Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

QUALITY INDICATORS WHEN COMBINING SURVEY DATA AND ADMINISTRATIVE DATA

Pierre Lavallée[1]

ABSTRACT

At Statistics Canada, business surveys, both annual and monthly, are using administrative data at an ever increasing rate. The administrative data come predominantly from the Canada Revenue Agency. These data are not only used to build and maintain the frame, the Business Register, or to assist imputation of survey data, they are now used to replace completely or partially data for subpopulations that would have traditionally been surveyed. The survey data are either replaced directly, or modelled using the administrative data, relying on a strong correlation between the administrative data and the survey data. As such, a typical survey estimate is now based on both survey and administrative information.

Traditionally, data quality indicators reported by surveys have been the sampling variance, coverage error, non-response rate and imputation rate. To obtain an imputation rate when combining survey data and administrative data, one of the problems is to compute the imputation rate itself when only imputation rates are available from both data sources, rather than imputation flags. This paper discusses how to solve this problem.

KEY WORDS: Imputation rates, combined rates, estimated expected value.

1. INTRODUCTION

At Statistics Canada, business surveys, both annual and monthly, are using administrative data at an ever increasing rate. The administrative data come from payroll deduction accounts, income tax reports or Goods and Services Tax reports collected by the Canada Revenue Agency. These data are not only used to build and maintain the frame, the Business Register, or to assist imputation of survey data, they are now used to replace completely or partially subpopulations that would have traditionally been surveyed. The primary goal is to reduce response burden and survey cost. Another goal is to possibly improve the quality of the data. The survey data are either replaced directly, or modelled using the administrative data, relying on a strong correlation between the administrative data and the survey data. As such, a typical survey estimate is now based on both survey and administrative information. The source of the data may vary by unit and even by variable.

Because estimates are based on data from both survey and administrative sources, some typical quality measures need to be revisited. For example, how to define the response rate when a large part of the data comes from an administrative source? If half of the sample comes from administrative data, do we say that 50 percent of the data has been imputed, or that 50 percent of the data has been collected via another collection mode? How to define then the imputation rate? In order to attack this problem, a Task Force on Quality Indicators was put in place in the Business Survey Methods Division of Statistics Canada during the winter of 2004. Some conceptual thinking was done in the spring and an adjustment of concepts with existing surveys in the summer. This lead, for instance, to the documents of Houle and Lavallée (2004) and Trépanier, Julien and Kovar (2005).

In the present paper, the first part will be devoted to measuring quality in general when both survey data and administrative data are used. We will then review some traditional quality indicators. We will also describe how administrative data are typically used within Statistics Canada. We will finally go through possible approaches for measuring quality, including quality profiles.

In the second part, we will discuss combined quality measures, i.e., quality measures taking into account both survey data and administrative data, and we will then restrict the discussion on defining combined rates. First, we will

[1] Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada, pierre.lavallee@statcan.ca.

discuss the desired properties when developing a rate in a general context. Second, we will develop some concepts and definitions that will help us to develop combine rates. Third, we will propose different combined rates for the case of imputation. We will finally present two different combined imputation rates based on estimated expectations, and we will discuss properties for each rate.

2. MEASURING QUALITY

2.1 Traditional Quality Indicators

Survey estimates are affected by sampling errors and non-sampling errors to various degrees. Sampling errors occur because the complete population has not been measured by the survey. To measure sampling errors, the usual indicators are variances, standard errors, coefficients of variation and confidence intervals.

Non-sampling errors can occur at any step of the survey process. With respect to the sampling frame, non-sampling errors are usually related to coverage. We then express coverage errors in terms of *coverage rate* and bias. At the collection step, non-response is one of the non-sampling errors. It is measured in terms of *non-response rate*, bias, *imputation rate*, and imputation impact. Another type of non-sampling error is due to response errors that occur when, for example, the respondent does not provide the correct value. Response errors are measured in terms of measurement error, collection mode effects, and response bias. During the processing step, keying errors can occur and they are expressed by *keying error rate*, and editing impact. Finally, at the modelling step, non-sampling errors also exist; they correspond to the model bias and variance.

