

DEALING WITH MISCLASSIFIED UNITS IN REPEATED BUSINESS SURVEYS: THE EXPERIENCE OF THE REDESIGNED CANADIAN MONTHLY WHOLESALE AND RETAIL TRADE SURVEY (MWRTS)

Hélène Bérard ¹

ABSTRACT

Business surveys dealing with skewed populations often resort to stratified sampling designs. The efficiency of the stratified sampling design depends largely on the quality of the stratification variables in terms of their accuracy and timeliness, and their stability through time in the case of repeated surveys with sample overlap. The Monthly Wholesale and Retail Trade Survey (MWRTS) of Statistics Canada uses a stratified sampling design that can suffer from the undesirable effects of misclassified units on the level and the precision of the estimates. Many innovative solutions have been developed within the sampling design, within the frame and sample update processes, and at the estimation stage to reduce the impact of such units for the redesigned MWRTS that will be implemented in 2004. We developed an improved size measure for stratification purposes using data from other surveys and administrative files like the Goods and Services Tax (GST) files. We used an enhanced version of the Lavallée-Hidiroglou algorithm which takes into account the impact of dead units when determining size boundaries and an optimal sample allocation. We use sample constraints such as maximum allowable weights to reduce the impact of misclassified units in take-some strata. New frame and sample update procedures allow the monitoring and treatment of units that are no longer part of the proper size stratum in an unbiased manner. To deal with misclassified units at the estimation stage we investigated the use of post-stratification based on counts and developed an outlier detection and treatment strategy. In this communication, each of the new measures developed to reduce the impact of misclassified units in the MWRTS is discussed in detail, as well as their implementation in the context of a monthly production cycle.

KEY WORDS: Misclassified units; Outliers; Repeated surveys.

RÉSUMÉ

Les enquêtes économiques dont les populations sont asymétriques utilisent souvent un plan d'échantillonnage stratifié. L'efficacité du plan stratifié dépend largement de la qualité des variables de stratification (dont l'exactitude et l'actualité) et, dans le cas des enquêtes répétées avec chevauchement d'échantillon, de leur stabilité dans le temps. L'Enquête mensuelle sur le commerce de gros et de détail (EMCGD) de Statistique Canada utilise un plan d'échantillonnage stratifié sensible aux effets indésirables des unités classées de façon erronée sur le niveau et la précision des estimations. Afin de réduire l'impact de ces unités dans l'EMCGD remaniée qui débutera en 2004, plusieurs solutions innovatrices ont été développées touchant le plan d'échantillonnage, les processus de mise à jour de la base de sondage et de l'échantillon, et le module d'estimation. Nous avons développé une nouvelle mesure de taille en utilisant des données provenant d'enquêtes, et de sources administratives dont les fichiers de données sur la Taxe sur les produits et services (TPS). Nous avons utilisé une version améliorée de l'algorithme de Lavallée-Hidiroglou qui tient compte des unités mortes lors de la détermination des bornes selon la taille et la répartition optimale de l'échantillon. Nous avons utilisé des contraintes d'échantillonnage comme la détermination des poids maximums afin de réduire l'impact des unités classées de façon erronée dans les strates à tirage partiel. De nouvelles procédures pour la mise à jour de la base de sondage et de l'échantillon permettent de mieux suivre l'évolution de la mesure de taille des unités et de traiter des unités qui ne sont plus dans la strate appropriée d'une façon non biaisée. La possibilité d'appliquer à l'estimation une post-stratification basée sur les comptes de population a aussi été évaluée et une stratégie pour la détection et le traitement des valeurs aberrantes lors de l'estimation a été développée. Cette communication présentera en détail chacune des mesures élaborées afin de réduire l'impact des unités classées de façon erronée et leur utilisation dans le cycle de la production mensuelle de l'EMCGD.

MOTS CLÉS : Classification erronée; valeurs aberrantes; enquêtes répétées.

¹Hélène Bérard (Helene.Berard@statcan.ca), Business Survey Methods Division, Statistics Canada, Tunney's Pasture, R.H. Coats Building, 11th floor, Ottawa, Canada, K1A 0T6,

1. INTRODUCTION

Statistics Canada conducts a major repeated survey known as the Monthly Wholesale and Retail Trade Survey (MWRTS). The survey produces estimates based on monthly data collected for sales and inventories for various geographical region and industry groups. The estimates derived from the survey form a substantial portion of the monthly estimates for the Gross Domestic Product (GDP) and the sales trend represents an important economic indicator.

