Using Quality Indicators to Manage Collection and Editing in Business Surveys

Lingyun Zhu, Serge Godbout
Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON, K1A 0T6, Canada

Abstract

Statistics Canada recently launched the Integrated Business Statistics Program (IBSP) project with the purpose of redesigning and integrating several annual and sub-annual business surveys. The objectives of the IBSP include improving efficiency and timeliness while ensuring high quality outputs. As part of the IBSP, a processing strategy that combines active collection management, editing, imputation, estimation and analysis is currently being defined. This strategy consists of periodically producing estimates and quality indicators based on available data, both reported and imputed. Quality indicators will be used to evaluate the level of quality achieved at different points of the collection process, to identify influential records requiring follow-up or micro-editing, and to serve as criteria to stop collection.

In this presentation, the proposed processing strategy will be described along with some key quality indicators that are under consideration. These indicators include the response rate, the coefficient of variation, the estimated bias and the R-indicator. Findings from simulation studies based on synthetic and survey data will be presented.

Key Words: Quality indicators, Survey processing, Integrated Business Statistics Program, Response representativeness

1. Introduction

In 2010, Statistics Canada reviewed its business methods and systems to identify opportunities to improve efficiencies, enhance quality assurance and increase responsiveness in delivering new statistical programs. To meet these objectives for business surveys, the *Integrated Business Statistics Program (IBSP)* was proposed. By 2016, it will provide a common survey framework for nearly 120 Statistics Canada business surveys.

The Unified Enterprise Survey (UES) will be one of the first surveys to be integrated into IBSP. UES includes nearly sixty different surveys covering manufacturing, services and distributive trade industries. Under the current UES model, the survey process is linear and occurs in a pre-determined set of periods as shown in figure 1. In general, active collection continues until the collection period is over. Once collection is closed, the processing starts, which includes editing, imputation and estimation. Manual inventions by subject matter specialist occur at three different places. First occurrence is before editing and imputation during the processing stage, and the second occurs after editing and imputation. The final manual invention is at the analysis stage, where analysts examine outliers, record consistency and estimates. Since we only produce estimates and quality indicators at the end of the survey cycle, follow-up activities and the first two manual interventions are done without knowing clearly their impact on overall estimates and quality. Furthermore, this process is not only long and labour-intensive, but also

batch processing is not feasible due to the alternation between automated and manual steps.

Figure 1 Current Unified Enterprise Survey (UES) Process

A number of studies also highlighted other weaknesses of the current UES model. One weakness is that significant amount of resources are needed for failed edit and non-response follow-up activities. Extensive data analysis and ensuring micro-data consistency are also very resource intensive (Cloutier, 2009).

IBSP addresses these problems by improving the efficiency and timeliness of the survey processes without significantly impacting on the accuracy of the resulting estimates. One key component of the IBSP is the Rolling Estimates (RE) model. It is a centralized processing model where estimates are produced and analyzed iteratively until an acceptable level of quality is reached. After each RE iteration, estimates and quality indicators for key domains and key variables are produced. If all the quality targets are met for a specific survey, active collection is closed early, so follow-up can be stopped; otherwise, follow-up or editing can be efficiently prioritized to units that influence key estimates and quality based on the quality indicators produced. A list of influential respondents is also produced using the quality indicators to reduce analysis burden. Hence, the RE process relies on a set of quality indicators and unit-level scores.

Godbout and Beaucage (2011) proposed many quality indicator options for the IBSP, but this paper mainly focuses on maximal absolute bias and maximal root mean square error (RMSE) estimation using the response propensities approach (Schouten, Cobben and Bethlehem, 2009). Section II describes the proposed indicators in more details. Section III describes the simulation study, and section IV discusses the findings from the study. Finally, section V addresses the future directions of the RE process.

2. Background and Notation

2.1 Quality Indicators

A quality indicator assesses the quality of one or many estimated statistics associated to a domain of interest. The quality indicators are grouped into 2 categories: covariate-based and item-based (Schouten, Calinescu and Luiten, 2011).

