WP.24
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (iv): Micro editing – methods and software

Collection Follow-Up Operation Using Priority Scores For Business Surveys

Invited Paper

Prepared by Hansheng Xie, Serge Godbout, Sungjin Youn and Pierre Lavallée, Statistics Canada

I. Introduction

1. Data collection is one of the most expensive operations in a survey process and has a direct impact on data quality. For business surveys at Statistics Canada, significant costs are spent on data collection, especially on the follow-up operation. In order to improve the cost-efficiency of data collection under the constraints of a limited budget, different collection methods have been developed and implemented.

2. These different collection methods were developed at different times and have been applied to various surveys. Previously, these surveys were designed and processed independently. To increase efficiency in the survey process and produce reliable and coherent estimates for key economic variables, Statistics Canada has gradually integrated many business surveys into a unified survey program. The Unified Enterprise Survey (UES) is such a program that unifies annual business surveys from the service, wholesale and retail, transportation, aquaculture, banking and manufacturing industries. For more information regarding the UES, see Brodeur *et al.* (2006).

3. Two systems based on two different methodologies are currently used for UES data collection. This is because a system was first developed for the Annual Survey of Manufacturers and Logging (ASML) in 2001 when the survey was not yet integrated into the UES program. In 2002, a second system was developed for all other UES surveys. In 2004, the ASML was integrated into the UES, and since then these two systems have been used separately during data collection. Specifically, one system is used for the manufacturing surveys, while the other system is used for the other surveys. Since the two systems are based on different methodologies, they can have different impacts on the final estimates and collection follow-up costs. For more information regarding the two systems, see Evra and DeBlois (2007), Philips (2003) and Pursey (2003).

4. In order to further improve the cost-efficiency of UES data collection and the redesign of an effective and unique system, we conducted a study on the different methodological approaches for collection. The collection methods that are considered are associated with maximizing an economically weighted response rate (EWRR). This is done by establishing follow-up priorities based on a collection

unit's contribution to the EWRR. High priorities are given to collection units with large contributions (i.e., high priority scores). To compare the different collection methods/options and determine which were optimal, a simulation study was conducted using Reference Year (RY) 2006 data obtained from the UES.

5. The remaining sections of this paper are organized as follows. Section II describes the collection methods, as well as their applications and estimation. In Section III, the different collection methods are applied to the data at certain levels of domain of interest, and the methods are then compared by measuring their impacts on the final estimates of economic variables, as well as on the cost associated with each collection method. In section IV, we assess the bias of the estimator which can explain some of the results of section III. Finally, in Section V, a discussion follows on the results of the simulations; describing which circumstances the use of priority scores improves the collection process while maintaining nearly unbiased estimates.

II. Collection Methods and Applications

6. For each survey in the UES, a sample of establishments (or groups of establishments) is selected using the method of stratified simple random sampling without replacement (SRSWoR). To simplify the discussion, it is assumed that sampling is done only at the establishment level. According to this sample design, the population of interest U be divided into H strata where stratum h contains N_h establishments. In each stratum h, a sample s_h of size n_h is selected using SRSWoR. Each establishment k is then selected with probability $\pi_k = n_h / N_h$ for $k \in h$. To estimate the population total $Y = \sum_{h=1}^{H} \sum_{k=1}^{N_h} y_{hk}$ for a variable of interest y, we can use the Horvitz-Thompson estimator:

$$\hat{Y} = \sum_{k \in s} w_k y_k \tag{1}$$

where $s = \bigcup_{h=1}^{H} s_h$ is of size $n = \sum_{h=1}^{H} n_h$, and $w_k = 1/\pi_k$ is the sampling weight.

7. In the UES, depending on the survey's requirements, data are collected at either the establishment level or the level of a group of establishments which are called collection units. Again, to simplify, it is assumed that data are collected only at the establishment level. The data collection process consists of designing questionnaires for each survey, mailing out questionnaires to selected units, and following up with the establishments that either did not respond to the mail-out questionnaires or did not pass specific collection edits. A collection method is applied at the collection stage to prioritize follow-up effort and provide a useful day-to-day operational plan for staff to follow.

