administrative sources; (b) to reduce survey costs by utilizing generalized software developed at Statistics Canada such as GSAM (Generalized Sampling System), GEIS (Generalized Edit and Imputation System) and GES (Generalized Estimation System), and the systems built for a sister survey, MWRTS (Monthly Wholesale and Retail Trade Survey). A secondary objective was to give the end user calendar year estimates at the national and provincial levels.

In the last decade or so most annual business surveys conducted at Statistics Canada used data collected by direct surveys supplemented with information obtained from administrative sources, namely tax data. This was time consuming. In addition, imperfections on the frame led to biases that were commensurate with the estimated reliability levels of the corresponding monthly surveys.

Over time the increasing reliance on tax data contributed to delays in the release of estimates. For instance, the previous AWRTS published data almost two years after the end of the reference period. Additionally, changes in content and processing raised concerns as to the quality of the tax data being blended with survey responses. As a result it was decided to use monthly surveys as a vehicle for carrying out annual surveys. This approach was much less time consuming, with preliminary estimates being available within six to nine months after the end of the reference period - the calendar year.

The quick release of final figures put the redesigned AWRTS into the realm of *fast track surveys*, a term used to describe surveys that could convert design specifications into the final product within a very short time (13 months for AWRTS). The idea, although not new, was deemed experimental and served as the proving ground for other surveys of this type.

Currently the redesigned AWRTS is the only annual business survey in production at Statistics Canada driven almost entirely by its monthly counterpart. All other annual surveys still produce estimates from a mixture of administrative and respondent data. This is, however,

changing as more business survey taking areas within the corporation are adopting the AWRTS approach. Transportation's Q5 is one such example; a quarterly survey will be the basis for its annual counterpart.

The previous AWRTS is described in Section 2. Some aspects of the sampling frame and MWRTS upon which the redesigned survey is based follow in Sections 3 and 4. The last three sections elaborate on the current incarnation of AWRTS, planned future enhancements that are already under way and some concluding remarks along with an evaluation of the new approach.

## 2. Previous Annual Wholesale and Retail Trade Survey

The previous annual survey came into existence in 1984. Its mandate was to produce limited operating statistics from a mixture of survey and administrative data. The objective of the program was to reduce the response burden imposed on small businesses without affecting the overall quality of the estimates - before 1984, the AWRTS was a census of retail and wholesale trade. An additional benefit of shifting the emphasis away from respondent data toward administrative data was the reduction in the cost incurred by the collection and editing processes.

Between 1984 and 1989, all businesses with a revenue above one million dollars were surveyed directly while those with a revenue under this threshold were estimated from tax records. In 1990 the survey portion was downsized to include only the **Large** (see Section 3) units on the Business Register. This change decreased the number of directly surveyed companies and increased the contribution of the tax portion to the estimates. For example, for wholesale trade, administrative sources represented 5.4% of sales in 1989 and 15.2% in 1990.

The non-surveyed portion of AWRTS consisted of tax reporting units. To avoid duplication of records, all tax reporting units belonging to businesses in the surveyed portion were to be removed. The process used to identify duplicates was done in two stages. First automated matching was invoked. Second a manual process by which a person experienced in record linkage and unduplication removed any units missed in the first stage. The success or failure of the manual process depended heavily on the person's own expertise in the field of record linkage. Often the smaller units eluded being linked and thus contributed to an upward bias and a larger mean squared error of the estimates.

Another source of bias stemmed from inaccurate or incorrect classification of units in the non-surveyed portion. Tax units were often miscoded in the third and fourth digits of the Standard Industrial Classification (SIC). Unfortunately, their sheer volume made accurate classification prohibitive. Uncertainties as to the quality of the coded economic activity resulted in switching to a two-phase sample of tax units. The realized sample was classified to the four-digit SIC level. This alleviated the problem of misclassification bias, but did not completely eliminate it as some businesses remained miscoded due to the lack of a detailed description of their economic activity.

