Survey Methodology, December 1995
Vol. 21, No. 2, pp. 125–135
Statistics Canada

Alternative Adjustments Where There Are Several Levels of Auxiliary Information

F. DUPONT[1]

ABSTRACT

Regression estimation and its generalization, calibration estimation, introduced by Deville and Särndal in 1993, serves to reduce *a posteriori* the variance of the estimators through the use of auxiliary information. In sample surveys, there is often useable supplementary information that is distributed according to a complex schema, especially where the sampling is realized in several phases. An adaptation of regression estimation was proposed along with its variants in the framework of two-phase sampling by Särndal and Swensson in 1987. This article seeks to examine alternative estimation strategies according to two alternative configurations for auxiliary information. It will do so by linking the two possible approaches to the problem: use of a regression model and calibration estimation.

KEY WORDS: Auxiliary information; Regression estimator; Calibration estimator; Two-phase sampling.

1. INTRODUCTION

Using the regression estimator studied by Fuller (1975), Cassel, Särndal and Wretman (1976), Särndal (1980), Gourieroux (1981), Isaki and Fuller (1982), and Wright (1983), it is possible to improve *a posteriori* – that is, after the sampling has been completed – the estimate of a total of a variable of interest on the basis of auxiliary variables $x_1, \ldots, x_k$ for which additional information is available. The variance in relation to the Horwitz-Thompson estimator is reduced by using the regression estimator, provided that one knows the true value of the target population totals of the auxiliary variables, which will constitute the additional information referred to as auxiliary information. Deville and Särndal in 1992 proposed a class of estimators derived from a reweighting approach that addresses the same issue of variance reduction: calibration estimators. By calibrating sampling weights it is possible to incorporate *a posteriori* auxiliary information of the type totals $X_1, \ldots, X_k$ of k variables $x_1, \ldots, x_k$ into the estimate made on the basis of the new weightings and thus to improve the estimate. This approach generalizes regression estimation, which is one of the elements of the class.

However, in surveys based on sampling, there is often usable additional information that is distributed according to a more complex schema than what has been described above, especially when the sampling is carried out in several phases. This article looks at different strategies for using this complex auxiliary information in the framework of two-phase sampling, with the possibility of generalizing to more than two phases.

When the sampling plan entails two phases, the auxiliary information consists of information known for the entire population, but also of information known for the sample resulting from the first sampling phase. These two bodies of information may concern different variables.

In their 1987 article, Särndal and Swensson propose an estimator that uses all the auxiliary information available for a two-phase sampling, with different auxiliary information for the total population and the population obtained from the first-phase sampling. This is an estimator adapting the principle of the regression estimator when the information known for the individuals obtained from the first-phase sampling is considered to be substitutable for the aggregated information and to be of better quality than the information available for the target population as a whole, for purposes of estimating the variable of interest. However, in practice it often happens that these two bodies of information are complementary rather than substitutable. We have thus sought in this study to develop the regression estimate in a context in which the bodies of auxiliary information are complementary.

Furthermore, insofar as calibration estimation generalizes regression estimation when the auxiliary information is at only one level, we have sought to adapt calibration estimation to this context. We review the various calibration strategies in order to propose the most suitable ones, seeking to relate them to generalizations of regression estimation that are possible in this context.

We show (Section 2) that the joint use of two different bodies of auxiliary information leads to two regression models and three associated decompositions of the variable of interest. The regression model assisted approach (RMAA) thus enables us to derive 3 alternative estimators.

In turn, the calibration approach (CA) (Section 3) enables us to derive 4 estimators. Each of these estimators may be related to (associated with) the three estimators derived from the regression model approach.

[1] F. Dupont, Unité Méthodes Statistiques, Institut National de la Statistique et des Études Économiques, (INSEE), 18 Blvd. Adolphe Pinard, 75675 Paris Cedex.

Thus (Section 4), the two approaches may be linked together and result in three classes of estimators, each associated with a decomposition of the variable of interest. The estimators of a given class have the same asymptotic variance.

When strategies are evaluated on the basis of the sampling plan alone, our choice is directed toward the third class of estimators, which is superior to the other two from the standpoint of variance.

When strategies are evaluated on the basis of a modelling of the variable of interest, the preferable class of estimators is the one associated with the modelling adopted.