The MWRTS was completely redesigned in part to provide estimates for the new North American Industry Classification System (NAICS) and to take full advantage of the availability of administrative data from the Goods and Services Tax program. The redesign also addressed the need to reduce cost and respondent burden, update computer systems, and harmonise various concepts with those used by other annual surveys. Estimates from the redesigned MWRTS will be released in 2004 after extensive testing. The testing will take place over a period of five months when both the old and the new MWRTS will be in production.

The new MWRTS will continue to use a stratified random sample design with monthly samples. The monthly samples will consist of ongoing units and new units (births). The stratified design with maximisation of monthly sample overlap is efficient to survey skewed populations where both quality level and trend estimates are required. The efficiency of the stratified sampling design depends largely on the quality of the stratification variables in terms of accuracy and timeliness, and in the case of trend estimates, their stability through time. This paper describes the impact of misclassified units and reviews the many innovative solutions that have been developed within the sampling design, within the frame and sample update processes, and at the estimation stage to reduce the impact of misclassified units on both level and trend estimates of the new MWRTS.

2. MISCLASSIFIED UNITS

In the context of a stratified design, a unit is considered misclassified if any of its stratification variables is inappropriate. The stratification variables used for MWRTS are the industry groups (under NAICS), the geographical regions and the size measure that represents a proxy for annual sales. A unit will be considered misclassified if the stratification information used causes the unit to be in the wrong industry group, wrong geographical region, or wrong size stratum.

A unit can be misclassified at birth or following its birth. In the latter case, the industry group, geographical location, or size measure used when the unit was originally stratified are no longer adequate. In the context of MWRTS, misclassification by size is more frequent than misclassification by industry or geography. Due to the dynamic nature of retailers and wholesalers, misclassification by size following birth is quite frequent.

Misclassified units result in increased variances of the estimates (Hidioglou and Laniel, 2001). For example, size misclassification of new units (births) with extremely large sales can cause important fluctuations in the estimates and a potential increase in the estimates of variance. Similar effects can be expected over time by units that are no longer in the appropriate size stratum because of their uncharacteristic sudden growth or decline. The presence of misclassified units in the sample can cause recurring problems in the estimates of monthly population totals and trends. These cases trigger considerable analysis by subject matter experts to determine if these fluctuations reflect movements in the economy or frame deficiencies.

3. MWRTS STRATEGY TO DEAL WITH MISCLASSIFIED UNITS

The strategy that we developed consisted of improving the methods historically used to efficiently deal with misclassified units in the old MWRTS and implementing new innovative solutions to reduce the occurrence and impact of misclassified units on estimates. We developed an improved size measure for stratification purposes to reduce the occurrence of misclassified units at birth. We used an enhanced version of the Lavallée-Hidioglou algorithm that takes into account the impact of dead units when determining size boundaries and optimal sample allocation. We applied more restrictive sample constraints (allowable maximum weights and minimum sample size in take-some strata) to reduce the potential impact of misclassified units on the estimates. New frame and sample update procedures allow us to monitor and treat units that are no longer part of the proper size stratum in an unbiased manner. To deal with misclassified units at the estimation stage we will continue to use domain estimation; however to treat size misclassification, we investigated the use of post-stratification based on counts and we developed an outlier detection and treatment strategy. These new solutions are

discussed as well as brief descriptions of the new MWRTS sample design and sample and population maintenance procedures.

3.1. Sample Design

The new MWRTS still extracts its frame from the Business Register (BR) maintained by Statistics Canada. The businesses on the BR are represented by a hierarchical structure that has four levels, with the enterprise at the top, followed by the company, the establishment, and the location. An enterprise can be linked to one or more companies, a company can be linked to one or more establishments, and an establishment can be linked to one or more locations. The new MWRTS sampling unit is defined as the cluster of establishments of the same enterprise that operate in the same industry group and same geography. The sampling units used in the Annual Wholesale Trade Survey (AWTS) and the Annual Retail Trade Survey (ARTS) that were recently redesigned are defined in the same manner (Parent and Simard, 2000).