The period for which data were collected was the company's fiscal year. These represented a wide spectrum of fiscal end dates ranging from April 1 of the **reference year** to March 31 of the **following year**. Revenue Canada Taxation allowed corporations to file their tax returns within six months of the end of their fiscal year; some companies responded as late as September or October of the **following year**, or after their financial statements had been prepared for taxation purposes. This contributed to the lengthy duration of the survey process.

For approximately 40% of the respondents the fiscal period corresponded to the calendar year. Despite the obvious inconsistency in the reported fiscal year ends no adjustments were applied to *calendarize* the data. It was assumed that no systematic biases were introduced because of not adjusting non-calendar year responses.

The mail-out commenced in March of the following year. Data collection was directed by Operations and Integrations Division, a central service area within Statistics Canada, by mailing a questionnaire to all surveyed companies. Mail and/or telephone follow-ups were done for late respondents. All returns were screened for completeness and, where necessary, a follow-up was initiated to obtain missing information, or to confirm, explain, and correct suspicious data.

Non-responses were manually imputed for units that were active during the survey year. Some imputation was also based on data reported to the monthly survey (if MWRTS data were reported by the respondent) by the delinquent units for the same period, or on tax data, growth rates and cell ratios calculated from data provided by respondents with similar characteristics.

AWRTS estimates were published at the SIC and geographical unit levels (a geographical unit was either a province or a territory). Given the detail at which the publication was produced, the application of confidentiality was done manually. This process was time consuming as many conditions and regulations required by Statistics Canada had to be observed to protect the

confidentiality of the respondent under the Statistics Act.

The heavy use of manual intervention throughout pre- and post-collection processing made the previous AWRTS highly susceptible to human error. In addition, differences between the annual survey and its monthly counterpart were often difficult to explain due to the lack of control of the analyst over data coming from administrative sources.

### 3. Some Aspects of the Sampling Frame

The target population for AWRTS, both previous and redesigned, consists of all businesses operating in Canada that have at least one retail location or wholesale establishment within their structure. These units make up the sampling list frame that resides on the Business Register (BR). The BR is a database that uses many different files from Revenue Canada Taxation and data collected by direct surveys to maintain and update existing information.

To manage the BR in a cost-effective manner, businesses on the frame are classified into three categories: Large, Small and Out-of-scope. A business is considered **Large** when its Gross Business Income (GBI - an annualized size measure derived from tax data) exceeds a predetermined threshold for its geographical and industrial classification. Businesses with a positive GBI that is less than the **Large** threshold belong in the **Small** portion. All others are classified as **Out-of-scope** and adjusted for at estimation.

### 4. Monthly Wholesale and Retail Trade Survey

The MWRTS is fully contained within the BR and as such reaps the benefits associated with a single business frame, i.e., unduplication, classification, and other survey related activities are highly automated, and designed in a way that is conducive to reducing non-sampling error.

The population is stratified by geography, trade group and size. The **geography** variable encompasses ten provinces and two territories for both retail and wholesale plus four metropolitan areas for retail only. **Trade group** denotes SIC aggregation of retail and wholesale activities. The **size** variable partitions the population into take-all (units selected with certainty), large take-somes and small take-somes.

The monthly sample contains approximately 3,500 large companies that operate in multiple regions and/or trade groups. For stratification these units are assigned their dominant geography and trade group (i.e., the origin of the largest contribution to their total GBI).

MWRTS employs rotation group sampling. The population in each stratum is randomly assigned to

clusters, or rotation groups. The number of clusters in and out of sample depends on the sampling fraction in the stratum, plus time-in and time-out constraints. The Retail Trade portion of the survey is currently using a rotation scheme, whereas its Wholesale counterpart has not yet begun.

The survey collects sales and the number of locations from all sampled companies; inventories are collected monthly from all wholesalers and a subset of retailers (e.g., companies whose annual sales exceed a preset threshold or which are in a specific trade group). Data collection is done by mail and by telephone with all follow-ups handled strictly by telephone. Except for a few head office respondents, the bulk of the data is collected in seven regional offices.

Data for non-respondents are imputed by an automated system that uses several imputation methods depending on what current/historical information is available. Most of the non-response is imputed systematically. Anomalies to which Subject Matter Officers are alerted are resolved manually.