8. During data collection, some units are found to be in-scope for the UES, and some others are identified as out-of-scope. In-scope units have been determined to belong to the target universe for the survey. Responding units include all units that are deemed to have responded by virtue of having provided usable information. Let L be the set of ℓ in-scope units in the sample and let R be the set containing the responding units in the sample. Then, the set $R \cap L$ contains the r in-scope responding units in the sample. Note that at the end of the collection process, some units are left as unresolved, i.e., they are not determined to be in-scope nor out-of-scope. Fortunately, for the UES, unresolved units correspond to a very small portion of the total sample size. For the present paper, it will then be assumed that unresolved units are non-existant.

A. Economically Weighted Response Rate

9. Many collection methods are associated with maximizing a EWRR. This weighted response rate at the data collection phase takes account of both the sample weight w_k and an economic variable x_k at the level of domain of interest. It is defined as follows:

$$\text{EWRR} = \frac{\sum_{k \in R \cap L} w_k x_k}{\sum_{k \in L} w_k x_k}. \quad (2)$$

10. For an in-scope unit in the sample, formula (2) can be used to determine a unit's priority if there is no sum in the numerator. This weighted rate ϕ_k at the level of an establishment *k* is usually called the priority score, and formula (2) can be rewritten in the following form:

$$\phi_k = \frac{w_k x_k}{\sum_{i \in L} w_i x_i}, \text{ for } k \in L. \quad (3)$$

11. Follow-up priorities are set up based on an establishment's contribution to the EWRR. Higher priorities are given to those establishments with large contributions, i.e., high priority scores. In each survey, the priority scores are calculated for all establishments belonging to the same domain of interest, such as industry and geography (usually province/territory). The calculation of priority scores is based on its specific collection methodology.

12. With a targeted EWRR set for each group, the establishments within each group are sorted by priority score in descending order. These ordered units are divided into two subsets using the targeted EWRR as the cut-off. The units belonging to the top subset will then be followed up in case of non-response ("follow-up units"). The remaining units in the bottom subset will not be followed up ("non-follow-up units").

13. The priority scores are recalculated periodically during data collection based on updated information. During recalculation, all the units that have responded contribute to the EWRR and become non-follow-up units. Confirmed non-responses also become non-follow-up units since they will not provide any data. Out-of-scope units are removed entirely from the process, and some non-follow-up units are promoted for follow-up to make up for the non-responses.

B. Collection Methods/Options

14. Formula (3) can be employed to determine priorities at the unit level using different economic weights when combined with the sample weight, such as different economic variables, the previous year's economic values, or available updated economic values. The priority score can also be computed at different levels of domain of interest.

15. Suppose that the economic data contains several output commodities. Let x_{kj} represent an economic value produced by the j^{th} commodity of the k^{th} establishment and w_k be the sample weight for x_{kj}. In addition, assume that this unit produces m_k output commodities and the annual sample contains *n* establishments. For calculating the priority score for the k^{th} establishment, formula (3) can be rewritten as follows:

$$\phi_k^t = \frac{\sum_{j=1}^{m_k} w_k^t x_{kj}^{t-1}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i^t x_{ij}^{t-1}} \quad (4)$$

where *t* denotes the current year and *t*-1 denotes the previous year.

16. Under the above assumptions, if the weights $Q_j^{t-1} = \left(\sum_{k=1}^{n} x_{kj}^{t-1}\right) \Big/ \left(\sum_{k=1}^{n}\sum_{j=1}^{m_k} x_{kj}^{t-1}\right)$ of an establishment's output commodities are further considered for this unit, the priority score for the k^{th} unit can be calculated using the following other formula:

$$\phi_k'^t = \sum_{j=1}^{m_k} \left\{ Q_j^{t-1} w_k^t x_{kj}^{t-1} \times \frac{\sum_{j=1}^{m_k} x_{kj}^{t-1}}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} x_{ij}^{t-1}} \right\}. \tag{5}$$

17. Note that this formula cannot be derived directly from the first three formulas. It has been used for ASML data collection.