Covariate-based quality functions are derived from statistics using the estimated response propensities given the covariates, X. They are independent from the variables of interest. Response rates, R-indicator (Schouten, Cobben and Bethlehem, 2009), standardized maximum absolute bias, variance and maximal root mean square error (RMSE) are some quality indicators that are considered for the RE process. The standardized maximal absolute bias gives the bias in the worst case scenario, so it can track the upper bound of the non-response bias. Covariate-based quality indicators are measured on the set of sampled units, and they require a well-specified response model based on available covariates. They can be used in the preliminary collection phase such as during pre-contact or to manage collection of a master sample from which a second

phase sample will be selected and variables of interest will be collected in a latter phase (Godbout and Beaucage, 2011).

Item-Based quality functions are specific to a variable of interest. For example, the item-based maximal absolute bias or the maximal estimated RMSE are considered for the RE process. They are estimated from the respondents only. Item-based quality functions allow accurate monitoring of key variables and assessing if targets are achieved during regular collection. For this paper, we mainly focus on maximal absolute bias and maximal RMSE estimation using the response propensities approach (Schouten, Cobben and Bethlehem, 2009).

2.2 R-Indicator

Schouten et al (2009) presented the R-indicator, which measures the similarity between the response to a survey and the sample under investigation. This similarity is referred to as "representative response". The R-indicator relies on the individual response propensities in the sample. However, in a survey, we do not know the response propensity of all units, because we only have information from the responding units. An alternative is to estimate the individual response propensities using available auxiliary variables through methods, such as, logistic regression models. $\hat{r}_{Xk} = \hat{f}(x_k)$ is the estimated probability that unit k responds using auxiliary variable, X, and below is the formula to calculate the estimated R-indicator under a simple random sample (SRS) design drawn from a population *U* of size *N*. The R-indicator takes a value on the interval [0, 1], with 1 being perfect representativeness and 0 being the maximum deviation from representativeness.

$$\textit{R-indicator:}\quad \hat{R}_s(\hat{r}_X) = 1 - 2\hat{S}_s(\hat{r}_X) \tag{1}$$

Where,

$$\hat{S}_s(\hat{r}_X) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{n}\frac{1}{p_k}(\hat{r}_{Xk} - \bar{\hat{r}}_X)^2}\ , \text{ and } \bar{\hat{r}}_X = \frac{1}{N}\sum_{k=1}^{n}\hat{r}_{Xk}\frac{1}{p_k},$$

p_k is the first-order inclusion probability of unit *k* in sample *s*.

2.2 Estimator

Let's start with a population U of N units. For this paper, y_{ik} is the i[th] variable of interest that the survey is collecting for the k[th] unit and the parameter of interest is the total of each variable y_i.

$$\textit{Total of } y_i: \qquad t_{y_i} = \sum_{k=1}^{N} y_{ik} \tag{2}$$

For missing data due to non-response, it is a common-use to either impute or reweight. In this paper, we will consider estimators using the reweighting method. Given a SRS sample, *s*, subject to non-response, the NR-S estimator is calculated from the set of respondents, *r*, and it is inflated based on an average response rate.

$$\textit{NR-S Estimator}: \qquad \hat{t}_{y_i}^{NR-S} = p^{-1}(m/n)^{-1}\sum_{k=1}^{m} y_{ik} \tag{3}$$

For the NR-S estimator, $p = n/N$ is the sampling probability and the response rate (m/n) is the number of respondents divided by number of sampled units at the reweighting class level; for simplicity, we assume that the reweighting classes correspond to the whole sample. For more details on construction of reweighting classes, see Haziza and Beaumont (2007). The NR-S estimator is potentially biased if the response propensities are correlated to the variable y_i; in that case, we expect to attach quality indicators that are suitable to monitor this potential bias.

