The Use of Administrative Data in Business Surveys: The Statistics Canada Experience

Martin Renaud, Richard Laroche[1]

Abstract

Over ten years ago, in an effort to reduce response burden in its monthly economic surveys, Statistics Canada introduced the use of administrative data, namely the Goods and Services Tax (GST) data, as part of its estimation process. The following four surveys adopted this new practice: Monthly Survey of Manufacturing, Monthly Wholesale Trade Survey, Monthly Retail Trade Survey, Monthly Survey of Food Services and Drinking Places. However, each survey decided on how to use the administrative data independently in a way that would best meet their needs. Since last year, an initiative has been launched to use the GST data as part of a ratio estimation strategy that is to be uniformly applied across all four surveys. This paper will start by presenting how each survey was initially using the administrative data before elaborating on the current harmonization effort in the use of the GST data. To conclude, some potential avenues which could be used in the future due to the increasing availability of administrative data will be discussed briefly.

Key Words: monthly surveys, Goods and Services Tax data, ratio estimation

1. Introduction

Statistics Canada has an extensive business survey program. Surveys in this program can be very different from one another. Some cover very large populations where more than 100,000 units can be found while others are censuses of population with less than 5 units. Topics are also very diversified ranging from agriculture surveys to environment surveys with transportation, energy, manufacturing, retail, wholesale, capital expenditures, research and development, construction, service industries, and many other topics in between. As well, some surveys are conducted on an annual basis while others occur at either a monthly or quarterly frequency.

Over the years, the number of business surveys has steadily increased in order to provide more statistics to government agencies, policy makers, and many other parties to help them in their planning and decision making process. Given the ever increasing number of surveys and the sometimes high frequency of contacts with respondents, the response burden has also increased dramatically in recent times. To provide some relief to its respondents, Statistics Canada started using administrative data, mainly tax data, in its business surveys program in the late 1990's.

Statistics Canada has an agreement with the Canada Revenue Agency (CRA) where the latter provides the former with the income tax return data for all Canadian businesses which have filed their income tax return. As well, the CRA transmits the amounts of the

[1] Martin Renaud, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6, email: martin.renaud@canada.ca. Richard Laroche, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6, email: richard.laroche2@canada.ca.

Goods and Services Tax (GST) collected by all businesses in Canada. At first, some tax data were used for direct replacement of survey data meaning that some questions were not asked on the survey questionnaire, their values being taken directly from the appropriate tax data files. With time, the use of tax data grew as it started to be used in the edit and imputation strategy for some surveys. Furthermore, tax data was used for validation and even in some sort of modelling exercises. Finally, tax data was introduced at the estimation stage in some surveys.

Sub-annual business surveys have been making extensive use of tax data for more than ten years now. The main reason behind the introduction of tax data was to reduce the response burden, especially for the mid-size businesses. Given that most sub-annual business surveys contact their respondents on a monthly basis and that units remain in the sample for approximately five years, one can see that being selected in one of those surveys is quite demanding.

In section 2 of this paper, we will present the sub-annual surveys which started using tax data in the early 2000's. The following section will explain how this auxiliary data was first used. Section 4 will describe the recent initiative to harmonize all monthly business surveys. Finally, section 5 will present results of implementing a ratio estimator for all monthly surveys covered in this paper. A short conclusion will summarize this paper and present some potential avenues for future use of administrative data.

2. Sub-annual surveys

Statistics Canada has been conducting a monthly retail trade survey, in a format that has evolved and improved through the years, since 1930. In 1947, a monthly survey of manufacturing was introduced. One year later, a monthly wholesale trade survey joined the business statistics programs. These three surveys are now part of Statistics Canada's mission critical program**.** In 1980, a monthly survey of food services and drinking places was added to the business surveys program. The following paragraphs present a short description of each of these four surveys and their sample design.