2.2 How is Administrative Data Used?

In many surveys, the use of administrative data may contribute a significant portion of the estimates (see Pelletier, 2004). Now, administrative data are not free of errors. The data are usually subject to a series of editing and imputation processes in order to make the data of sufficient quality to be usable for the production of estimates. Processes such as imputation introduce sources of variability that should not be ignored when assessing the quality of the estimates. For example, even though there are available on a census basis, some variables of tax data on incorporated enterprises are imputed up to 40%. Therefore, even if there is no sampling error associated with these data, there might be non-negligible imputation errors. The different sources should then be reported and measured in a way that properly informs users and managers about the quality of the data.

Even if administrative data contain some sources of errors, they are relatively cheap compared to surveys data. They are also often available for the complete population. They are therefore more and more considered for the production of statistical information. At Statistics Canada, administrative data are used in several ways. First, they can be used to provide estimates for a non-surveyed portion of the population. This portion corresponds usually to very small businesses that would be too costly to sample in comparison to their economic importance. Second, administrative data can be used for imputation of non-responding units. Third, data replacement (or direct substitution) of survey data can be done using administrative data. This practice has become quite important for business surveys, especially for measuring financial variables of businesses. A fourth way to use administrative data is to model part of the survey sample. Modelling is performed for those cases where direct substitution cannot be done. The process is to send one part of the sample to regular data collection, and build a model using these data to finally predict the non-collected part using administrative data. Finally, administrative data can be used in mass imputation to recreate a census micro-data file from a sample. This turns out to be useful when producing, for instance, small area estimates (see Godbout and Grondin, 2005).

2.3 Possible Approaches for Measuring Quality

The quality of an estimate can be reported in terms of data sources and/or processes used. We then give some statistics on the errors associated with each source and process. The idea is to measure the main sources of errors for each data source and process, and to report them in an appropriate way for the data users. Now, when survey data and administrative data have been combined, it can be useful to produce combined quality indicators, i.e., quality indicators that take into account errors in both the survey and administrative sources. These combined quality indicators can then be used in quality profiles.

A quality profile is a set of quality statements and measures associated to the data sources and processes of an information product. The product can be aggregated estimates, micro-data files, etc. As mentioned by the U.S. Office of Management and Budget (2001), the three main functions of quality profiles are to provide qualitative and quantitative information about total survey error and the principal components of those errors; to summarize research and information on the quality of a survey; and to give a systematic account of error sources to affect the estimates, which can then be used to direct improvement activities. For further details, one can also see Jabine (1991). In measuring quality, quality profiles offer a good compromise, compared to modelling of total error, or data confrontation with another source. Modelling total error is trying to summarize in one statistic the quality of the data by considering simultaneously all sources of error (coverage, non-response, response, etc.). In practice, this turns out to be quite difficult to compute. Also, summarising the error in one statistic produces too limited a view of the errors. Data confrontation with another source should be done if data are available, but it turns out to be only one aspect of quality profiles.

One problem with quality profiles is their lengthy write-up. They involve the description of the whole survey process together with the results of the assessment of the errors. While this is detailed and informative, most of the users do not take the time to read through the quality profile of the data that they use. One solution is then to try to describe processes and errors graphically. For example, a pie chart describing the proportions of the data sources (including, for example, the imputed portions) can be built, in which case combined measures of quality could be used.

3. COMBINED RATES

Combined quality measures are useful when considering the different data sources (survey and administrative) at the same time. One type of combined quality measures is combined rates, i.e., rates taking into account both survey data and administrative data. For simplicity, the present section will focus on computing combined rates in relation to imputation, as opposed to combined response rates or other rates.

As a first example, let us suppose that a data set has been obtained by merging the variables from a survey source and an administrative source. Some variables of the sources might have been imputed, which results in a combined data set with some imputed data. We might then be interested in obtaining a global imputation rate (all variables considered together) for this data set. A second example is one where one part of the observations of the data set comes from a survey, and the other part from an administrative source. We might want to compute an imputation rate either globally, or considering each variable separately. In both examples, this involves computing combined imputation rates.

3.1 Desired Properties

Combined imputation rates can be developed in different ways. Let τ be a combined imputation rate. The following properties should hold for τ : (i) $\tau \in [0,1]$; (ii) $\tau \geq \min(\tau_A, \tau_S)$, where τ_A and τ_S are the imputation rates for the administrative source and the survey source, respectively; (iii) $\tau = 0$ when $\tau_A = 0$ and $\tau_S = 0$; (iv) $\tau = 1$ when $\tau_A = 1$ and $\tau_S = 1$; (v) the number of observations to be used in the denominator of τ is the number of observations from the union of two sources; (vi) τ can be generalized to more than two data sources.