2.2 Variance and Mean Square Error

Using the formula from Särndal et al. (1992), we can estimate the variance of the total, $\hat{t}_{y_i}^{NR-S}$:

$$\textit{Estimated Variance of the total, } \hat{t}_{y_i}^{NR-S}: \hat{V}(\hat{t}_{y_i}^{NR-S}) = N^2 \frac{1-m/N}{m} \hat{S}_{y_i r}^2 \qquad (4)$$

Where, *m* is number of respondents, and $\hat{S}_{y_i r}^2$ is the variance of y_i for all the respondents. Given that NR-S has a potential non-zero bias, the relative root mean square error (RRMSE) is derived from the estimated variance.

$$\hat{RRMSE}(\hat{t}_{y_i}^{NR-S}) = (\hat{t}_{yi}^{NR-S})^{-1} \sqrt{\hat{B}^2(\hat{t}_{y_i}^{NR-S}) + \hat{V}(\hat{t}_{y_i}^{NR-S})} \qquad (5)$$

This highlights the need to have an accurate estimator of the non-response bias. As described by Schouten et al (2009), the bias can be derived from the true response propensities r:

$$\begin{aligned} B(\hat{t}_{y_i}, r) &= N\, S(y_i, r)/\bar{r} \\ &= N\, S(y_i) S(r)\, r(y_i, r)/\bar{r} \end{aligned} \qquad (6)$$

Where $S(y_i), S(r)$, $S(y_i, r)$ and $r(y_i, r)$ are respectively the standard deviation of the y_i and the r, the covariance between y_i and the r and their coefficient of correlation. Assuming a maximal correlation of $r(y_i, r) = 1$ and substituting $S(y_i)$ and $S(r)$ by $\hat{S}_r(y_i)$ and $\hat{S}_s(\hat{r}_X)$, a quality indicator for $\hat{t}_{y_i}^{NR-S}$ can be formulated using the R-indicator as the upper bound of the non-response bias, which shows the impact under worst-case scenarios.

$$\textit{Item-based Maximal Bias: } \hat{B}_m(\hat{t}_{y_i}^{NR-S}, \hat{r}_X) = N\, \hat{S}_r(y_i)\left(1 - \hat{R}_s(\hat{r}_X)\right)/2\bar{\hat{r}}_X \qquad (7)$$

Finally, the maximal RRMSE can be estimated:

$$\hat{RRMSE}_m(\hat{t}_{y_i}^{NR-S}, \hat{r}_X) = (\hat{t}_{yi}^{NR-S})^{-1}\left[\sqrt{\hat{B}_m^2(\hat{t}_{y_i}, \hat{r}_X) + \hat{V}(\hat{t}_{y_i})}\right] \qquad (8)$$

3. Simulation Study Methodology

In order to study the quality indicators more closely, especially the maximal absolute bias estimation using response propensities, a simulation study is conducted. The objective of this study is to see how well the item-based maximal absolute bias and maximal RRMSE estimated using the response propensities approach performs under the conditions:

- The variables collected, y_i, are correlated at different levels with the response propensity of the units.
- The response model is correctly or incorrectly specified.

First, a population is created by generating 500 random numbers from the uniform distribution between [0, 1], and they are stored in the auxiliary variable, *X*. For each unit, the response propensity is created following a logistic model:

Unit Response Propensity: $$r_{Xk} = \frac{\exp(x_k)}{1+\exp(x_k)} \quad (9)$$

Variables of interest, y_1 to y_5, are created for each unit using *X* as well. These variables of interests correlate at different levels as shown in table 1.

	$\mathbf{y_1}$	$\mathbf{y_2}$	$\mathbf{y_3}$	$\mathbf{y_4}$	$\mathbf{y_5}$
COR(X, y)	1	0.99	0.73	0.16	0.04
COR(r , y)	0.99	0.99	0.73	0.16	0.04

Table 1: Table of Correlation

A simple random sample (SRS) of 100 units is selected from the population considering SRS design, and 1000 Monte Carlo replicates are generated.

We then use a logistic regression with covariate *X* to estimate $\hat{r}_{Xk}$ for each Monte Carlo replicate. The item-based maximal absolute bias and root mean square error using the response propensities approach are calculated using formula (7) and (8). Dividing by the estimate of the population total, $\hat{t}_{y_i}$, we get the estimated relative maximal absolute bias ($R\hat{B}_m = \hat{t}_{y_i}^{-1}\hat{B}_m$) and the estimated RRMSE ($RR\hat{M}SE_m = \hat{t}_{y_i}^{-1}R\hat{M}SE_m$), and they are averaged over all replicates and then compared to two Monte Carlo measures of goodness, the relative bias and the RRMSE, using formula (10) and (11).

