# INTEGRATING QUALITY INDICATORS IN COMPLEX SURVEYS

James Brennan[1], Jack Lothian[2] and Pierre Daoust[3]

## ABSTRACT

The Quarterly Survey of Financial Statements (QFS) is Statistics Canada's principal source of current financial information concerning the Canadian-private-incorporated business sector. This information is used extensively by governments, banks, and other institutions. A significant portion of the QFS data is imputed from previous responses or administrative data and it is increasingly important to determine the impact of this imputed data on the quality of the QFS estimates. This paper examines a strategy for estimating the combined variance of the QFS sampling and imputation by using Statistics Canada's System for Estimation of Variance due to Non-response and Imputation (SEVANI).

KEY WORDS: Estimation , Imputation, Variance.

## RÉSUMÉ

Le Relevé trimestriel des états financiers (RTEF) est la principale source actuelle de renseignements financiers de Statistique Canada en ce qui concerne le secteur canadien des entreprises privées constituées en société. Ces renseignements sont utilisés abondamment par les gouvernements, les institutions financières et d'autres établissements institutionnels. Une portion considérable des données du RTEF est imputée à partir de réponses antérieures ou de données administratives, et il s'avère de plus en plus important de déterminer la conséquence de ces données imputées sur la qualité des estimations du RTEF. Cet article examine une stratégie pour estimer la variance combinée de l'échantillonnage du RTEF et des stratégies d'imputation au moyen du Système d'estimation de la variance due à la non-réponse et à l'imputation (SEVANI) de Statistique Canada.

MOTS CLÉS : Estimation; imputation; variance.

## 1. INTRODUCTION

### 1.1 Survey background

The Quarterly Survey of Financial Statements (QFS) is an enterprise-based quarterly survey of the corporate sector that obtains information on corporate income statements and balance sheets, measures financial position and performance in Canada and flow of funds (net borrowing and savings) between economic sectors. It is a critical input to the System of National Accounts and is thus a major input into the production of the quarterly Gross Domestic Product estimates.

The sampling strategy for the QFS involves stratification by industry based on the North American Industrial Classification System (NAICS) for 80 industry aggregations and further size stratification based on assets and revenue. There can be up to 4 size strata in each industry aggregation: a take-all stratum (TA) two take-some strata (TS1 and TS2) and a take-none stratum (TN). Industries may have two, one or no take-some strata depending on the distribution of enterprises and the desired quality of estimates. The surveyed population of roughly 21000 units is made up of the enterprises in the TA, TS1 and TS2 strata. The core sampling design consists of taking a census in the TA stratum and a simple random sample within the TS1 and TS2 strata, which is usually selected in the first quarter of a calendar year. A small number of births are selected by Bernoulli sampling for the quarters between core samples. The sampling strategy varies between industries with some industries having a new sample selected each year and some getting a new sample after two or more years depending on the observed characteristics of the sample and population. The TN population released variables are derived based on annual tax data and quarterly movements estimated using the response data of the

[1] James Brennan, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, James.Brennan@statcan.gc.ca
[2] Jack Lothian, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, Jack.Lothian@statcan.gc.ca
[3] Pierre Daoust, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, Pierre.Daoust@statcan.gc.ca

surveyed population. Calibration estimation is used to adjust to population totals for the quarter in question with calibration adjustments typically being less than 1% of the estimate.

The QFS has a complex imputation process which is comprised of several different imputation strategies. The main form of imputation is historic trend imputation which is based on the trend observed between the current and previous quarters for respondents in a particular industry aggregation. Additionally, the imputation process uses donor imputation, manual imputation and, since it is a financial survey, there are deterministic balancing edits that are done to ensure individual records balance in terms of accounting formulas.

### 1.2 Total variance estimation

The QFS currently uses what is commonly referred to as naïve variance estimation where imputed values are treated as responses and variance is based solely on sampling. While the quality of imputation is considered to be quite good, especially for the main variables of interest, it is known that naïve variance estimation tends to underestimate the variance of the survey estimates. Variance that takes into account non-response and imputation as well as sampling will hereafter be referred to as the total variance. Development of theory and supporting software over time has made it possible to obtain total variance estimates for several types of imputation. However, it is important to determine if the estimation of total variance is practical in terms the QFS production, imputation process and the evaluation tools available.