Most of the above properties are evident, except maybe for property (ii). This property means that the imputation rate computed for the combined sources should be at least equal to the minimum of the imputation rates in each of the survey and administration sources. Given that an imputation rate is a type of quality measure, it would not be natural to see the result of combining data being of better quality than its components.

3.2 Concepts and Definitions

Define the imputation rate in the imputation class *h* of the survey source as follows:

$$\tau_{Sh} = \frac{\sum_{i=1}^{n_{Sh}} Q_{Si}\delta_{Si}}{\sum_{i=1}^{n_{Sh}} Q_{Si}} \qquad (1)$$

where Q_{Si} is an economic weight for the survey source, $\delta_{Si} = 1$ if unit *i* is imputed in the survey source, and 0 otherwise, and n_{Sh} is the number of units in the class *h* of the survey source. Similarly, define the imputation rate in the imputation class *g* of the administrative source by the following:

$$\tau_{Ag} = \frac{\sum_{i=1}^{n_{Ag}} Q_{Ai}\delta_{Ai}}{\sum_{i=1}^{n_{Ag}} Q_{Ai}} \qquad (2)$$

where Q_{Ai} is an economic weight for the administrative source, $\delta_{Ai} = 1$ if unit *i* is imputed in the administrative source, and 0 otherwise, and n_{Ag} is the number of units in the class *g* of the administrative source. Note that the economic weights are used to reflect the importance of each unit entering into the imputation rates.

For the combined data set, the traditional imputation rate τ_{unit} would have the following form:

$$\tau_{\text{unit}} = \frac{\sum_{i=1}^{n} Q_i\delta_i}{\sum_{i=1}^{n} Q_i} \qquad (3)$$

where $\delta_i = 1$ if unit *i* has been imputed, and 0 otherwise, Q_i is an economic weight associated to the combined data set, and *n* is the total number of units in the combined data set. This imputation rate represents the proportion of units imputed in at least one source. It should be noted that $\delta_i = \max(\delta_{Ai}, \delta_{Si})$, and thus, in order to compute the rate (3), we need to have the imputation flags δ_{Si} and δ_{Ai} from each of the survey and administrative sources. If they are available, it is clear that the best way to compute the combined imputation rate is to use equation (3). Unfortunately, the imputation flags are not always available and we then need to define a combined imputation rate using a different approach.

Let us suppose that the imputation rates τ_{Ag} and τ_{Sh} are available from each of the sources, but not the imputation flags δ_{Si} and δ_{Ai}. This is a situation often encountered in practice since the produced microdata files do not usually carry the imputation flags, but some imputation rates are provided in the documentation. In order to develop a combined rate, one can use the estimated expected value $\hat{E}(\delta_i)$, instead of the imputation flag δ_i itself. With $E(\delta_i) = P(\delta_i = 1) = p_i$, rather than considering if unit *i* has been imputed or not, we work with the probability p_i that the unit has been imputed. For applying the new rate in practice, we estimate p_i by some estimator $\hat{p}_i$. The *expectation-based imputation rate* is then defined as:

$$\tau_{\exp} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i \qquad (4)$$

Before estimating the imputation probability p_i, we first need to define the *unit rates* τ_{Ai} and τ_{Si} from each administrative and survey sources to be associated to unit *i*. Let $\tau_{Ai} = \tau_{Ag}$ for $i \in$ imputation class *g*, and let $\tau_{Si} = \tau_{Sh}$ for $i \in$ imputation class *h*. Both τ_{Ai} and τ_{Si} can be justified using the concept of imputation probabilities (or the probabilities of being imputed). We can indeed see τ_{Ai} and τ_{Si} as estimates of the probabilities p_{Ai} and p_{Si} of unit *i* to be imputed in the administrative and the survey sources, respectively. These probabilities p_{Ai} and p_{Si} of unit *i* are then estimated using the imputation rates τ_{Ag} for $i \in$ imputation class *g*, and τ_{Sh} for $i \in$ imputation class *h*. Note that the same discussion can be done in the context of response probabilities, where the response rates are the estimated response probabilities obtained through Response Homogeneity Groups (see Särndal, Swensson and Wretman, 1992).