For estimation the reported data are distributed into domains to measure more accurately the activity by individual geography and trade group. These domains may differ from the original strata because of businesses changing industry, location or size. The parameters of interest that we wish to estimate are total sales, inventories and the number of locations. The corresponding statistics are stratified  $\pi$ -weighted domain estimates adjusted for the realized sample size.

MWRTS gives the end user estimates at the trade group and geography levels. The publication describes monthly and annual trends as well as deseasonalized economic activities. Confidentiality is applied systematically with all Statistics Canada regulations strictly observed to protect the confidentiality of the respondent as required under the Statistics Act.

The monthly survey is well established with response rates of more than 95%. Given its successful track record, a decision was made in 1992 to carry out a pilot study to assess the feasibility of using the monthly sample as a vehicle for conducting an annual survey.

## **5. The Redesigned Annual Wholesale and Retail Trade Survey**

In the pilot survey the emphasis was placed on finding out whether the monthly contacts would be able to provide sufficient information to compute gross margin and the provincial distribution of key variables. A sample of 1,500 units was selected for this purpose. The conclusions were favourable and 1993 was to be the first year of the

redesigned survey driven almost entirely by its monthly counterpart.

The 1993 annual sample contained all units sampled by MWRTS at any time during 1993. To accommodate the requirements of the Annual Retail Chain Survey, the monthly sample was augmented to include out of sample retail chains. Some grain and petroleum dealers were also added to produce estimates for a sector that is not normally covered by MWRTS.

The sampling frame used for conducting the monthly survey represents a snapshot in time. The annual frame being a union of twelve monthly frames must integrate several snapshots in time because it has to convey the dynamic nature of the economy by adjusting for births and deaths as well as promotions and demotions initiated by the BR. This information is used to produce nearly unbiased annual weights.

It was mentioned previously that duplication on the frame may cause an upward bias in the estimates. Even with a single frame, duplication does occur. This takes place when the BR promotes a unit from the Small to the Large sector within the frame, i.e., the Small unit is deceased and in its place a new Large unit is birthed. The absence of direct links between the two portions dictates that both units remain on the frame causing monthly estimates, although unbiased, to be less efficient. However, there is no easy way for the system to unduplicate these records. This also applies to the corresponding collection entities (reporting units in the US). If they were not correctly linked, estimates for the annual survey would be biased due to double counting.

### **5a. Sampling Weights and Estimation**

The preceding paragraph describes a fictitious transaction as it does not reflect the real world, but rather attempts to improve the efficiency of monthly sampling. Real changes in the business world (e.g., takeovers or buy-outs) impact the quality of the sampling frame in much the same way. Both real and fictitious changes result in erroneous estimates of reduced efficiency. To combat this problem the frame has to be unduplicated in an unbiased manner using a source that is independent of the survey.

The unduplication of the sampling frame may be done by identifying Payroll Deductions Accounts (used by employers for remitting taxes on behalf of their employees) that have gone from Small to Large portion of the frame during the survey reference year. These accounts are then linked to businesses to which they belong. When a business is promoted, it is placed on the frame under a new identifier while it remains associated with the same Payroll Deduction Account. The new business appears as a birth and as such is subject to

sampling. This has to be reflected in the annual sampling weights.

MWRTS uses weights representing the inverse of the probability of selection adjusted for the realized sample size. A similar approach is adopted for AWRTS where simple probabilities of selection are replaced with combined probabilities of selection defined by

$$p_i^{comb} = 1 - \prod_{j=1}^{\# \text{ sampled}} (1 - p_j),$$

where  $p_j$  is the probability of selection on a given occasion and  $n_j$  is the number of times a unit  $i$  has been eligible for sampling. A set of preliminary inversion weights may be obtained by taking the inverse of  $p_j$ , i.e.,

$$W_i^{comb} = 1 / p_i^{comb}$$

The individual inclusion probabilities may be computed using the information generated by the sampling system. Included therein are three stratification variables: trade group, geography and size as well as parameters used by the rotation scheme:  $P^{in}$  (number of rotation groups in sample) and  $P$  (number of rotation groups in the population).  $P^{in}$  and  $P$  give us the probabilities of selection  $p_i = P_i^{in} / P_i$ .