The Monthly Retail Trade Survey (MRTS) collects sales and the number of retail locations by province and territory from a sample of retailers. Sales estimates obtained from retailers are a key monthly indicator of consumer purchasing patterns in Canada. Furthermore, retail sales are an important component of the Gross Domestic Product, which measures Canada's production, and are part of many economic models used by public and private agencies. The Bank of Canada relies partly on monthly retail sales estimates when making decisions that influence interest rates. Businesses use retail sales estimates to track their own performance against industry averages and to prepare investment strategies. The target population consists of all statistical establishments on Statistics Canada's Business Register (BR) that are classified to the retail sector using the North American Industry Classification System (NAICS). The MRTS uses a stratified design with simple random sample selection in each stratum. The stratification is done by industrial groups and provinces/territories where each stratum is further stratified by size. The size strata consist of one take-all (census), at most two take-some (partially sampled) strata, and one take-none (none sampled) stratum. Take-none strata serve to reduce respondent burden by excluding the smaller businesses from the surveyed population. These businesses represent at most ten percent of total sales. Instead of sending

questionnaires to these businesses, the estimates are produced through the use of administrative data. In total, approximately 10,000 units are sampled from the more than 300,000 in-scope establishments.

The Monthly Wholesale Trade Survey (MWTS) provides information on the performance of the wholesale trade sector and is an important indicator of the health of the Canadian economy. In addition, the business community uses the data to analyse market performance. This survey presents estimates of monthly sales and inventory levels for wholesale merchants in Canada and in each province and territory. A variety of organizations, sector associations, and levels of government make use of the information. Governments are able to understand the role of wholesalers in the economy (5-6% of the Gross Domestic Product, depending on the year), which aid in the development of policies and tax incentives. The MWTS target population consists of all statistical establishments on the BR that are classified to the NAICS wholesale sector. Its design is identical to the MRTS but its sample size differs slightly as approximately 7,500 units are selected from a population of 100,000 in-scope establishments.

The Monthly Survey of Manufacturing (MSM) publishes statistical series for manufacturers – sales of goods manufactured, inventories, unfilled orders and new orders. The data collected by the MSM are used to analyze the Canadian economic situation and the short- and medium-term health of specific industries. They also serve as inputs to Canada's Gross Domestic Product. The information is used by both private and public sectors including Statistics Canada, federal and provincial governments, business and trade entities, international and domestic non-governmental organizations, consultants, the business press and private citizens. The MSM target population consists of all statistical establishments on the BR that are classified to the manufacturing sector, or more precisely which have a NAICS code beginning with 31, 32, or 33. Like its retail and wholesale trade counterparts, the MSM uses a stratified design with sample random sampling in each stratum. Again, strata are further stratified by size with one take-all at most three take-some, and one take-none strata. The MSM sample contains close to 12,000 establishments which were selected from a population of about 117,000 units (Laroche and Chen, 2015).

The Monthly Survey of Food Services and Drinking Places (MSFSDP) collects data on sales and the number of locations of restaurants, caterers, and drinking places. Estimates of the value of sales and the number of locations are produced by province and territory and by industry at the NAICS four-digit or six-digit level, but only the estimates of the value of sales are published. These data are used by federal and provincial governments, private associations and food service businesses for consulting, marketing and planning purposes. The provincial and federal governments use the information to estimate provincial taxation shares. The MSFSDP target population consists of all establishments on the BR which have a NAICS code beginning with 722. This survey has negotiated reporting arrangements with several major restaurant chains to collect data for all of its outlets through the head office. Thus the head office becomes the sampling unit and collection entity, rather than the individual outlets. This approach has the advantage of decreasing response burden. The sampling unit for the MSFSDP is the cluster of establishments. It is defined as in-scope establishments that belong to the same enterprise and have the same chain agreement status (within the enterprise, all establishments with a chain agreement are grouped into one cluster and those without a chain agreement are grouped into another cluster). The MSFSDP uses two different estimators and because of

that, its design has two different approaches to match these estimators. First, the traditional Horvitz-Thompson estimator is used in some strata defined by geography and NAICS. Those strata are further stratified by size with one take-all, some take-some (usually two), and one take-none strata. The other estimator that is used is the ratio estimator. For certain combinations of geography and NAICS, simple units (units present in only one province and only one NAICS) are separated from complex units (units present in more than one province and/or more than one NAICS). Simple units are then matched to a file of auxiliary data, in this case, tax data which is used in the ratio estimator. Close to 97,000 establishments are identified each month as being in-scope for the MSFSDP and the sample is composed of approximately 11,500 clusters of establishments (Laroche and Muenz, 2015).