There are at least two different approaches to estimate the imputation probability p_i with the units rates τ_{Ai} and τ_{Si}. These two approaches are described in the next sections.

3.3 Combined Rates Assuming Independent Imputation

Let us assume that imputation is done independently in each of the two sources. More precisely, it is assumed that if a unit *i* is imputed in the administrative source, it may or may not be imputed in the survey source. The probability that unit *i* is imputed in either source is then given by

$$p_i = 1 - (1 - p_{Ai})(1 - p_{Si}) \tag{5}$$

where p_{Ai} and p_{Si} are the probabilities that unit *i* is imputed in the administrative source and the survey source, respectively. To estimate p_{Ai}, one can simply use the imputation rates computed in the *G* imputation classes of the administrative source, i.e., $\hat{p}_{Ai} = \tau_{Ai} = \tau_{Ag}$ for $i \in g$, *g*=1,...,*G*. Similarly, for the *H* imputation classes of the survey source, $\hat{p}_{Si} = \tau_{Si} = \tau_{Sh}$ for $i \in h$, *h*=1,..., *H*. Using these estimators, we have

$$\hat{p}_i = 1 - (1 - \tau_{Ai})(1 - \tau_{Si}) \tag{6}$$

Using (5) and (6), we then get

$$\begin{aligned} \tau_{\text{indep}} &= 1 - \frac{1}{n}\sum_{i=1}^{n}(1 - \tau_{Ai})(1 - \tau_{Si}) \\ &= 1 - \frac{1}{n}\sum_{g=1}^{G}\sum_{h=1}^{H} n_{gh}(1 - \tau_{Ag})(1 - \tau_{Sh}) \end{aligned} \tag{7}$$

The information required to produce this combined rate is the imputation rate in each of the classes from the two sources, as well as the size of the classes n_{gh} obtained by crossing the two sources. It can happen in practice that the sizes n_{gh} (and therefore n_{gh}/n) are unknown. In this case, one can replace in (7) the ratios n_{gh}/n by some class weights ω_{gh} where $\omega_{gh} \in]0,1[$ and $\sum_{g=1}^{G}\sum_{h=1}^{H}\omega_{gh} = 1$. This yields

$$\tau^{*}_{\text{indep}} = 1 - \sum_{g=1}^{G}\sum_{h=1}^{H}\omega_{gh}(1 - \tau_{Ag})(1 - \tau_{Sh}) \tag{8}$$

The class weights ω_{gh} can be chosen arbitrarily. For example, one can choose $\omega_{gh} = 1/(G \times H)$ for $g = 1,...,G$ and $h = 1,...,H$. It should finally be noted that, in general, $\tau^{*}_{\text{indep}} \neq \tau_{\text{indep}}$.

3.4 Combined Rates Assuming Dependent Imputation

Let us now assume that imputation is done in some dependent way between the two sources. That is, it is assumed that if a unit *i* is imputed in the administrative source, it is likely to be imputed in the survey source, and vice versa. For a given unit *i*, this can be translated into the following probability:

$$p_i = \max(p_{Ai}, p_{Si}) \tag{9}$$

The rationale behind (9) can come from Poisson Sampling. Suppose that unit *i* has been imputed in the administrative source. In a sampling perspective, this would correspond to "selecting" unit *i* for imputation according to a Bernoulli trial with probability p_{Ai}. For this selection, we generate a random number $u_i \sim U(0,1)$ that we compare to p_{Ai} and unit *i* is "chosen" to be imputed if $u_i \le p_{Ai}$. Now, if we assume that imputation is not independent between the two sources, this can be translated into keeping the random number u_i for the Bernoulli trial within the survey source. That is, the same random number u_i is compared to p_{Si} and unit *i* is "chosen" to be imputed in the survey source if $u_i \le p_{Si}$. According to this "selection" process, the probability that unit *i* be imputed in either source is given by (9).