The new MWRTS continues to use a stratified design with simple random sample selection in each stratum. The stratification is done by industry groups using the NAICS four digit level, and the geographical regions consisting of the provinces and territories, as well as three sub-provincial regions for the retail component. We further stratify the population by size. The size strata consist of one take-all, at most two take-some strata, and one take-none stratum. Take-none strata serve to reduce respondent burden by excluding the smaller units from the surveyed population. These units should represent at most five percent of total sales. Questionnaires will not be sent to these units instead, estimates will be produced through the use of administrative data. Three methods implemented within the sample design to deal with misclassified units are described below.

3.1.1. New Size Measure for Stratification

Different size measures are available from our frame. For employers and non-employers that have annual revenues above \$30,000, an annual size measure is calculated (since 1997) using the sales reported through the Goods and Services Tax program (GST sales). For incorporated businesses, another annual size measure is available that is calculated based on the annual revenue from the income tax reports of corporations (T2-revenue). In this case, the size measure represents a reference period of the previous year or older. For employers only, an annual size measure is also obtained throughout the derivation of the annual Gross Business Income (GBI) that is modelled using information on the number of employees and the value of employer remittances to the Canada Customs and Revenue Agency (CCRA).

We developed a new size measure for stratification purposes in order to reduce the occurrence of size misclassification. The GBI was the size measure used in the old survey developed in 1988. For the new MWRTS, the new size measure is created using a combination of independent survey data and three administrative variables: the GBI, the GST sales, and the T2-revenue. The independent survey data consist of the annual sales available from respondents to the old MWRTS, the AWTS, and the ARTS. The new MWRTS sample is drawn independently of the samples from the old MWRTS, the AWTS, and the ARTS. For respondents to the new MWRTS that were also respondents in the old MWRTS, the AWTS, or the ARTS, the size measure is set to the annual sales obtained from the most up to date information that can be obtained from these surveys. Recent respondents' data are deemed to be a better indicator of size than administrative data. When respondents' data from independent surveys are not available, the size measure is set equal to the largest of the administrative variables among the GBI, the GST sales, and T2-revenue that are available for that business. All three administrative variables may be available for employers. For non-employers, the GBI is not available; only the GST sales and the T2-revenue may be present. Note that employers represent respectively 93% and 92% of wholesale total sales and retail total sales. For employers, a 10% reduction in the misclassification rate is observed when the size measure relies on the three administrative variables instead of the GBI alone. This new size measure is particularly efficient in identifying large businesses that should be classified in the take-all stratum. The use of independent survey data is restricted to the initial sample allocation since independent survey data are not available for units born following the initial stratification. However, the new size measure derived from the administrative sources is used to reduce size misclassification at birth for units entering the population following the initial sample selection.

3.1.2. Use of the Enhanced Lavallée-Hidiroglou Algorithm

Sample allocation in a stratified design is often based on fixed constraints in terms of costs or precision. Our goal for MWRTS was to minimise the sample size given target coefficients of variation for the various domains of interest. Since

nonresponse, undetected deaths, and misclassification will occur, sample sizes are usually inflated after they have been allocated to compensate for the nonresponse and the potential effect of misclassified units on the expected coefficients of variation. For the new MWRTS, we took into account the expected percentage of undetected dead units (units with zero sales) within the sample allocation and size stratification determination to have a more efficient sample. To achieve this we used a new enhanced version of the Lavallée-Hidiroglou (1988) algorithm developed by Ferland (2003). The enhanced algorithm incorporates the expected percentages of undetected dead units and their impact on the variance calculation to find optimal boundaries such that the sample size will be minimum for a fixed coefficient of variation. We defined the expected undetected death rate at 10% in the large take-some strata and 20% in the small take-some strata. These rates were determined based on our experience with the old MWRTS. Within the algorithm, the samples for the survey portion (above the take-none) are allocated proportionally to the square root of the new size measure.

As well, to compensate for nonresponse, we inflated the sample by 10%, the observed nonresponse rate in the old MWRTS. We also derived oversampling rates for industry misclassification based on the observed misclassification rate in the annual surveys that were also NAICS based. These rates range between 10% and 35% depending on the industry group. We could not use the old MWRTS to determine the expected rate of industry misclassification by NAICS since it is based on the Standard Industrial Classification (SIC).