3. Administrative data use, the beginning

In the late 1990's, Statistics Canada, like many statistical agencies around the world, had already started using administrative data in various ways for its business surveys. Even though the use of such data had numerous advantages, it also presented many challenges. For example, complete digital information was not available for all businesses in Canada and hence could not be used directly to replace survey data. On top of that, the program was also dealing with some conceptual, technical, operational, methodological, and legal issues. For more information with regards to the early days of tax data use in business surveys at Statistics Canada, the reader is directed to Smith, 2000. Because of those challenges, the use of tax data was quite limited. While it could be used in a few instances to directly replace survey data when all proper conditions were aligned, tax data was mostly used to confront and validate survey results. At other times, it was also used as part of the imputation strategy but in a fairly limited way.

Fast forward to the early 2000's when some of these early challenges had been resolved. Around that same time, the MSM, the MWTS, and the MRTS were starting to see a decline in their response rates, especially in their mid-size businesses. The situation was becoming so pressing that something had to be done before those surveys were due for a restratification. It's important to mention that at Statistics Canada, monthly surveys go through a restratification approximately every five years. In a nutshell, this process involves recalculating the strata boundaries and selecting a new sample. Of course, large units in the population are so important to the economy that they will always be part of the sample, even after a restratification. But the mid-size units are removed from the old sample and replaced in the new sample with randomly selected units from the population. Therefore, given that it was not time to restratify these surveys, something else had to be done to relieve the mid-size businesses of their burden. The idea was put forward to use tax data to model their survey data.

Studies were conducted for the three previously mentioned surveys to see how well tax data was correlated to survey data for the mid-size businesses. Not surprisingly, the correlation between their income on tax data and their total sales reported on survey data was very high. Based on this result, the following strategy was adopted (for more details, the reader is invited to consult Haziza and Yung, 2006). It was decided that some of the smaller mid-size businesses that had reported stable sales since joining the survey would not receive a questionnaire anymore. These units were commonly referred to as the S2 units. Instead, their survey data would be imputed with a value obtained from a regular

linear regression model between tax and survey data fitted to the businesses of similar size who did receive a questionnaire and responded to the survey. This second group of units was referred to as the S1 units. Overall, there were approximately 1,000 units in the S2 group for each survey. Although not methodologically perfect, this stop gap measure served its purpose and still provided very good quality data given the very strong relationship observed between tax and survey data. However, it was clear that this solution would have to be replaced eventually with a more sound methodological approach. In recent years, efforts were made to harmonize monthly surveys as much as possible. One aspect of this harmonization was the use of tax data at the estimation step through the implementation of a ratio estimator.

4. Harmonization of monthly surveys

The MSFSDP, MRTS, MWTS and MSM have a number of things in common: they are surveys whose main purpose is to measure the monthly sales of Canadian businesses; Statistics Canada's BR is used as the frame; a sample of new births is added each month to an initial sample that is used from month to month; after data collection, the data are calendarized, edited and imputed, if necessary; estimates are produced for a number of industries and provinces, and for a combination of these two dimensions; some estimates are seasonally adjusted; confidential cells are deleted; and six to eight weeks after the reference month, the data are published in Statistics Canada's official release bulletin, The Daily, and in the CANSIM (Canadian Socioeconomic Information Management System) database.

Despite their numerous similarities, the surveys were developed and, until recently, conducted independently of each other (except for the MRTS and MWTS, which have always shared the same systems and methodology). Therefore, different computer programs existed for executing a given procedure when only one program should suffice. As well, some methods differed slightly between surveys and could definitely be harmonized.

Various initiatives were undertaken or are currently in progress to further harmonize these monthly surveys.

4.1 First phase: Harmonization of the imputation and estimation systems

In 2007, a data quality assurance review conducted by Statistics Canada identified some concerns about monthly surveys. The MSM production system was deemed at risk because of its age, its complexity and the multiple manual adjustments that had been made over the years. As well, the systems used for the MRTS and MWTS lacked key analysis-related functionality (Andrews et al. 2011).

In response to these results, the Industry Statistics Branch Monthly Survey Systems Integration Project (ISBMSSIP) was created. The purpose of this major project was to harmonize the imputation and estimation systems of the MSFSDP, MRTS, MWTS and MSM.

In December 2011, the MSFSDP became the first of the four monthly surveys to use the new system. It was followed by the MSM in September 2012 and then the MRTS and MWTS in April 2014.