Despite the complexity of the QFS imputation process, historic trend imputation dominates; therefore it was decided to conduct this study by treating all non-response cases as historically imputed. It was assumed that such an approach would approximate well the total variance of QFS estimates when non-response and imputation were taken into account.

The historic trend or ratio imputation is itself somewhat complicated. In practice trends are determined for the 80 industry aggregations based on NAICS codes. The estimate of the ratio $\hat{R}$ within an industry aggregation is calculated using non outlier respondents ($A$) to both the current ($t$) and prior ($t$-$1$) quarters $\hat{R} = \dfrac{\sum_{k \in A} x_k^t}{\sum_{k \in A} x_k^{t-1}} = \dfrac{\bar{x}^t}{\bar{x}^{t-1}}$.

For non-respondents this ratio would then be applied to the previous quarter value to impute the current quarter value $x_{+k}^t = \hat{R} x_k^{t-1}$, where the subscript + denotes an imputed value. A further complication of the imputation process is that the ratio is calculated for only three main variables: assets, revenue and expenses. The ratio for one of three main variables is then applied to a subset of variables $y_{+k}^t = \hat{R} y_k^{t-1}$.

The application of the ratio of one variable to another is done for efficiency and to balance records approximately in terms of accounting formulas and therefore minimize balancing edits required after imputation. This form of imputation is potentially biased and makes the task of estimating the total variance more difficult as discussed in section 2.2.

In practice, the determination of the trends is not fully automated as subject matter experts are involved in a manual determination of outliers and may determine the trend based on industry study and analysis when there are few respondents available. For this study, the process of calculating trends was fully automated, with trends determined for 24 industry aggregations, which is a higher level of aggregation than the 80 aggregations used for production. The higher level industry aggregation provided enough respondents to have automated outlier detection and to estimate the trends more robustly. Outlier detection for the study was done using the Hidiroglou-Berthelot (1986) historical trend method as implemented in Statistics Canada's Banff[4] system (Banff Support Team 2008). This fully automated imputation approach is thought to give comparable results to what is done in practice and allows for a reasonable estimation of total variance for the survey. Automation of the imputation process for the study had the added benefit of indicating some of the practical issues that would be involved with implementing changes in actual production.

4 Banff is a generalized system that offers methods of editing and imputing survey data in the form of SAS procedures.

## 2. LITERATURE REVIEW and METHODOLOGY

### 2.1 Key Approaches

Total variance estimation has been studied extensively over the preceding 20 years with several key papers coming out in the 1990s. While the theory has been around for a while, implementation necessarily lags behind as it takes time to adapt basic theory to the specific practical problems posed by complex surveys as well as to develop and test programs that are capable of use in production. Key references are Särndal (1992) and Shao and Steel (1999) which give different approaches to the problem in terms of breakdown of variance.

### 2.2 SEVANI

Statistics Canada methodologists have developed a system of SAS macros known as the System for Estimation of Variance due to Non-response and Imputation (SEVANI). The framework for SEVANI is based on Särndal (1992) and has been developed and adapted to many of the most common types of imputation done at Statistics Canada, the methodology being described by Beaumont (2010). SEVANI is the principal tool employed in this study and it uses the following breakdown of total error into sampling error and non-response error

$$\hat{\Theta}_I - \Theta = (\hat{\Theta} - \Theta) + (\hat{\Theta}_I - \hat{\Theta}). \quad (1)$$