18. If the updated economic values are available for this unit, formula (5) takes the following form:

$$\phi_k''^t = \sum_{j=1}^{m_k} \left\{ Q_j^{t-1} w_k^t x_{kj}^t \times \frac{\sum_{j=1}^{m_k} x_{kj}^t}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} x_{ij}^{t-1}} \right\}. \tag{6}$$

19. When a unit has been contacted but does not respond, the responding probability decreases after each unsuccessful contact. Based on an experimental formula used for ASML data collection, the unit's priority score can be reduced using the following formula:

$$\phi_k'''^t = \phi_k^t \times \frac{1 - (\text{number of contacts} \times \alpha)}{\beta} \tag{7}$$

where $\beta > \alpha > 0$.

C. Estimation

20. Once the collection follow-up process is completed, the sample *s* contains three sets of units:
   (a) R_1 : set of n_{R1} units (in-scope or out-of-scope) that responded without the need of the follow-up process;
   (b) R_2 : set of n_{R2} units (in-scope or out-of-scope) that responded after being followed up;
   (c) NR : set of n_{NR} non-responding units (in-scope or out-of-scope).

21. Note that $R = R_1 \cup R_2$. We also define the set $F = s \setminus R_1$ of $n_F = n - n_{R1}$ units that went into the follow-up process. We need to distinguish between R_1 and R_2 because of the nature of the responding process. Before any follow-up, it can be assumed that the establishments that are part of the stratum sample s_h have an equal response probability, i.e., $p_k^{(1)} = P(\text{unit } k \text{ responds} \big| k \in s_h) = p_h^{(1)}$, say. This response process is called *missing at random* (MAR) (Little and Rubin, 1987). Now, for the units of s_h that are not part of R_1, the follow-up process using the priority scores gives higher responses probabilities to units that have higher priority scores. Therefore, these units should have response probabilities that are proportional to their priority score, i.e., for $k \in s_h$ and $k \notin R_1$ (i.e., $k \in F_h$), we should have $p_k^{(2)} = P(\text{unit } k \text{ responds} \big| k \in F_h, s_h) \propto \phi_k$. This response process is *not* MAR. Assuming that the number of follow-up units $n_{h,R2}$ ares fixed (and $n_{h,R2} \le n_{h,F}$), we could have $p_k^{(2)} = n_{h,R2}\phi_k / \sum_{i \in F_h} \phi_i$ for $k \in h$. Note that for the UES, the follow-up is made by considering the n_{R2} largest priority scores, and ordering them

by descending order. The distinction between R_1 and R_2 should normally be taken into account in the estimation process to insure unbiasedness.

22. After the collection follow-up process, nearest-neighbour donor imputation method is applied to the non-responding units. The auxiliary variables used for determining the nearest donor are *the same as the economic variables x* used in the determination of the priority scores. For each in-scope unit *k* part of *U*, we then obtain an imputed value $\hat{y}_k$. Note that out-of-scope units do not need to be imputed since, by definition, their variables of interest y_k are set to zero.

23. Taking into account non-response and imputation, estimator (1) can then be split in three components:

$$\begin{aligned}\tilde{Y} &= \sum_{k\in R_1} w_k y_k + \sum_{k\in R_2} w_k y_k + \sum_{k\in NR} w_k \hat{y}_k \\ &= \tilde{Y}^{(1)} + \tilde{Y}^{(2)} + \tilde{Y}_{imp}^{(NR)}\end{aligned} \tag{8}$$

where the sampling weights $w_k = 1/\pi_k = N_h / n_h$ for $k \in h$ remain unchanged.

24. The properties (bias, variance and mean square error) of the estimator are studied in the next section. In addition to the quality of estimation, the follow-up cost associated with each collection method/option is also assessed.