Monte Carlo Relative Bias: $$R\hat{B}_{MC} = \left| t_{y_i}^{-1}\frac{1}{1000}\sum_{j=1}^{1000}\hat{t}_{y_i,j} - t_{y_i} \right| \quad (10)$$

Monte Carlo RRMSE: $$RR\hat{M}SE_{MC} = t_{y_i}^{-1}\frac{1}{1000}\left(\sum_{j=1}^{1000}(\hat{t}_{y_i,j} - t_{y_i})^2\right)^{1/2} \quad (11)$$

The NR-S estimator proposed in (3) will be compared to an alternate estimator, NR-U, based on estimated response propensity at the unit level derived from covariates (Godbout, Beaucage and Turmelle, 2011).

NR-U Estimator: $$\hat{t}_{y_i}^{NR-U} = p^{-1}\sum_{k=1}^{m}\hat{r}_{xk}^{-1}y_{ik} \quad (12)$$

The NR-U estimator is unbiased but has a large variability if some $\hat{r}_{Xk}$ are estimated close to 0 or if the response propensities are not correlated to the variable y_i. More importantly, NR-U is biased if the response propensity model is incorrect. The NR-S and NR-U estimators are the same if we set $\hat{r}_{Xk} = m/n$ in the case where there are no covariates available or significant (Godbout, Beaucage and Turmelle, 2011). Overall, we expect that NR-S estimator will perform better in term of mean square error (MSE) due

to the robustness of the response rate (m/n) especially if the response propensities are homogeneous enough.

4. Results Discussion

First we exam our model when the response propensity is estimated using the correct model. Since we created the response propensity following a logistic model, using a logistics regression having X as the independent variable should give us the correct estimation.

Table 2 shows the Monte-Carlo and item-based estimated maximal absolute relative bias and RRMSE. For estimator NR-U, we see that the Monte-Carlo relative bias is very small, as expected since the non-response model is well specified in this case. From y_1 to y_5, we see that the Monte Carlo estimated RRMSE is getting higher. Recall y_1 correlates the most with X (and r_{Xk}) while y_5 correlates the least, so a lower correlation between y and the response propensities increases the variance of the NR-U estimates.

For estimator NR-S, the Monte-Carlo relative bias decreases from y_1 to y_5, because the decline in correlation between response propensity and y lowers the bias. We see $\hat{RB}_m$ is slightly bigger than the highest Monte Carlo relative bias, for y_1, which suggests that the maximal relative bias calculated using the response propensities is able to provide an upper bound of the bias under this situation.

Variables	Relative Bias			Root Mean Square Error			R-Indicator
	NR-S		NR-U	NR-S		NR-U	
	RB_{MC}	RB_m	RB_{MC}	$RRMSE_{MC}$	$RRMSE_m$	$RRMSE_{MC}$	
y_1	6.1%	6.5%	0.1%	9.0%	9.3%	5.2%	0.86
y_2	5.6%	5.9%	0.0%	8.2%	8.7%	4.7%	
y_3	3.4%	4.8%	0.1%	5.9%	7.5%	4.3%	
y_4	1.1%	6.0%	0.2%	6.4%	8.7%	6.4%	
y_5	0.6%	6.7%	0.2%	7.0%	9.4%	7.2%	

Table 2: Estimated Maximal Absolute Relative Bias and estimated RRMSE under the correct response propensity model

Now we want to study what happens when the response propensity model is incorrect. In figure 1, the bottom blue line shows a graph of the auxiliary variable, X, against true response propensity from the population, ρ. The top red line is created following the model: $r_{Xk} = 0.5 + 0.25 * x_k$. The graph shows that this model preserves the order of the response propensities, their average and their standard deviation, but it loses the curvature seen in the logistic model. Since we want to study when the response propensity is modeled incorrectly, we now use $\hat{r}_{Xk} = 0.5 + 0.25 * x_k$ to estimate the response propensity instead of the logistic regression.