For retail, inclusion probabilities have to be adjusted for the number of rotation groups present in the AWRTS sample. Hence, the final working formula for  $p_i$  may be written as

$$p_i^{comb} = 1 - \prod_{j=1}^{\text{times sampled}} (1 - \min(1, (P_j^{in} + M) / P_j)),$$

where  $M$  is the number of rotations considered.

The inverse of the inclusion probabilities gives us a set of preliminary weights. These are adjusted for the realized sample size to produce the final estimation weights defined as

$$W_{hi}^{final} = N_h W_{hi}^{comb} / \sum_j W_{hj}^{comb},$$

where  $h$  and  $N_h$  denote the stratum and stratum population count.  $W_{hi}^{final}$  is the weight used in estimation. The adjustment is used to reduce the variance of the estimates and may be thought of as the  $g$ -weight in a Generalized Regression setup, i.e.,

$$g_{hi} = N_h / \sum_{i=1}^n w_{hi}^{comb} = N_h / \hat{N}_h.$$

The domain estimate of a total is then

$$\hat{Y}_h^k(d) = \sum_{i=1}^{n_h} g_{hi} w_{ih}^{comb} y_{hi}^k(d)$$

for variable  $k$ .

Unequal weights for units in many strata prevent us from using the simple stratified SRSWOR formulation of variance. Moreover, given the unavailability of joint probabilities of selection it cannot be computed exactly. A good approximation for the variance of the parameters of interest (totals and ratios of totals) in this hybrid SRSWOR / PPSWOR setup may be obtained employing the theory developed in Hidioglou (1993).

This is done by first transforming the  $g$ -weights along with the  $y$  and  $x$  vectors by  $t_{hi} = \left\{ w_{hi}^{final} + \left( \frac{N_h}{n_h} \right) \right\}^{1/2}$ , and then by simply choosing the GREG variance for SRSWOR option in the Statistics Canada Generalized Estimation System used for estimation, i.e.,

$$\hat{v}(\hat{Y}_h^k(d)) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\hat{D}_{hi}(d) - \bar{D}_h(d))^2,$$

where after some manipulation

$$\hat{D}_{hi}(d) = g_{hi} w_{hi}^{comb} \left\{ y_{hi}(d) - \mathbf{x}_{hi} \left[ \left( \sum_i w_{hi}^{comb} \mathbf{x}_{hi} \mathbf{x}_{hi}^T \right)^{-1} \sum_i w_{hi}^{comb} \mathbf{x}_{hi} y_{hi}(d) \right] \right\}.$$

The variable  $y_{hi}(d)$  is equal to  $y_{hi}$  if the  $hi$ -th unit belongs to domain  $d$  and zero otherwise;  $\mathbf{x}_{hi}$  denotes the auxiliary count variable.

## 5b. Redesigned AWRTS - continued

The redesigned annual survey may be thought of as the 13th period of the monthly survey. The surveyed AWRTS sample units represent virtually the same units as those surveyed by the December 1993 MWRTS, augmented by the inclusion of non-MWRTS retail chains and some grain and petroleum dealers. The monthly survey collects only sales, inventory and the number of locations' data, whereas its annual counterpart is expected to solicit other financial data as well.

Data collection was done by mail and by telephone using a CATI system; non-respondents were followed up strictly by telephone. With the exception of a few head office collections and the retail chain portion, the bulk of the data was solicited in seven regional offices.

The period for which businesses were asked to provide data was the 1993 calendar year. There were some

companies that could not report in said manner and would give fiscal year data. Estimates based on a mix of reporting periods were often difficult to interpret and meant very little to the user. For this reason, a *calendarization* system was developed to align all responses to the calendar year.