III. Simulation Study

A. Methodology

25. In the study, the operational environment of collection follow-up was simulated using the economic data of RY2006 obtained from several UES surveys. These data cover the following industries: Fabricated Metal Product Manufacturing (12 economic variables), Repair and Maintenance (9 economic variables), Advertising and Related Services (9 economic variables), and Store Retailing (11 economic variables). The populations for these data had relatively large numbers of establishments (>300). Note that for the simulations, only in-scope establishments were used.

26. One thousand samples were randomly selected from each population using a stratified SRSWoR design. Size stratification was executed using the Lavallée-Hidiroglou algorithm (Lavallée and Hidiroglou, 1988). After setting a targeted EWRR from 50% to 100%, respectively, each collection method/option was applied to a random sample at the level of domain of interest to obtain a group of follow-up units. The collection follow-up process was then simulated through several dynamic runs until it reached its target.

27. In the study, the seven methods/options listed below were compared:

- M1 – using the previous year's revenue (i.e., Total Revenue) as an economic weight, i.e., based on formula (4).
- M2 – using the same economic weight as M1 but establishing priorities at a lower level, i.e., size stratum.
- M3 – using the same economic weight as M1 but reducing a unit's priority after several unsuccessful contacts, i.e., based on formulas (4) and (7).
- M4 – using the previous year's Total Sales of Goods and Services Produced (TSGSP) and Commodity as an economic weight, i.e., based on formula (5).
- M5 – using the same economic weight as M4 but updating TSGSP, i.e., based on formula (6).

M6 – using the same economic weight as M5 but reducing a unit's priority after several unsuccessful contacts, i.e., based on formulas (6) and (7).
M7 – randomly choosing some establishments for follow-up at the beginning of collection.

28. During data collection, a unit could be in one of the three possible statuses: responding, non-responding, or still-in-process. A follow-up unit has a higher likelihood of response than a non-follow-up one. In a dynamic run, the two types of units (follow-up and non-follow-up) would have the assumed probabilities of being in a given status given in Table 1.

Table 1: Assumptions of collection status of each unit in a dynamic run

	Status	Probability
Follow-up units	Responding	25%
	Non-responding	5%
	Still-in-process	70%
Non-follow-up units	Responding	2%
	Non-responding	0%
	Still-in-process	98%

29. In order to make the simulation less time-consuming, about 10 dynamic runs were simulated in the study for reaching a targeted EWRR, rather than having over 50 dynamic runs in production. With these simulation runs, as well as the probabilities shown in the above table, the response rates and the EWRR obtained at the end of the simulation would be similar to those rates achieved at the end of the collection in production.

30. After simulating the collection follow-up process, the nearest-neighbour donor imputation method was applied to non-response units, and total estimates were then produced using (8).

31. In the study, the standard for comparisons was based on both the quality of estimation and the follow-up cost associated with each collection method/option, assuming that each contact had a unit cost. For measuring the quality of estimation, the Relative Bias (RB) and the Relative Root Mean Square Error (RRMSE) were computed for each sample relative to a collection method/option. The sample estimates were compared to the *pseudo true values* obtained from the UES surveys.

32. The RB measures the deviation of an average estimate from the reference value and is defined as

$$\mathrm{RB} = \frac{1}{Y}\left|\left(\frac{1}{1000}\sum_{r=1}^{1000}\tilde{Y}_r\right) - Y\right| \tag{9}$$

where $\tilde{Y}_r$ is the r^{th} estimate based on a collection method/option and Y is the pseudo true value of the total of an economic variable.

33. For measuring the total variation from the reference value, the RRMSE is defined as

$$\mathrm{RRMSE} = \frac{1}{Y}\sqrt{\frac{1}{1000}\sum_{r=1}^{1000}\left(\tilde{Y}_r - Y\right)^2}\,. \tag{10}$$

B. Analysis of the Main Results

34. For simplicity, this section presents the main results of the study based on the data of Fabricated Metal Product Manufacturing and Store Retailing. In terms of Total Revenue, the population for the

former data set is skewed (skewness: 8.8), while the population for the latter data set is extremely skewed (skewness: 15.8).