4.2 Second phase: Harmonization of sampling methods

The second phase of the ISBMSSIP started in 2013. It involved harmonizing the sampling methodology of the monthly surveys and building a single sampling system that could be used by all the surveys.

For this project, the existing sampling methodologies were carefully examined and documented (Blanchard 2013). Best practices were identified and recommendations were made to the project's steering committee (Blanchard 2014). The recommendations included harmonizing the criteria required to extract the target population from the Business Register, including certain types of units in the frame that had until then been excluded from the target population, defining the sampling unit, determining the take-none portion, deriving the size measure, reducing the number of strata for the MSM, regularly using an unbiased process for removing dead units from the frame, determining the process for a mini-restratification and producing diagnostics.

Another recommendation put forward was to use a ratio estimator and drop the S1/S2 strategy described in Section 3. This is a major change that would affect the way in which administrative data are used, among other things.

5. Ratio estimator

5.1 Background

Using a ratio estimator in monthly surveys is not a new idea. Studies on the topic have been conducted since 2000 (Marchand et al. 2000). At that time, GST data had been available for only a few years for use as auxiliary variables in surveys, and some concepts underlying these data were not completely understood or documented. In addition, the system for processing these data was not as well-developed as the one that is used now.

Over the last 15 years, the methodology for processing GST data has continually improved, in terms of calendarization, imputation and allocation of business data to establishments. Everything is well documented and data quality is now excellent.

Therefore, using GST data through a ratio estimator is now a promising avenue.

5.2 Definition

To use ratio estimation, the two variables y_i and x_i must be available for each sample unit (y_i is measured during the survey, and x_i normally comes from administrative data). In addition, the control totals of the variable x must also be known (these totals may be at the Canada, provincial or industry level, for instance).

In a population U, let $t_y = \sum_U y_i$ and $t_x = \sum_U x_i$ be the totals of variables y and x, respectively, where y is the study variable and x is an auxiliary variable.

The ratio estimator of t_y can then be defined as

$$\hat{t}_{yr} = \hat{t}_y \cdot \frac{t_x}{\hat{t}_x}$$

where $\hat{t}_y$ and $\hat{t}_x$ are estimates of t_y and t_x, respectively, obtained using the sample.

The ratio estimator takes advantage of the correlation that may exist between x and y. The greater the correlation, the more efficient the estimator will be in terms of variance (Lohr 1999).

Various studies have been done to examine the correlation between the variable of interest (monthly sales) and the auxiliary variable (revenues from the GST file) for each of the four monthly surveys. For the MSM, Yung et al. (2004) showed that, once outliers have been removed, the correlation between sales and revenues in the GST file is 0.91. For the MSFSDP, Pritchard and Tardif (2006) showed that the correlation between sales and revenues in the GST file was 0.94 in February 2006. As well, Majkowski and Trépanier (2005) showed that, once outliers had been removed, the correlation between sales and revenues in the GST file was 0.92 for the MWTS and 0.96 for the MRTS.

5.3 Empirical results

5.3.1 Auxiliary variable

The auxiliary variable used in the simulations comes from the GST data received from the CRA. The file from the CRA does not contain data for all enterprises in the target population of the monthly surveys. In Canada, businesses with less than $30,000 in annual revenue don't have to collect GST from their customers and do not appear in the CRA file.

The following table gives the percentage of enterprises in the CRA file along with their contribution to the total revenue.

Table 5.1: Enterprises in the CRA file

Survey	Number of enterprises in the population[1]	Percentage of enterprises in the population[1] found in the CRA file	Percentage of total revenue[2] coming from enterprises in the CRA file
MSFSDP	93,351	71.9%	95.4%
MRTS	216,336	62.5%	97.7%
MSM	103,547	62.9%	98.4%
MWTS	91,667	69.9%	98.7%

1. As of January 1st, 2016

2. Revenue coming from Statistics Canada's Business Register (available for all units in the population)

Even if the auxiliary variable is missing for a third of the enterprises, those enterprises are generally small and do not contribute much to the total revenue. For those units, the auxiliary variable was imputed with a model using the revenue from Statistics Canada's Business Register.