Continuous variables such as total operating revenue and cost of goods sold were calendarized by simply taking the product of reported AWRTS data and the ratio of calendar MWRTS sales to fiscal MWRTS sales, i.e. ,

$$y_{i,cal}^{AWRTS} = y_{i,fiscal}^{AWRTS} \left\{ \frac{\sum_{i \in cal} Sales_i^{MWRTS}}{\sum_{i \in fiscal} Sales_i^{MWRTS}} \right\}.$$

Stock variables such as opening and closing inventories were calendarized in a similar manner, preserving the cost of goods sold identity, either by adjusting their net difference or by adjusting purchases. The corresponding method depended on whether MWRTS inventories were deemed accurate or merely approximate. The number of locations was accepted as reported.

Calendarization of AWRTS data did not only lend uniformity to the estimates, it also facilitated comparisons with MWRTS. Note that at the national level calendar year alignment represented approximately a 2% difference with annual fiscal data.

Non-response was imputed by an automated system using several imputation methods depending on what current/historical data were available. Initially total operating revenue was imputed using MWRTS data, trend adjusted previous year annual data or cell means computed using complete respondents of similar characteristics. Sales ratios were used to impute all other missing variables. Complete non-response was handled using donor imputation. Anomalies were brought to the attention of Subject Matter Officers and resolved manually.

To reduce the number of gross response and typographical errors, a *multivariate outlier detection* routine was applied to all complete and partial respondents prior to imputation. The method used robust Mahalanobis distance

$$(\mathbf{x}_i - \bar{\mathbf{x}}_{rob})^T \Sigma_{rob}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{rob})$$

to identify units for which this statistic exceeded a prespecified percentile of a  $\chi_p^2$  distribution, where  $p$  was the number of degrees of freedom. The  $\bar{\mathbf{x}}_{rob}$  was the  $L_1$ -estimator of multivariate location (Rousseeuw and Leroy, 1987) and  $\Sigma_{rob}$  a robust estimator of scatter computed using a method similar to projection pursuit. The

sensitivity of the process could be adjusted to suit the survey's needs. The detected outliers were excluded from the computation of imputation parameters.

Included in the estimates were all surveyed units as well as those businesses that ceased their operation during the calendar year, if they showed any economic activity measured by MWRTS. To facilitate the process of quantifying deaths along with imputation and calendarization, a Parametric Data File (PDF) was created. This file contained all businesses sampled by MWRTS along with their monthly sales, the number of locations, inventories and all known linkages for units that underwent changes.

To cover all possible reporting periods, data extraction was done for 24 months (April, previous year to March, following year). The creation of the PDF is fairly complex and any interested reader should refer to a Statistics Canada internal document entitled "*The Setting Up of the Period 13 Parametric Data File*".

The publication put out by the redesigned AWRTS provided the user with estimates computed at the trade group and geography levels, as well as the 4-digit SIC, so that continuity with the previous annual survey was maintained. Again, confidentiality was applied systematically in a manner similar to the monthly.

## 6. Qualitative Analysis and Future Enhancements

The timely availability of a consistent set of limited operating annual statistics complimented MWRTS estimates in giving the user a more complete picture of the performance of wholesale and retail trade. The timeliness factor also greatly enhanced the relevance of the published data making them more marketable.

The format of the questionnaire was well received by respondents and interviewers alike. The collection period was reduced from eight months required for the previous annual survey to approximately forty five days, without sacrificing data quality. The average absolute difference between retail portions of MRTS and ARTS declined from 3.3% in 1991 and 3.9% in 1992 to only 1.0% in 1993.

Similar improvements have also been realized at the trade group level. The remaining discrepancies can be attributed largely to the respondents' interpretation of what data are being solicited by the two surveys. This is discussed in more detail in Davis et al. (1995).

While in the past micro and macro level comparisons were hampered by the wide range of fiscal periods reported by the respondents, the current calendarization module has greatly simplified the task by aligning all responses to the

calendar year. Table 1 shows the impact of calendarization on Wholesale Trade data at the national level. Also presented are two industry levels where we contrast the minimum and maximum effects of the procedure.