B.1 Results based on Fabricated Metal Product Manufacturing data

35. Fabricated Metal Product Manufacturing data contains 12 economic variables and 1,037 establishments. Except for Expenses of Vehicle and Aircraft Fuel (EVAF), 11 of the 12 economic variables are highly correlated to each other, so that they have similar patterns. Since Total Sales of Goods and Services Produced (TSGSP) is one of the most important of these 11 variables, it is only necessary to discuss the results for the following two economic variables: TSGSP and EVAF.

36. For the simulation, two economic weights were used: the previous year's Revenue (i.e., Total Revenue); the previous year's TSGSP. The current TSGSP has *high correlations* with the two economic weight variables, while EVAF has *low correlations* with them.

37. The cost corresponding to a targeted EWRR is usually different for the different collection methods/options, except for the EWRR of 100% when all the units in a sample are followed up.

38. Any modifications done to the comparable original method/option, such as establishing priorities at a lower level, reducing a unit's priority after several unsuccessful contacts, and updating the economic weight variable, will usually increase the cost. Therefore, as expected, the higher the targeted EWRR and/or the more complex the collection method/option, the higher the follow-up cost.

39. For our simulations with TSGSP, the cost ranges from 74 to 296, the RB ranges from 0.002 to 0.080, and therefore, for all methods, the RB is relatively small. The RRMSE ranges from 0.095 to 0.167. As the cost increases along with the targeted EWRR, the RB and RRMSE usually go down consistently, or vary within a very small range for all the collection methods/options. For most of the targeted EWRRs, M3 has the best performance among the seven collection methods/options in terms of the quality of estimation. The associated follow-up cost is also only slightly higher than that of using M1. It can also be seen that using M7 is expensive in terms of achieving a high-targeted EWRR. This indicates that, with the same cost, one can achieve a higher targeted EWRR when using methods other than M7.

40. For EVAF, the RB's are found to be relatively small for all collection methods/options. The range of the cost is the same as that obtained for TSGSP. However, a lot of variability appears in the quality of estimation for the different collection methods/options when there is a low correlation between the economic weight variables and the estimated economic variables.

41. As the cost increases, the RB and RRMSE usually increase and decrease over a large range. The RB ranges from 0.000 to 0.047, while the RRMSE ranges from 0.269 to 0.612. The range of the RRMSE is much larger (4.8 times) than that obtained for TSGSP, even though the range of the RB obtained for EVAF is smaller (0.6 times) than that obtained for TSGSP. In general, M3 is still optimal in terms of the quality of estimation. M7 could also be used to improve the quality of estimation while achieving a low-targeted EWRR.

B.2 Results based on Store Retailing data

42. In terms of Total Revenue, the population of Store Retailing data is extremely skewed. This population contains 11 economic variables and 1,458 establishments. Except for TSGSP, 10 of the 11 economic variables correlate highly with each other. Total Operating Revenue (TOR) is one of the most important of the 10 variables. Therefore, only the results for the two economic variables need to be obtained: TOR and TSGSP.

43. For this dataset, no commodity dimension is involved and the economic weight variable is the previous year's revenue (i.e. Total Revenue). TOR has a *high correlation* with Total Revenue, while TSGSP has a *low correlation* with it.

44. Since no commodity dimension is involved, only the first four methods/options were compared. For TOR, the cost ranges from 22 to 137, the RB ranges from 0.00 to 0.49, while the RRMSE ranges from 0.23 to 2.82. As the cost increases, the RB and RRMSE usually decrease or are relatively stable for all collection methods/options. For most of the targeted EWRRs, M2 has the best performance among the four collection methods/options in terms of the quality of estimation. However, the associated follow-up cost is slightly higher than that of using M1. Using M7 is still expensive in terms of achieving a high-targeted EWRR.