5.3.2 Calibration groups

The calibration groups form a mutually exclusive and exhaustive partition of the population within which there exists a good relationship between the auxiliary variable and the variable of interest; they should also ideally be as close as possible to the domains for which estimates are produced in order to avoid large variance estimates (Marchand et al., 2001).

The take-all units (the ones with a selection probability of one) are part of the same calibration group, the goal being to keep their weight always equal to one. The take-all units represent 42% of the total estimate for the MSFSDP, 58% of the total estimate of the MRTS, 72% of the total estimate for the MWTS and 74% of the total estimate of the MSM.

For the remaining units (the ones being part of the take-some and take-none strata), the calibration groups were defined at the industry level (NAICS code at the 3, 4 or 5-digit level) for the MSM, the MRTS and the MWTS and at the industry and province level for the MSFSDP to make them as close as possible to the domains of interest as previously stated. This way of defining the calibration groups is not final and could change depending on the results observed in future studies.

5.3.3 Cut-off sampling

All four monthly surveys have a take-none stratum in which each unit has a zero inclusion probability. This is common in business surveys where a large number of small businesses have a small contribution to the total of the variable of interest. The deliberate exclusion of part of the target population from sample selection is called cut-off sampling (Särndal et al., 1992). A nice feature of the ratio estimator proposed in section 5.2 is that it can be used to get an estimate for the whole population.

In a population U, let U_c denote the cut-off portion of the population and let U_0 be the rest of the population, from which we assume that a probability sample is selected in the normal way and where π_i denotes the selection probability of unit i. Let x be the auxiliary variable and y be the variable of interest. Let $\hat{R}_{U_0} = \frac{\sum_{s_0} y_i/\pi_i}{\sum_{s_0} x_i/\pi_i}$ be the estimator of R_{U_0} based on the probability sample from U_0. Assuming that $R_{U_0} = R_U = \sum_U y_i / \sum_U x_i$, then $\hat{R}_{U_0}$ can serve to estimate R_U as well, and by ratio adjustment we arrive at $\hat{t}_{cut} = (\sum_U x_i)\hat{R}_{U_0}$ as an estimator of $t = \sum_U y_i$ (Sarndal et al., 1992).

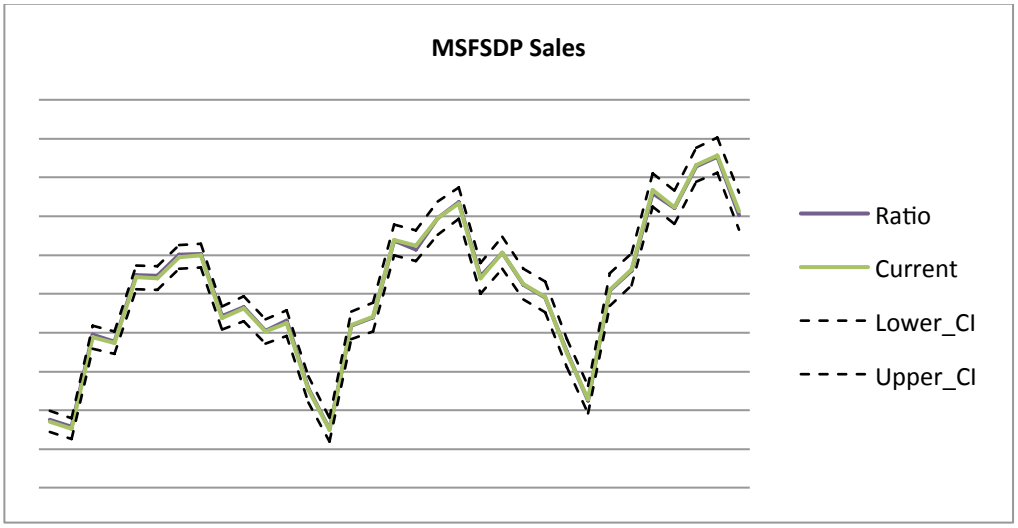
5.3.4 Results

Simulations were run for all four surveys. The current estimates were compared with the ones obtained using the ratio estimator as defined in section 5.2. Comparisons were made for approximately 24 months. The level of the estimates, as well as the month over month trends, were looked at for many domains. Sales, as well as inventories (for the MWTS and the MSM) were analyzed.