Using $\hat{p}_{Ai} = \tau_{Ai} = \tau_{Ag}$ for $i \in g$, and $\hat{p}_{Si} = \tau_{Si} = \tau_{Sh}$ for $i \in h$, in equation (4), we then get the following combined rate:

$$\begin{aligned}\tau_{\text{depend}} &= \frac{1}{n}\sum_{i=1}^{n} \max(\tau_{Ai}, \tau_{Si}) \\ &= \frac{1}{n}\sum_{g=1}^{G}\sum_{h=1}^{H} n_{gh} \max(\tau_{Ag}, \tau_{Sh})\end{aligned} \tag{10}$$

Again, it can happen in practice that the sizes n_{gh} (and n_{gh} / n) are unknown. In this case, one can replace in (10) the ratios n_{gh} / n by some class weights ω_{gh} where $\omega_{gh} \in]0,1[$ and $\sum_{g=1}^{G}\sum_{h=1}^{H}\omega_{gh} = 1$. The gives

$$\tau_{\text{depend}}^{*} = \sum_{g=1}^{G}\sum_{h=1}^{H} \omega_{gh} \max(\tau_{Ag}, \tau_{Sh}) \tag{11}$$

As for the case of independent imputation, in general, $\tau_{\text{depend}}^{*} \ne \tau_{\text{depend}}$.

3.5 Weighted Versions

The combined imputation rates (3), (7) and (10) can also be developed in weighted versions using estimation weights. When using the estimation weights, the computed imputation rates can be seen as estimates of what these rates might have been if all the units of the population were surveyed.

The weighted version of τ_{unit} is given by the following:

$$\tau_{\text{unit},w} = \frac{\sum_{i=1}^{n} w_i Q_i \delta_i}{\sum_{i=1}^{n} w_i Q_i} \tag{12}$$

where w_i is the estimation weight obtained from the administrative source, the survey source, or a combination of the two sources. A weighted version of τ_{indep} is given by the following:

$$\tau_{\text{indep},w1} = 1 - \left[\frac{\sum_{i=1}^{n} w_i (1-\tau_{Ai})(1-\tau_{Si})}{\sum_{i=1}^{n} w_i}\right] \tag{13}$$

A second weighted version of τ_{indep} can also be given by the following:

$$\tau_{\text{indep},w2} = 1 - \frac{1}{n}\sum_{i=1}^{n}(1-\tau_{Awi})(1-\tau_{Swi}) \tag{14}$$

where $\tau_{Awi} = \tau_{Awg}$ for $i \in$ class *g*, with $\tau_{Awg} = \dfrac{\sum_{i=1}^{n_{Ag}} w_{Ai} Q_{Ai} \delta_{Ai}}{\sum_{i=1}^{n_{Ag}} w_{Ai} Q_{Ai}}$, and where w_{Ai} is the estimation weight obtained from the administrative source. Similarly for τ_{Swi}.

The first weighted version $\tau_{\text{indep},w1}$ of the combined imputation rate can be viewed as an estimation of the imputation rate as if the *n* units involved would represent a census. The second weighted version of the combined imputation rate is a mixture of the weighted imputation rates of both administrative and survey sources. Weighted versions of the combined rates assuming dependent imputation ($\tau_{\text{depend},w1}$ and $\tau_{\text{depend},w2}$) can be obtained in a similar way as (13) and (14), respectively.

3.6 Graphical Comparison of τ_{indep} and τ_{depend}

In this section, an example of the results that would be obtained with τ_{indep} and τ_{depend} is illustrated graphically. Figure 1 presents the results obtained for τ_{indep} with τ_{Ai} and τ_{Si} varying between 0 and 1. Figure 2 shows the results obtained for τ_{depend}. Finally, Figure 3 illustrates the difference between the two rates, i.e., $\tau_{indep} - \tau_{depend}$. As it can be seen, τ_{indep} is a smooth function while τ_{depend} is not. Also, $\tau_{indep} \geq \tau_{depend}$ and the maximum difference (0.25) between the two is attained when τ_{indep} and τ_{depend} are both at 0.5. The difference is relatively large when $\tau_{Ai} = \tau_{Si}$, but it decreases rapidly when $\tau_{Ai} \neq \tau_{Si}$.