45. For TSGSP, the range of the cost is the same as that obtained for TOR. Since there is a low correlation between the economic weight variables and the estimated economic variables, a lot of variability occurs in the quality of estimation for the different collection methods/options. As the cost increases, the RB and RRMSE usually fluctuate widely. For TSGSP, the RB ranges from 0.002 to 0.818, while the RRMSE ranges from 2.75 to 11.64. The range of the RB for TSGSP is 1.7 times larger than that obtained for TOR, and the range of the RRMSE for TSGSP is 3.5 times larger than that obtained for TOR.

46. For most of the targeted EWRR's, M2 is still optimal in terms of the quality of estimation, and a cost increase can only be seen for low EWRRs in comparison with M1. Nevertheless, further reducing a unit's priority after several unsuccessful contacts results in a slight cost increase and minimal improvement in the quality of estimation. In contrast, using M7 could still improve the quality of estimation while achieving a low-targeted EWRR.

47. When comparing the results based on Fabricated Metal Product Manufacturing data with those based on Store Retailing data, the follow-up cost for the former dataset is higher than that of the latter dataset, i.e. the maximum cost is 296 vs. 137. At the same time, the quality of estimation for Fabricated Metal Product Manufacturing is better than that of Store Retailing, e.g., the maximum RRMSE for a highly correlated variable is 0.17 vs. 2.82 and the maximum RRMSE for a weakly correlated variable is 0.61 vs. 11.64. This is because the former population is less skewed to the right than the latter, and contains fewer large units. In order to achieve a targeted EWRR, more units need to be followed up for Fabricated Metal Product Manufacturing than for Store Retailing, which increases the cost but benefits the quality of estimation.

IV. Unbiasedness of the estimation process

48. This section provides the theoretical justifications (in terms of bias) to discuss the performance of estimator (8) in the simulation study. For simplicity, because all computations are performed stratum by stratum, we will omit the subscript *h*.

49. We propose to use a step-by-step approach. First, let us assume that we would produce an estimate of *Y* using only the sample of respondents R_1. The convenient estimator to be used would be the post-stratified estimator:

$$\hat{Y}_{post}^{(1)} = \frac{\sum_{k \in R_1} w_k y_k / p_k^{(1)}}{\sum_{k \in R_1} w_k / p_k^{(1)}} N = \frac{\sum_{k \in R_1} w_k y_k / p^{(1)}}{\sum_{k \in R_1} w_k / p^{(1)}} N = \frac{\hat{Y}^{(1)}}{\hat{N}^{(1)}} N \qquad (11)$$

It can be shown that this estimator is asymptotically unbiased.

50. Let us now assume that we would produce an estimate of Y using only the two samples of respondents, R_1 and R_2, without imputation. The convenient estimator to be used in this case would be the post-stratified estimator:

$$\hat{Y}_{post}^{(1+2)} = \sum_{k \in R_1} w_k y_k + \frac{\sum_{k \in R_2} w_k y_k / p_k^{(2)}}{\sum_{k \in R_2} w_k / p_k^{(2)}} \sum_{k \in F} w_k = \tilde{Y}^{(1)} + \frac{\hat{\tilde{Y}}^{(2)}}{\hat{\tilde{N}}^{(2)}} \tilde{N}^{(F)} = \tilde{Y}^{(1)} + \tilde{Y}_{post}^{(F)} \tag{12}$$

As for (11), it can be shown that this estimator is asymptotically unbiased.

51. Using again only the two samples of respondents R_1 and R_2, without imputation, a different estimator than (12) would arise naturally. This estimator is:

$$\hat{Y}_{natural}^{(1+2)} = \frac{\sum_{k \in R_1} w_k y_k + \sum_{k \in R_2} w_k y_k}{\sum_{k \in R_1} w_k + \sum_{k \in R_2} w_k} N = \frac{\tilde{Y}^{(1)} + \tilde{Y}^{(2)}}{\tilde{N}^{(1)} + \tilde{N}^{(2)}} N \tag{13}$$

Estimator (13) combines the total sample of respondents, R_1 and R_2, to estimate the total Y. It can be shown that $Y_{natural}^{(1+2)}$ is biased.