Table 3 shows the Monte-Carlo and estimated maximal absolute relative bias and relative RMSE. For estimator NR-U, the Monte-Carlo relative bias shows a positive relative bias (between 1.0% and 1.5%) because the response propensity is estimated using the wrong

model and its RRMSE became higher than NR-S's RRMSE. On the other hand, since NR-S is not based on the response propensity model, the Monte-Carlo measures of goodness are the same. In this scenario, in which the model is misspecified but close to the true model, the estimated maximal absolute relative bias didn't change that much so it still gives an upper bound of the Monte-Carlo estimated relative bias.

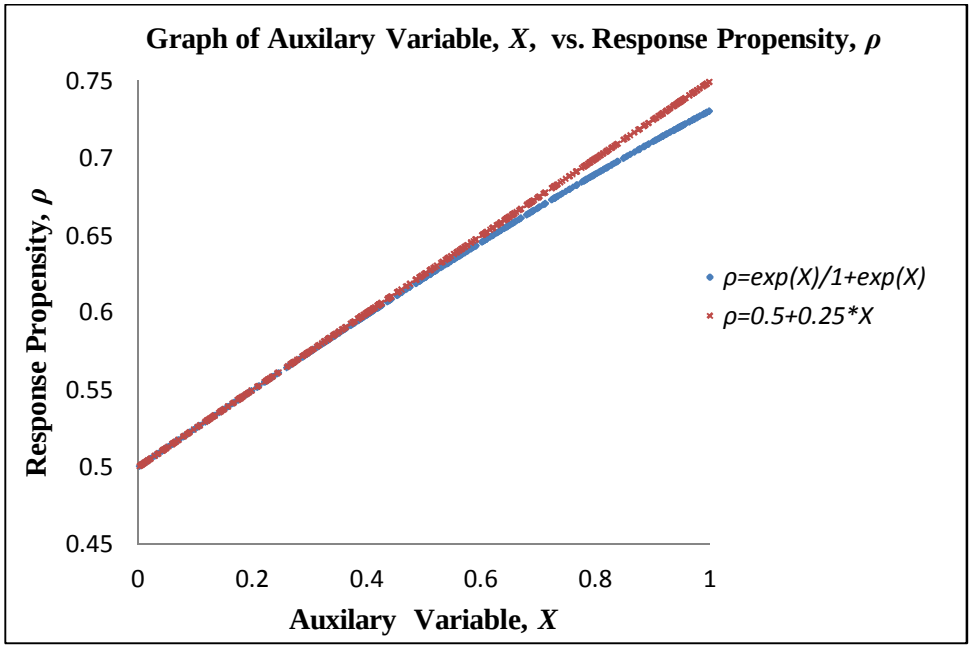

Figure 1: Graph of Auxiliary Variable, *X*, vs. Response Propensity, ρ

Variables	Relative Bias			Root Mean Square Error			R-Indicator
	NR-S		NR-U	NR-S		NR-U	
	RB_{MC}	RB_m	RB_{MC}	$RRMSE_{MC}$	$RRMSE_m$	$RRMSE_{MC}$	
y_1	6.1%	6.4%	1.5%	9.0%	8.8%	9.3%	0.86
y_2	5.6%	5.9%	1.5%	8.2%	8.3%	9.0%	
y_3	3.4%	4.7%	1.2%	5.9%	7.1%	8.6%	
y_4	1.1%	6.0%	1.0%	6.4%	8.3%	9.9%	
y_5	0.6%	6.6%	1.0%	7.0%	8.9%	10.4%	

Table 3: Estimated Maximal Absolute Relative Bias and estimated RRMSE under the incorrect response propensity model

This result is important in real survey processes, because the true underlying response propensity model is generally unknown although relevant covariates are often available from collection specifications, paradata and/or historical data. The model used to estimate the response propensity might be incomplete but realistic. This simulation suggests that the NR-S estimator is more robust to response model misspecifications than NR-U and that the estimated maximal relative bias can still be accurate if the response model is close to the true model. On the other hand, other simulation scenarios showed that the estimated maximal relative bias can be too high (resp. too low) if the standard deviation of the response propensities is significantly overestimated (resp. underestimated).