The following graphs present results observed for some selected domains and the confidence interval displayed is for the current estimator:

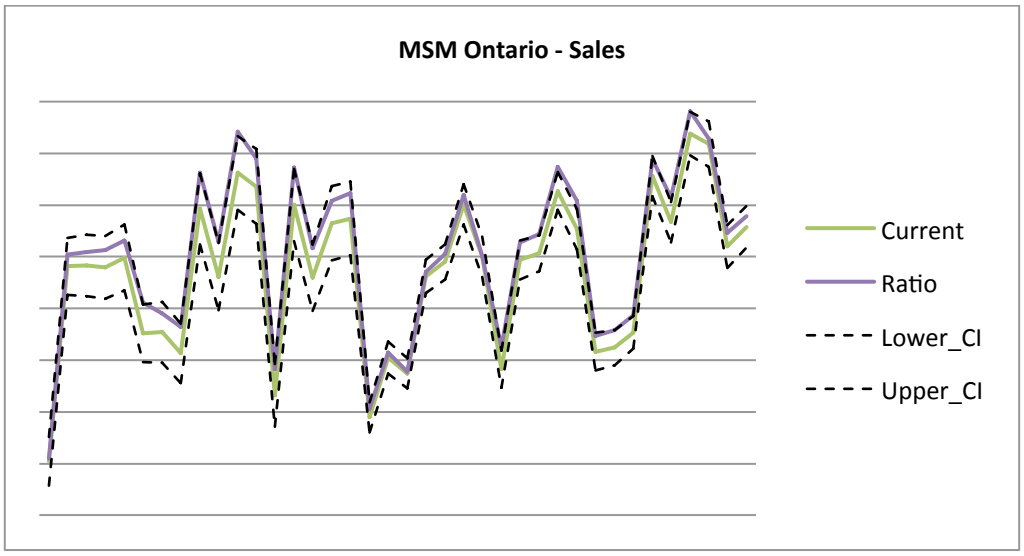
- The current estimator and the ratio estimator give very similar results for most NAICS and provinces in the MSFSDP

Graph 5.1: Sales for MSFSDP



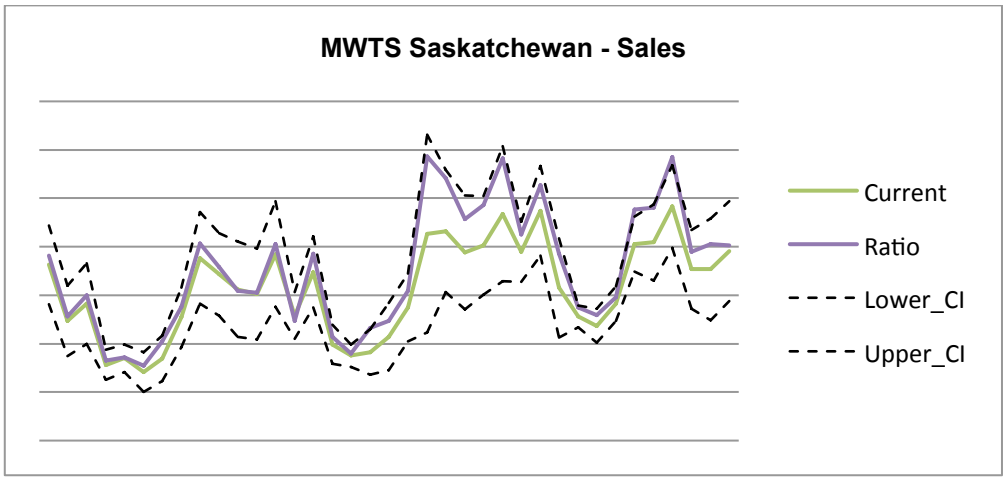
- For the MRTS, the MWTS and the MSM, the levels of estimates are often different ; however, the month over month trends are very similar

Graph 5.2: Sales for Ontario (MSM)