Figure 1. τ_{indep}

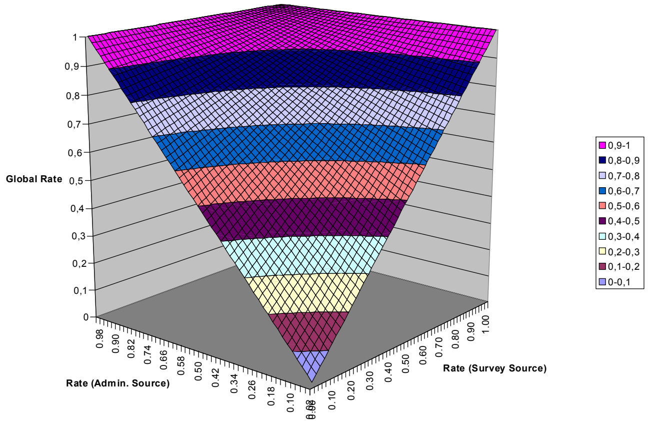


Figure 2. τ_{depend}

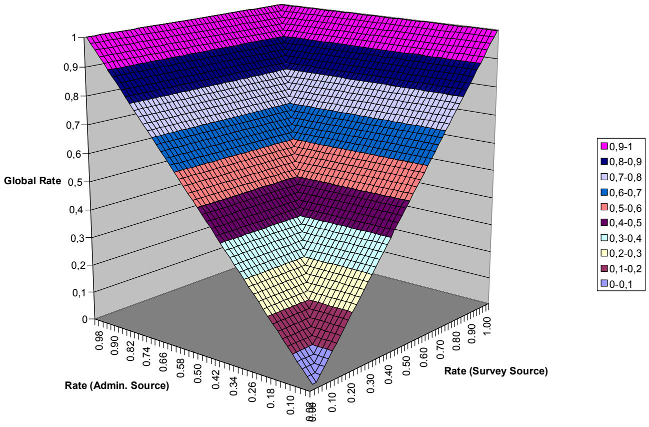
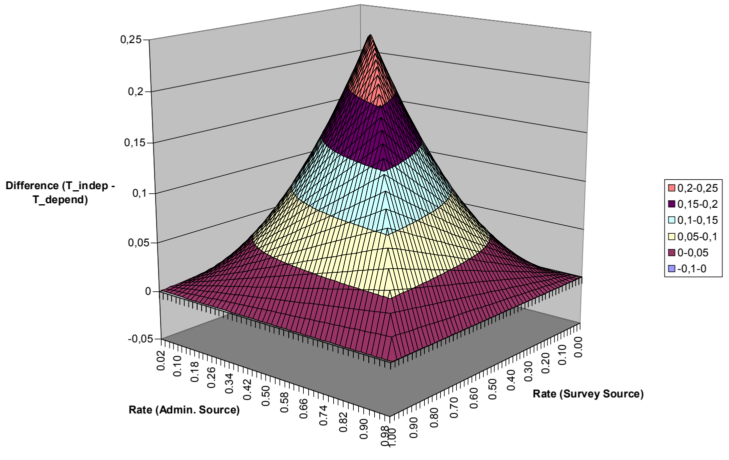


Figure 3. Difference between τ_{indep} and τ_{depend}



AKNOWLEDGEMENT

Thanks are due to John Kovar, Hélène Bérard and Julie Trépanier who were part of the Task Force on Quality Indicators that worked on the problem of defining quality measures when both survey and administrative data are used. A special thank you to Anne-Marie Houle that produced the graphs. Without our fruitful discussions, none of this work would have been possible.

REFERENCES

Godbout, S. and Grondin, C. (2005), "*Contourner une différence de concepts entre deux sources de données pour la production d'estimations*", Paper presented at the Colloque francophone sur les sondages, Québec, May 24-27, 2005.

Houle, A.-M. and Lavallée, P. (2004), "*BSMD Task Force on Quality Indicators — Survey of Employment, Payroll, and Hours*", Statistics Canada internal document, November 20th, 2004.

Jabine, T. (1991), "The SIPP Quality Profile", *Seminar on the Quality of Federal Data, Part 1 of 3*, Statistical Policy Working Paper 20, U.S. Office of Management and Budget, Washington, DC, pp. 19-28.

Pelletier, É. (2004), "*L'utilisation accrue des données fiscales dans le cadre de l'Enquête unifiée sur les entreprises*", Paper presented at the annual conference of the Statistical Society of Canada, Montréal.

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

Trépanier, J., Julien, C. and Kovar, J. (2005), "*Reporting Response Rates when Survey and Administrative Data are Combined*", Paper presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, Virginia, November 14-16, 2005.

U.S. Office of Management and Budget (2001), "*Measuring and Reporting Sources of Errors in Surveys*", Statistical Policy Working Paper 31, Subcommittee on Measuring and Reporting the Quality of Survey Data of the Federal Committee on Statistical Methodology, June 2001.