52. It should be noted that in the case where the priority scores are all equal (which implies that $p_k^{(2)} = p^{(2)}$), it can be shown that $E(\hat{Y}_{natural}^{(1+2)}) \approx Y$. Therefore, in this case, estimator (13) is asymptotically unbiased. Having $p_k^{(2)} = p^{(2)}$ means that the follow-up process gives equal response probabilities to all units in the follow-up process. We then have MAR for both response sets R_1 and R_2. Note that if the priority scores are proportional to the variable of interest (which implies that $p_k^{(2)} \propto y_k$), $E(\hat{Y}_{natural}^{(1+2)}) \neq Y$ and thus, estimator (13) remains biased.

53. Let us now consider the "real" situation, i.e., where we produce an estimate of Y using estimator (8). This estimator uses the two samples of respondents, R_1 and R_2, and imputed values for the set NR of non-respondents.

54. As mentioned earlier, after the collection follow-up process, nearest-neighbour donor imputation method is applied to the n_{NR} non-responding units. Unfortunately, this imputation method does not have, in general, a clear statistical distribution. However, a possible statistical distribution for nearest-neighbour imputation can be defined for the two extreme cases: (i) the variable of interest y is highly correlated with the auxiliary variable x used to find the nearest neighbour; (ii) the variable of interest y is not correlated with the auxiliary variable x used to find the nearest neighbour. These two cases correspond to the situations used in the simulation study.

A. Case (i): *y* highly correlated with *x*

55. When the variable of interest y is highly correlated with the auxiliary variable x used to find the nearest neighbour, we can expect the imputed value $\hat{y}_k$ to be relatively close to the true value y_k. That is,

$$E_m(\hat{y}_k \,|\, s, R_1, R_2) \approx y_k \tag{14}$$

where subscript m refers to the expectation under the imputation process.

56. Under model (14), considering the last term of estimator (8), we have directly $E(\tilde{Y}) \approx Y$. This result means that if the imputation method is highly accurate, estimator (8) is nearly unbiased for Y,

irrespective of the response mechanism and the follow-up process. That is, whether or not scores are used in the follow-up process, estimator (8) remains approximately unbiased.

57. It is important to note that this assumes that the set of donors allows finding donors relatively close to the non-responding units to be imputed. For a skewed population, the right portion of the distribution (large units) should be mainly respondents, while more non-response should be in the left portion (small units). However, this last portion contains more units, and therefore, we should be able to find donors that are relatively close to the true value of the non-responding units. For very skewed populations, the closest donor can be very far from the true value when imputing for large units. This can cause serious bias in the final estimates, even if there are only a few large non-responding units.

58. In the first part of the simulation study (Fabricated Metal Product Manufacturing data), relatively low RB's were obtained when variable of interest *y* was highly correlated with the auxiliary variable *x* used. In the second part (Store Retailing data), the relatively high RB's can be mainly explained by the high skewness of the population that caused close donors to be difficult to find.

B. Case (ii): *y* is not correlated with *x*

59. When the auxiliary variable *x* used to find the nearest neighbour is not correlated with the variable of interest *y*, nearest-neighbour imputation is almost equivalent as picking a donor randomly among the set $R = R_1 \cup R_2$ of responding units. This means that

$$E_m(\hat{y}_k \,|\, s, R_1, R_2) \approx \frac{1}{n_R} \sum_{k \in R} y_k = \bar{y}_R \qquad (15)$$

60. It is important to note that in (15), donors really need to be randomly selected. If the auxiliary variable used to find a donor turns out to be moderately correlated with the variable of interest, then (15) might not hold. This might create a bias. For a much skewed population, this is even more the case.

61. Recalling that the auxiliary variables used for determining the nearest donor are *the same as the economic variables x* used in the determination of the priority scores, we also have that the variable of interest *y* is not correlated with the priority score ϕ. In addition, because $p_k^{(2)} \propto \phi_k$, we then have that the variable of interest *y* is not correlated with the response probabilities in the follow-up process.