5. Conclusion and Future Work

The Rolling Estimates model is one way to achieve higher efficiency, quality assurance and responsiveness in survey processing, and it relies on a list of quality indicator. The maximal absolute bias using the response propensities proposed by Schouten et al provides an option to estimate the upper bound of the non-response bias.

From the simulation results, we see that the maximal absolute bias gives a good estimate of the upper bound of the bias if the response propensity model is well specified, and it is less affected by small model misspecifications than the NR-U estimator. Nevertheless, the maximal absolute bias could be used to identify the units with low response propensities having a negative impact on the RRMSE through the bias even though the bias itself might be incorrect. The simulation study presented in this paper is a feasibility study conducted using a simplified uniform population with one stratum sampling design. To enhance the study, we can use skewed population or real survey data with stratified sampling design and different scenarios of model misspecifications.

We are currently conducting a Rolling Estimates prototype that involves all the surveys from the reference year 2010 of the current UES program. The objectives of this prototype are to assess the potential benefits of this new process model and set up new procedures to maximize quality under a given collection budget. At this time, we are doing four iterations at July, August, September and October. For this prototype, only basic quality indicators are implemented based on a variety of key variables. The prototype will be repeated for reference years 2011 and 2012, and bias measurement will definitely be implemented in the survey process in the future and the response propensity model is a valuable option.

References

Cloutier, L. (2009). Selective Editing for Business Surveys at Statistics Canada. Statistics Canada, Conference of European Statisticians, Work Session on Statistical Data Editing (Neuchâtel, Switzerland, October 2009).

Enterprise Statistics Division (2010). Integrated Business Statistics Program Blueprint. Statistics Canada, Internal document.

Godbout, S., Beaucage, Y. and Turmelle, C. (2011) Quality and efficiency using a top-down approach in the Canadian's Integrated Business Statistics Program. Conference of European statisticians, Work session on Statistical Data Editing (Ljubljana, Slovenia, 9-11 May 2011).

Godbout, S. and Beaucage, Y. (2011) Using Quality Indicators to Manage Collection and Editing for the Integrated Business Statistics Program.

Granquist, L. and Kovar, J.G. (1997). Editing of Survey Data: How Much Is Enough? Survey Measurement and Process Quality, Lyberg et al (eds.), New York, Wiley, pp. 415-435.

Haziza, D. and Beaumont, J.-F. (2007), On the construction of imputation classes in surveys, International Statistical Review (2007), 75, 1, 25–43

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistics, Vol. 19, No. 2, 2003, pp. 177-199.

Latouche, M. and Berthelot, J.-M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. Journal of Official Statistics, Vol. 8, No. 3, 1992, pp. 389-400.

Laflamme, F. and Mohl, C. (2007). Research and Responsive Design Options for Survey Data Collection at Statistics Canada. American Statistical Association, Proceedings of the Section on Survey Research Methods.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. New York, Springer-Verlag, 694 p.

Schouten, B., Calinescu, M. and Luiten, A. (2011). Optimizing Quality of Response Through Adaptive Survey Designs. Discussion paper, Statistics Netherlands, Den Haag, available soon through http://www.cbs.nl/nl-NL/menu/methoden/onderzoek-methoden/discussionpapers/archief/2011/default.htm

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. Survey Methodology, June 2009 101, Vol. 35, No. 1, pp. 101-113, Statistics Canada, Catalogue No. 12-001-X

Statistics Canada (2000). Policy on Informing Users of Data Quality and Methodology. http://www.statcan.gc.ca/about-apercu/policy-politique/info_user-usager-eng.htm

Statistics Canada (2009). Statistics Canada Quality Guidelines. Fifth Edition – October 2009, Catalogue no. 12-539-X. http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf

Xie, H., Godbout, S., Youn, S. and Lavallée, P. (2011). Collection Follow-Up Operation Using Priority Scores for Business Surveys. Conference of European Statisticians, Work Session on Statistical Data Editing (Ljubljana, Slovenia, 9-11 May 2011).