**Table 1. Impact of Calendarization on Wholesale Trade Data**

| Canada (\$000,000)          | Calendar | Fiscal  | % Dif |
|-----------------------------|----------|---------|-------|
| Sales                       | 253,956  | 251,419 | 1.01  |
| Purchases                   | 202,336  | 200,174 | 1.08  |
| Open Inv                    | 28,686   | 28,505  | 0.63  |
| Clos Inv                    | 30,543   | 30,350  | 0.97  |
| COG                         | 200,479  | 198,428 | 1.03  |
| Salaries                    | 21,469   | 21,219  | 1.18  |
| Gross Margin                | 53,477   | 52,992  | 0.92  |
| <b>Hardware (\$000,000)</b> |          |         |       |
| Sales                       | 14,929   | 14,894  | 0.23  |
| Purchases                   | 11,983   | 11,952  | 0.27  |
| Open Inv                    | 2,204    | 2,218   | -0.64 |
| Clos Inv                    | 2,371    | 2,379   | -0.35 |
| COG                         | 11,816   | 11,790  | 0.22  |
| Salaries                    | 1,580    | 1,580   | 0.03  |
| Gross Margin                | 3,112    | 3,104   | 0.27  |
| <b>Food (\$000,000)</b>     |          |         |       |
| Sales                       | 54,942   | 54,477  | 0.85  |
| Purchases                   | 46,529   | 46,169  | 0.78  |
| Open Inv                    | 3,568    | 3,361   | 6.15  |
| Clos Inv                    | 3,559    | 3,389   | 5.02  |
| COG                         | 46,538   | 46,141  | 0.86  |
| Salaries                    | 3,167    | 3,128   | 1.26  |
| Gross Margin                | 8,404    | 8,337   | 0.81  |

The work to enhance further the AWRTS is continuing. For the 1994 survey occasion, almost every part of the annual system has undergone a facelift. The improvements start with an updated questionnaire. It now resembles more accurately a company's financial statement; both revenues and expenses are fully itemized. This reduces response burden as data manipulation required from the respondent is kept to a minimum. The edit parameters have been changed to integrate the 1993 data within the CATI system to provide a more complete suite of current and historical edits.

The imputation module has been rewritten with more emphasis placed on ratio imputation. For donor imputation extreme observations in each cell are removed prior to selecting a suitable donor. The PDF has been improved to represent more accurately the MWRTS historical data series down to the provincial level. The module now also includes a step for annual survey purposes whereby missing data are imputed (e.g., after a business has been rotated out of the monthly sample).

## 7. Concluding Remarks

The changes to the Statistics Canada's annual survey program covering wholesale and retail trade provide a textbook example of the advantages inherent in a sample-based survey. More timely data are available at a lower cost; response burden is reduced and greater attention is paid to the quality of the responses, with little loss of reliability of the estimates. In fact, there are early indications that the change in the survey vehicle improved the reliability of the results.

The use of Statistics Canada generalized systems reduced manual intervention and the chance of a human error creeping into survey processing. Other refinements such as calendarization, outlier detection and the use of a historical Parametric Data File in several post-collection steps make the AWRTS a significant venture in the improvement of annual business survey methods.

## 8. References

- Davis, K and Patak, Z (1995). *The Setting Up of the Period 13 Parametric Data File*. Statistics Canada Internal document, April 6, 1995.
- Davis, K., Murphy, L. And Simard, G. (1995). *Survey Coverage Improvements Using a Common Frame: The Case of Conducting an Annual Survey from a Monthly Frame*. The Proceedings of the American Statistical Association.
- Hidirolou, M.A., Hoyt, P., Nemes, B. and Theriault, B. (1992). *Proposed Annual Wholesale and Retail Trade Principal Statistics Survey*. Internal Statistics Canada document, July 7, 1992.
- Hidirolou, M.A. (1993). *Using the Current Stratified SRS Version of GES for Stratified PPS (WOR or WR) Designs*. Internal Statistics Canada Memorandum, February 8, 1993.
- P-13 Steering Committee (1993). *Period 13 Pilot Survey*. Internal Statistics Canada document, October 1, 1993.
- Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons.
- Ryten, J.R. (1992). *My Four Thoughts of the Week*. Internal Statistics Canada Memorandum, March 3, 1992.