62. Under model (15), for no specific sampling design, we can show that $E(\tilde{Y}) \neq Y$. Since sample *s* is selected using (stratified) SRSWoR, we have $\pi_k = n/N$, and in this case, we can show that $E(\tilde{Y}) \approx Y$. This result means that if the variable of interest *y* is not correlated with the priority score ϕ, nor with the auxiliary variable *x* used to find the nearest neighbour, estimator (8) is nearly unbiased for *Y*.

63. Note that using similar arguments, it can be shown that the natural estimator $\hat{Y}_{natural}^{(1+2)}$ given by (13) is also approximately unbiased for *Y*. That is, even if no imputation is performed, estimator (13) should provide, in this case, nearly unbiased estimates.

64. In the first part of the simulation study (Fabricated Metal Product Manufacturing data), relatively low RB's were obtained when the variable of interest *y* was weakly correlated with the auxiliary variable *x*. In the second part (Store Retailing data), the relatively high RB's can be mainly explained by the high skewness of the population. Because of this, even a low correlation between *x* and *y* created a substantial bias in the estimates. Note that this bias decreased rapidly as the EWRR increased because of the reduction of the amount of imputation performed.

V. Conclusions

65. In the study, seven collection methods/options associated with maximizing the EWRR are compared in terms of the quality of estimation and the cost of follow-up. As expected, follow-up costs goes up along with the increase of a targeted EWRR, and if a collection method/option is complex, it usually increases the follow-up costs.

66. Our study has shown that it is important to have a high correlation between the economic weight variables and the estimated economic variables. Compared with a low correlation, high correlation allows for a better quality of estimation as well as less variability when using different collection methods/options. As the cost increases, the RB and RRMSE usually decrease or stay relatively stable when the correlation is high, whereas they usually fluctuate greatly when the correlation is low.

67. When the population is not extremely skewed, the quality of estimation could be improved by reducing a unit's priority after several unsuccessful contacts with slightly increased cost, i.e., using Method/Option M3, especially when there is a low correlation between the economic weight variables and the estimated economic variables.

68. The follow-up cost is higher for a moderately skewed population than that for an extremely skewed population, but the quality of estimation is better for the former than for the latter.

69. The use of the random method for follow-up (i.e. Method/Option M7) increases significantly the cost of achieving a specific targeted EWRR, but it could improve the quality of estimation when there is a low correlation between the economic weight variables and the estimated economic variables, or when the population is extremely skewed.

70. The results of the study have shown that Method/Option M3 is optimal for most cases when the population is not extremely skewed, whereas Method/Option M2 is usually optimal when the population is extremely skewed. These two methods are also relatively simple. It should be noted that our results are based on data of relatively large population sizes. The effects of different collection methods could vary or not be obvious when the population size is small.

71. In the future, studies could be conducted on the use of increased auxiliary data, e.g. appointment cases, and on other collection methods that are not associated with maximizing the EWRR, e.g. methods of Probability Proportional to Size, as part of adaptive collection in order to improve further cost efficiencies.

References

Brodeur, M., Koumanakos, P., Leduc, J., Rancourt, É. and Wilson, K. (2006), *The Integrated Approach to Economic Surveys in Canada, Statistics Canada*, Ottawa (Canada). Catalogue 68-514.

Evra, R. and DeBlois, S. (2007), Using Paradata to Monitor and Improve the Collection Process in Annual Business Surveys, *Proceedings of the International Conference on Establishment Surveys III,* pp. 227-232.

Lavallée, P. and Hidiroglou, M. (1988), On the Stratification of Skewed Populations, *Survey Methodology*, Volume 14, no 1, 33-43, Statistics Canada, Ottawa (Canada).

Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Philips, R. (2003), The Theory and Application of the Score Function for Determining the Priority of Follow Up in the Annual Survey of Manufactures, *Proceedings of the Survey Methods Section, Statistical Society of Canada,* pp. 121-126.

Pursey, S. (2003), Use of the Score Function to Optimize Data Collection Resources in the Unified Enterprise, *Proceedings of the Survey Methods Section, Statistical Society of Canada,* pp. 117-120.