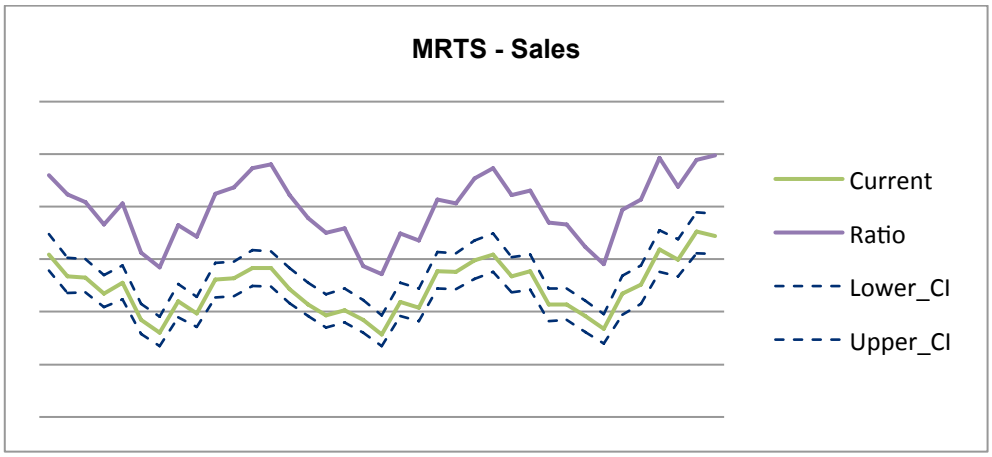
- Differences between the current estimator and the ratio estimator can sometimes be explained by the growth coming from the out-of-sample units

Graph 5.3: Sales for Saskatchewan (MWTS)



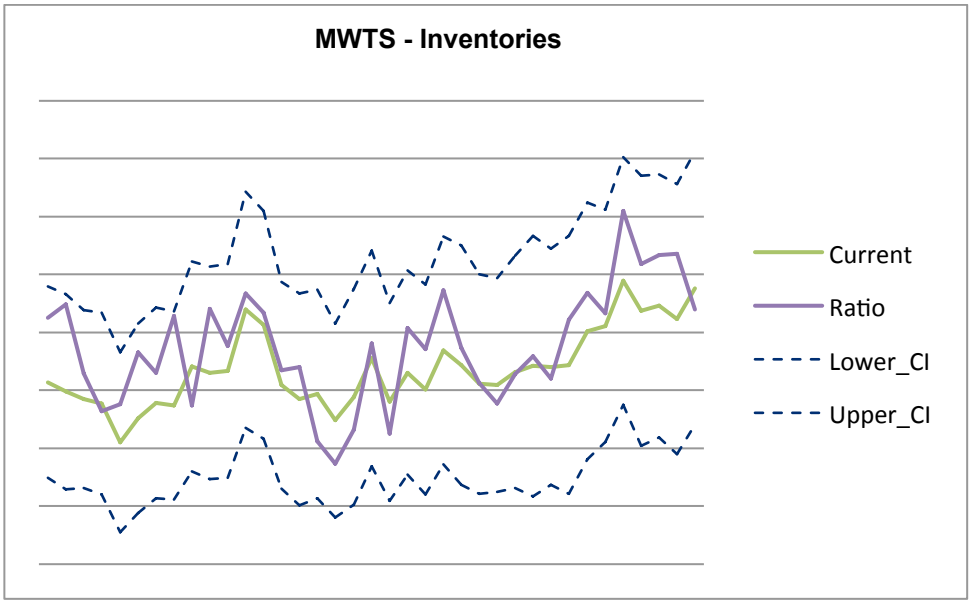
- Large differences between the current estimator and the ratio estimator can be explained by the fact that the current sample is not representative anymore

Graph 5.4: Sales for MRTS



- When looking at the inventories, the current estimator and the ratio estimator are often different for both the levels and the trends, but the ratio estimator is within the current confidence interval

Graph 5.5: Inventories for MWTS



5.4 Advantages and disadvantages of the ratio estimator

There are a number of advantages to using the ratio estimator:

- The method is available in Statistics Canada's generalized system, G-EST.
- The variance accounts for the take-none portion (the take-none portion is currently treated as a census).
- Administrative data are used for all population units.
- S2 subsamples (currently used for the MSM, MRTS and MWTS) are no longer needed.
- Monthly surveys are more harmonized (estimation method, use of GST file).
- The sampling and estimation process for the MSFSDP is simpler than the current method which combines two different types of designs and estimators.
- Under certain conditions, the ratio estimator is more precise than the Horvitz-Thompson estimator.