In the above equation $\hat{\Theta}_I$ is an imputed estimate of $\Theta$ the true parameter value and $\hat{\Theta}$ is the sampling estimate of $\Theta$. Squaring both sides and taking the expectation with respect to the imputation model($m$), the sampling design($p$), the non-response model($q$) and the non-response mechanism(*) gives:

$$E_{mpq*}(\hat{\Theta}_I - \Theta)^2 = E_{mpq*}(\hat{\Theta} - \Theta)^2 + E_{mpq*}(\hat{\Theta}_I - \hat{\Theta})^2 + 2E_{mpq*}(\hat{\Theta} - \Theta)(\hat{\Theta}_I - \hat{\Theta}) \quad (2)$$

$$\approx E_m \operatorname{var}_p(\hat{\Theta}) + E_{pq}E_{m*}\left[(\hat{\Theta}_I - \hat{\Theta})^2 \mid s, s_r\right] + 2E_{pq}E_{m*}\left[(\hat{\Theta} - \Theta)(\hat{\Theta}_I - \hat{\Theta}) \mid s, s_r\right]. \quad (3)$$

In the above equations, and those that follow, $s$ denotes the sample and $s_r$ denotes the sample respondents. The approximation in (3) is due to $E_p(\hat{\Theta} - \Theta) \approx 0$ since the two main estimation methods in SEVANI are Horvitz-Thompson and calibration estimation which are unbiased or approximately unbiased, at least asymptotically. The first term on the right hand side $V_{samp} = E_m \operatorname{var}_p(\hat{\Theta})$ is the sampling variance, the second $V_{NR} = E_{pq}E_{m*}\left[(\hat{\Theta}_I - \hat{\Theta})^2 \mid s, s_r\right]$ is the non-response variance and the third $V_{mix} = 2E_{pq}E_{m*}\left[(\hat{\Theta} - \Theta)(\hat{\Theta}_I - \hat{\Theta}) \mid s, s_r\right]$ is a mixed component. Note that the sampling variance $V_{samp}$ is not equivalent to the naïve variance estimator as $V_{samp}$ is dependent on the imputation model.

In SEVANI, ratio estimation is a special case of deterministic linear regression imputation with the following model $x_k^t = \beta x_k^{t-1} + \varepsilon_k$ where $V_m(\varepsilon_k) = \sigma^2 x_k$. For non-respondents the imputed value is $x_{ik}^t = \hat{\beta} x_k^{t-1}$ where $\hat{\beta} = \hat{R}$. SEVANI assumes the imputation method is unbiased, i.e. $E_{m*}\left[(\hat{\Theta}_I - \hat{\Theta}) \mid s, s_r\right] = 0$, however, this may not be the case when we impute using the ratio or trend for another variable unless it happens that the trend for the imputed variable is the same as the trend of the variable used for imputation. For assets, revenue and expenses the expected value is $E_{m*}(x_{+k}^t) = E_{m*}(x_k^{t-1}) \bar{x}^t / \bar{x}^{t-1} = \bar{x}^t$ therefore $E_{m*}(x_{+k}^t - x_k^t) = 0$ with an unbiased imputation model whereas for any variable other than assets, revenue and expenses the expected value is $E_{m*}(y_{+k}^t) = E_{m*}(y_k^{t-1}) \bar{x}^t / \bar{x}^{t-1}$ therefore $E_{m*}(y_{+k}^t - y_k^t) \neq 0$.

The imputation strategy of using the trend of one variable for imputing another does not correspond to any of the common imputation methods available in SEVANI so the following approximation was attempted for such variables. The trends

that were calculated for one of assets, revenue or expenses were supplied to SEVANI and the ratio model was used to estimate the variance. Note that SEVANI includes a squared bias term in the non-response component of the variance since $E_{m*}\left[(\hat{\Theta}_I - \hat{\Theta})^2 \mid s, s_r\right]$ can be expressed as $\text{var}_{m*}\left[(\hat{\Theta}_I - \hat{\Theta}) \mid s, s_r\right] + \left[E_{m*}\left[(\hat{\Theta}_I - \hat{\Theta}) \mid s, s_r\right]\right]^2$.

The inclusion of the bias term can have a large impact on the reported variance when the actual trend of the variable and the trend applied for imputation are quite different thus violating the model assumption of unbiased imputation.

## 3. DISCUSSION of RESULTS

The results of this study using SEVANI show that the automated imputation strategy for assets, revenue and expenses is good in terms of precision, as there is, in most industry groups, only a small increase in CVs when taking into account non-response and imputation. In most cases the small increase in CV does not translate into a change of quality indicator.