3.1.3. Use of Sample Constraints

Expected fluctuations in the population through time are also an issue to consider when allocating the sample. For example, too small of a sample in a given stratum may lead to fluctuations in the estimates from month to month depending on the heterogeneity of the stratum and the number of influential units entering (births) and leaving (deaths) the sample. In order to reduce the potential fluctuations in the estimates, we used a minimum sample size of 8 and maximum weights ranging from 5 to 10 in the large take-some strata and from 10 to 30 in the small take-some strata. The different maximum weights were defined in specific industry groups based on subject matter expertise. These constraints are applied once the sample overallocation has been performed to compensate for nonresponse and misclassification. Although these constraints contribute to increase the sample size, they should help in maintaining the quality of the survey.

3.2. Population and Sample Update

MWRTS is a repeated survey with maximisation of monthly sample overlap. The sample is kept month after month and every month births are added to the sample and dead units (see below) are identified for future treatment. The advantage of repeated surveys with sample overlap is the production of more accurate trend estimates. Monthly sample updates with births are necessary to have representative level estimates every month. These objectives require balance between frequent frame and sample update to obtain more accurate levels and frame and sample stability to reduce undesirable fluctuations in the trend estimates that are not linked to movements in the economy.

The sample updates for the new MWRTS are performed using collocated sampling methodology implemented in the Generalised Sampling System application developed at Statistics Canada (Statistics Canada, 1996). MWRTS births, i.e., new clusters of establishment(s), are identified every month via the BR's latest universe. They are stratified according to the same criteria as the initial population. A sample of these births will be selected according to the sampling fraction of the stratum to which they belong and will be added to the monthly sample. Deaths occur on a monthly basis. A death can be a cluster of establishment(s) that have ceased their activities (out-of-business) or whose major activities are no longer in the retail or wholesale trade (out-of-scope). The status of these businesses is updated on the BR using administrative sources and survey feedback, including feedback from the MWRTS. Methods to treat dead units and misclassified units as part of the sample and population update procedures are described below.

3.2.1. Treatment of Dead Units

The BR regularly updates the live/dead status of the units; however, the status of MWRTS "in-sample" units is usually more up-to-date since units are contacted every month. Dead units can not be removed as soon as they are detected since MWRTS is a non-independent source of information that contributes to update the BR. Deaths identified by MWRTS will remain in the sample but will not be sent a questionnaire. They contribute to the derivation of the level and the variance estimates with a value of zero. To reduce the impact of the dead units on the coefficients of variation, the same techniques that were developed for the old survey that allow removal of some dead units in an unbiased manner, will be

used twice a year (Trépanier et al.,1998). Essentially, the same percentage of dead units is removed in the “in-sample” portion and the “out-of-sample” portion.

3.2.2. Treatment of Misclassified Units

Stratification variables are updated on a continual basis on the BR; however, the stratification information is not updated for ongoing units of the MWRTS population. For example, even if the geography has been updated for an ongoing unit, it will continue to be stratified in its original stratum. We decided to implement regular restratification exercises to reduce the potential impact of misclassified units on the variance estimates. Restrartification consists of the reallocation of units to the adequate stratum using the latest information available from the stratification variables (industry, geography, size). Three types of restratification exercises are being considered; monthly, annual, and every five years (Majkowski 2001).

Monthly size restratification will take place from take-none strata to take-all strata for legitimately large units. These units are identified through monthly monitoring of their size measure and are investigated by subject matter specialists. This process was implemented since legitimate large self-representing units should be included in the take-all stratum.

Annual size restratification may be performed on some sample units and out-of-sample units in an unbiased manner in specific geography and industry group. If deemed necessary, all the units within a certain geography and industry group could be restratified by size using an independent size measure. Restrartification will potentially reduce the variance but will also cause the estimates to fluctuate in the month that it is implemented. For this reason, we plan to apply the restratification during a specific month where other planned updates that cause fluctuations in the estimates are also taking place.

In conjunction with the previous two restratifications there should be a full-scale restratification every five years on all the units. Full-scale restratification would take into account the changes to industry groups, geographical regions and size measure.

3.3. Estimation

Despite all our efforts to obtain adequate classification information some misclassification will occur. This misclassification causes problems that need to be addressed at estimation until corrective methods, such as restratification, can take place. Misclassification by geography and industry will continue to be treated though domain estimation. Domain estimation techniques allow the production of unbiased estimates. Since misclassification by size has historically led to the most problematic cases, we tested different methods to reduce its impact at estimation.