However, some disadvantages should be noted:

- On rare occasions, some take-all units may have weights other than 1 (in calibration groups with only one take-all stratum and one take-none stratum).
- The estimation process will become more dependent on potential problems in processing the GST file.
- For some domains, the current level of estimates may change.
- Revenues in the GST file need to be imputed for some units, which will affect the quality of the estimates (in a way that is currently unknown).
- A preprocessing step for the GST file must be developed, to identify outliers (for example cases in which cents are mistaken for dollars).
- Changes are required to operationalize the ratio estimator.
- The ratio estimator may not be effective for variables that are less closely correlated with sales.
- Weights may vary from one month to the next.

There are numerous advantages to using the ratio estimator, some of which align directly with the general principles of Statistics Canada's Corporate Business Architecture (CBA). One of the CBA's fundamental principles is to maximize reuse by minimizing the number of separate computer systems. In the current situation at Statistics Canada, the advantages of using the ratio estimator in monthly surveys outweigh the disadvantages.

6. Conclusion

As presented in this paper, one can see that the use of tax data has gradually increased in sub-annual business surveys at Statistics Canada. What started out as very basic methods evolved through the years, culminating in the recent adoption of the ratio estimator. Based on our current state of knowledge, what can be envisioned for the future? For one thing, tax data can and will be used more extensively in business surveys (Cloutier, 2010). Certain programs will use tax data instead of traditional survey data. But tax data is only part of the solution as more and more administrative data sources become

available. For example, scanner data could be used if widely available. There are already initiatives in place to see how scanner data could be used in the calculation of the Consumer Price Index. This type of data could also be very useful in the retail industry for a survey like the Retail Commodities Survey. Furthermore, some agriculture programs are developing methods to incorporate satellite data in the estimation of different crops around the country. With big data being on everyone's lips, this is a very exciting time to consider all sorts of different sources of administrative data. The main challenge will be to develop methods on how to properly use them in surveys (not only business ones) and once this is done, making sure that we have the digital power to store them and process them in a timely fashion.

Acknowledgements

The authors would like to thank Marie-Claude Duval and Margaret Wu for their valuable comments in improving the quality of this article as well as Wesley Yung for presenting it at the conference.

References

Andrews, J., Brisebois, F., Delahousse, I., Dochitoiu, C., Lachance, M., Philips, R. and Pursey, S. (2011), "Harmonizing methodologies through a system integration project: Challenges and lessons learned", Proceedings of Statistics Canada Symposium 2011.

Blanchard, T. (2013), ISB Front End Sampling Methodology Review, Statistics Canada, internal document.

Blanchard, T. (2014), ISB Front End Methodological Analysis, Statistics Canada, internal document.

Cloutier, M. (2010), A Strategic Vision for the Use of Administrative Data at Statistics Canada, Proceedings of Statistics Canada Symposium 2010.

Haziza, D., Yung, W. (2006), Tax Replacement Strategy in Business Surveys, Statistics Canada, internal document presented at the Advisory Committee on Statistical Methods

Laroche, R., Chen, S.X. (2015), Methodology of the Monthly Survey of Manufacturing, Statistics Canada, internal document.

Laroche, R., Muenz, J. (2015), Methodology of the Monthly Survey of Food Services and Drinking Places, Statistics Canada, internal document.

Lohr, S. (1999), Sampling: Design and Analysis, Duxbury Press.

Majkowski, M., Trépanier, J. (2005), Use of GST Data in the Monthly Wholesale and Retail Trade Survey, Statistics Canada, Internal document presented at the Advisory Committee on Statistical Methods.

Marchand, I., Gagnon, F. and Trépanier, J. (2000), "A study of calibration estimation using administrative data for the Canadian Monthly Wholesale and Retail Trade Survey", International Conference on Establishments Survey.

Pritchard, Z., Tardif, C. (2006), The Use of the Goods and Services Tax By the Monthly Restaurants, Caterers and Taverns Survey, SSC Annual Meeting.

Särndal, C., Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling, Springer.

Smith, P. (2000), Challenges Involved in Using Tax Data for Statistical Purposes, Statistics Canada, internal document.

Yung, W., Cook, K., Thomas, S. (2004), Use of GST Data by the Monthly Survey of Manufacturing, ASA Section on Survey Research Methods.