The QFS reports the quality of estimates based on a letter that corresponds to a range of CVs as outlined in Table 1.

**Table 1- QFS quality indicators**

| Excellent- A | Very Good- B | Good-C | Acceptable-D |
|---|---|---|---|
| 0.00-4.99% | 5.00-9.99% | 10.00-14.99% | 15.00-24.99% |

Table 2 and 3 are examples of the change in the quality indicators for assets and revenue respectively for a typical quarter. For assets although there is a slightly higher CV in general it is not usually enough to change the quality indicator. For revenue there is in general a small increase in CV and in a couple of industries the quality indicators change. When there is a change in quality indicator, it is usually going to the next higher percentage range as shown here but, occasionally the change could be more than one percentage range.

**Table 2- Cross tabulation of quality indicators for assets for a typical quarter naïve approach vs. evaluation by SEVANI**

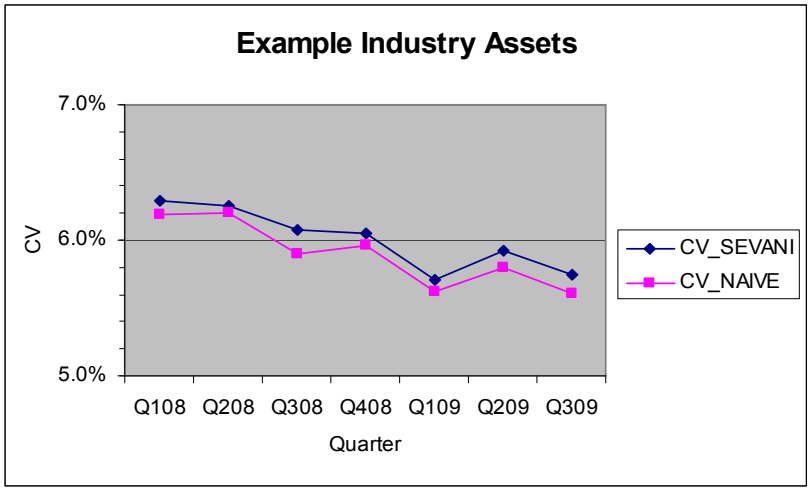
| **Assets** | **CV range SEVANI** | | | |
|---|---|---|---|---|
| **Naïve CV range** | **excellent** | **very good** | **good** | **acceptable** |
| **excellent** | **19** | **0** | **0** | **0** |
| **very good** | | **5** | **0** | **0** |
| **good** | | | **0** | **0** |
| **acceptable** | | | | **0** |

**Table 3- Cross tabulation of quality indicators for revenue for a typical quarter naïve approach vs. evaluation by SEVANI**

| **Revenue** | **CV range SEVANI** | | | |
|---|---|---|---|---|
| **Naïve CV range** | **excellent** | **very good** | **good** | **acceptable** |
| **excellent** | **17** | **1** | **0** | **0** |
| **very good** | | **5** | **1** | **0** |
| **good** | | | **0** | **0** |
| **acceptable** | | | | **0** |

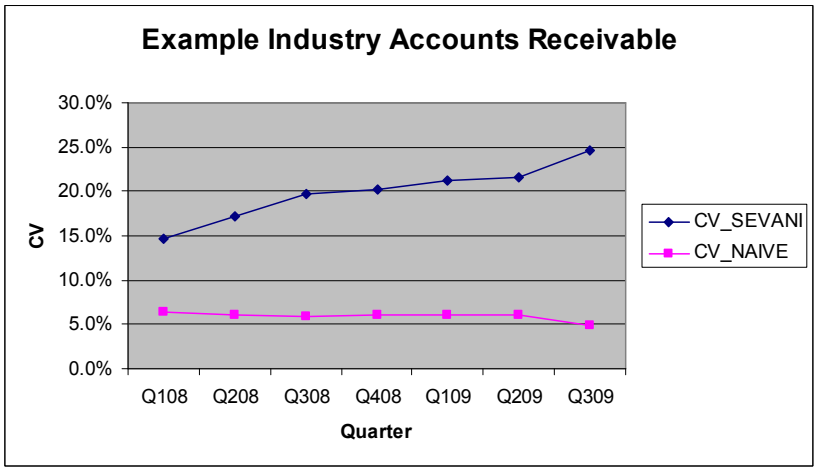
Looking at assets CVs over time for an example industry in Figure 1 it is clearer that there is typically a slightly higher CV based on the total variance than that produced by the naïve approach while the difference is fairly stable over time.