3.3.1. Post-Stratification by Population Counts

We attempted to deal with misclassification by size using post-stratification by population counts to take advantage of independent frame updates that have occurred since the population was stratified. We decided not to investigate the use of geography and industry group updates to the frame as many of these are a result of survey feedback, and thus are not independent of our survey. We evaluated forming post-strata using size boundaries based on those used for stratification. Our study showed that there was no clear gain in using post-stratification based on population counts in the MWRTS context when compared to the use of the simple expansion estimator (Matthews, 2002). We decided to continue using the simple expansion estimator but to use an outlier detection and treatment method to treat problems that arise from size changes.

3.3.2. Outlier Detection and Treatment

The outlier detection and treatment strategy consists of four steps that focus on the primary output of the survey, monthly sales by industry group and geography. The first step identifies suspicious domains by comparing preliminary sales to forecasted sales (using time series methods) based on data from previous months. Domains for which large discrepancies exist between the preliminary and forecasted estimates (beyond a specified tolerance) are considered suspicious.

The second step consists in identifying influential units within the suspicious domain using an outlier detection and treatment module. Different methods to identify and treat outliers were tested and two were implemented for the new MWRTS: a modified Fuller “Test and Treat” method, described in Fuller (1991) and the Deflation factor Method (Matthews and Bérard, 2002). These two methods were retained for their general properties in terms of mean square error. We also decided to keep two methods since we observed that different units are sometimes identified by the two methods. Hence, we can provide different alternatives to the users. Any outlier treatment will induce some bias;

therefore, the decision to apply a treatment on one unit and not another relies heavily on subject matter knowledge of the unit and its economic environment.

Extensive analysis of the proposed outlier treatment is done in the third step. Since both geography and industry estimates are produced by the survey, the treatment of an outlier in a specific industry group will automatically affect the estimates of the geographic region to which it belongs (and vice-versa). A report is generated that includes the proposed treatments, and their impact on both the level (total monthly sales) and the trend (month-to-month change) estimates at the industry and geographic level. This report serves to assist subject matter experts in the choice of the units that should be treated and the type of treatment that should be applied.

The final treatment (fourth step) is determined by subject matter experts who select whether to implement the proposed treatment or some other treatment based on other sources of information. Once a treatment is determined, it is applied to the unit at the estimation stage via a correction factor that is applied to the reported value.

4. FUTURE WORK

The methodology for restratification will be finalised shortly. Our objective is to develop an unbiased method that will allow the selection of units to be restratified and will lead to a gain in estimate precision while disturbing trend estimates as little as possible.

5. ACKNOWLEDGEMENT

This work would not have been possible without the collaboration of many methodologists from Statistics Canada who directly contributed to the development of the new MWRTS: Michel Ferland, Susie Fortier, Mark Majkowski, Steve Matthews and Jean-Sébastien Provençal. I also would like to thank Serge Legault, Steven Thomas and especially Julie Trépanier for their helpful comments when writing this paper.

6. REFERENCES

- Ferland, M. (2003). Enhanced Lavallée-Hidiroglou algorithm. *Unpublished Manuscript, Ottawa: Statistics Canada.*
- Fuller, W. A. (1991). Simple estimators for the mean of skewed populations . *Statistica Sinica*, **1**, 137-158
- Hidiroglou, M. and Laniel, N. (2001). Sampling and estimation issues for annual and sub-annual Canadian business surveys *International Statistical Review*, **69**, 3, 487-504.
- Lavallée, P. and Hidiroglou, M.A. (1988). On the stratification of skewed population . *Survey Methodology*, **14**, 33-45.
- Majkowski M. (2001). Maintaining estimate quality and easing response burden in a subannual business survey. *Proceedings of the Business and Economic Statistics Section, American Statistical Association, on CD-ROM.*
- Matthews, S. (2002). Post-stratification studies for the redesigned Monthly Wholesale and Retail Trade Survey, *Unpublished Manuscript, Ottawa: Statistics Canada.*
- Matthews, S. and Bérard, H. (2002). The outlier detection and treatment strategy for statistics Canada's Monthly Wholesale and Retail Trade Survey . *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting*, 63-68.
- Parent, M.-N. and Simard, M. (2000). Sampling with a unified approach: the case of the Unified Enterprise Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting*, 229-233.
- Trépanier, J., Babyak, C., Marchand, I., Bissonnette, J. and St-Pierre, M. (1998). Enhancements to the Canadian Monthly Wholesale and Retail Trade Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 487-492.
- Statistics Canada (1996). *Generalized Sampling System (GSAM) User Guide*, Ottawa: Statistics Canada.