**Figure 1- Assets CVs over time**




For variables other than assets, revenue and expenses there was in general a larger gap between the naïve CV and the CV taking into account non-response and imputation and the difference was not stable over time. However, because of the approximation tried in SEVANI for other variables it was not clear if the approximation used was adequate or if the CVs were plausible. It was suspected that in many cases the CVs reported by using the approximation were too large. Figure 2 shows CVs for an example industry, for the variable "accounts receivable", where imputation is based on the trend of assets and the CVs based on the approximation in SEVANI show a larger increase over the naïve approach CV in comparison to the increase seen for assets.

**Figure 2- Accounts Receivable CVs over time**




## 4. FUTURE

There are many interesting topics for study in terms of the QFS imputation and variance estimation. Future work may include testing possible changes to the imputation process that allow us to use more of Statistics Canada's generalised systems. A study which is in progress will include variance estimation by a without replacement bootstrap described in Shao and Sitter (1996). SEVANI and the bootstrap will be used to obtain variance and CV estimates for samples from a simulated population and compared to variance and CV estimates obtained by Monte-Carlo methods. Another interesting area of study would be looking at the potential bias of different imputation strategies for the survey.

## 5. CONCLUSIONS

For the three key variables of interest, assets, revenue and expenses, where the ratio for each variable is applied to the previous quarter value to impute, the quality of the automated imputation of this study is good, in terms of precision, and variance can be properly evaluated by SEVANI. The CVs using the total variance are usually slightly larger than the CV using the naïve approach; however, it is usually not enough to change the quality indicator which corresponds to a range of CVs.

For other variables, where the strategy is to impute based on the ratio of another variable, there is no corresponding imputation model in SEVANI and the resulting CVs of the approximation outlined in section 2.2 may be too high. While it looks possible to use SEVANI for evaluating total variance for such variables, more testing is needed to see if the results of imputation model approximations are reasonable.

There is the potential that more of the QFS imputation process could be automated, as was done using Banff for outlier detection in this study. Adaptation of the imputation strategy for all variables to methods currently available in SEVANI would allow evaluation of total variance but, it is not clear if possible changes to the imputation will be practical in the context of the survey. More work is needed to determine the benefits and drawbacks of changes to the variance estimation and to the imputation process itself.

## ACKNOWLEDGEMENTS

The authors would like to thank Joël Bissonnette and Jean-François Beaumont for their assistance during this project and to thank Danielle Lebrasseur, Joseph Duggan and Danielle Lafontaine-Sorgo for their helpful comments which contributed to improving this paper.

## REFERENCES

Banff Support Team. (2008). "Functional Description of the Banff System for Edit and Imputation". *Statistics Canada Technical Report.*

Beaumont, Jean-François (2010). "SEVANI v2.3 - Methodology Guide". *Technical report, Methodology Branch, Statistics Canada.*

Deville, Jean-Claude and Särndal, Carl-Erik (1992) "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, Vol.87, No. 418 June, pp. 376-382.

Hidiroglou, M.A. and Berthelot, J.M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". *Survey Methodology*, No.12 June, pp. 73-83.

Särndal, Carl-Erik (1992). "Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used". *Survey Methodology*, Vol. 18, No. 2 December, pp. 241-252.

Shao, Jun and Sitter, Randy R. (1996) "Bootstrap for Imputed Survey Data" *Journal of the American Statistical Association,* Vol.91, No.435, September, pp. 1278-1287

Shao, Jun and Steel, Philip (1999) "Variance estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions". *Journal of the American Statistical Association*, Vol. 94 No.445 March, pp. 254-265.