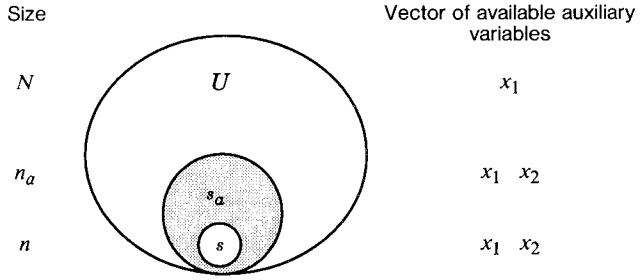
In a situation in which we wish to adjust a survey, and in which we wish simultaneously to correct the biases that would result from the use of gross weightings and to reduce the variance, the findings must be adapted: the changes introduced in the weightings to correct the biases are greater than the corrections for variance reduction. Hence the variables will be incorporated into the calibration once it appears that they are affecting the probability of selection and thus participating in the creation of the bias.

When the auxiliary variables are qualitative, the choice between *a priori* and *a posteriori* use of the auxiliary information – that is, between its use at the sampling stage and at the adjustment stage – still rests on the distinction between the two modellings of the variable of interest.

These findings may be extended to samplings of more than two phases.

2. NOTATIONS

The framework is that of a two-phase sampling. Assume that auxiliary information is available at two different levels: the target population and the population obtained from the first-phase sampling. The situation may be diagrammed as follows:



where U represents the target population for which the values of the vector of variable x_1 are known or, failing that, the total $X_1 = \sum_{i \in U} x_{i1}$. s_a represents an intermediate level of sampling for which the values of the vectors of k_1 variables x_1 and k_2 variables x_2 are known for all individuals. We denote as π_{ia} the probability of selection from the sample associated with the first phase of the sampling. s represents the final sample for which are available the values of the variable y, the total of which we are trying to estimate, as well as the values of the vectors of the auxiliary variables x_1 and x_2. This is denoted as $\pi_i = P(i \mid s_a)$.

We hope to make optimum use of all this auxiliary information in order to improve the estimates that will be made on the basis of the data gathered from the sample that results from the second sampling phase s.

An obvious first idea is to try to generalize the regression estimator in this context.

3. REGRESSION ESTIMATION APPROACH

3.1 The Information Contained in x_1 is Considered to be Substitutable for the Information Contained in x_2 for Estimating y and to be of Lesser Quality

In their work, Särndal, Swensson and Wretman propose the following regression estimator for estimating the total of y:

$$\hat{Y}_1 = \sum_{i \in s} \frac{y_i}{\pi_i \pi_{ai}} + \left(\sum_{i \in s_a} \frac{x'_{i2} \hat{b}_2}{\pi_i} - \sum_{i \in s} \frac{x'_{i2} \hat{b}_2}{\pi_i \pi_{ai}} \right) + \left(\sum_{i \in U} x'_{i1} \hat{b}_1 - \sum_{i \in s_a} \frac{x'_{i1} \hat{b}_1}{\pi_{ai}} \right)$$

where the second term is the correction for poor estimation on s_a and the third is the correction for poor estimation on s.

The estimation can also be written:

$$\hat{Y}_1 = \sum_{i \in U} x'_{i1} \hat{b}_1 + \sum_{i \in s_a} \frac{(x'_{i1} \hat{b}_1 - x'_{i2} \hat{b}_2)}{\pi_{ai}} + \sum_{i \in s} \frac{(y_i - x'_{i2} \hat{b}_2)}{\pi_i \pi_{ai}}$$

where the second term is the correction for poor approximation of y_i by $x'_{i1} \hat{b}_1$ and the third is the correction for poor approximation of y_i by $x'_{i2} \hat{b}_2$;

with $$\hat{b}_1 = \left(\sum_{i \in s_a} \frac{x_{i1} x'_{i1}}{\pi_{ai}} \right)^{-1} \left(\sum_{i \in s} \frac{x_{i1} y_i}{\pi_i \pi_{ai}} \right)$$

and $$\hat{b}_2 = \left(\sum_{i \in s} \frac{x_{i2} x'_{i2}}{\pi_i \pi_{ai}} \right)^{-1} \left(\sum_{i \in s} \frac{x_{i2} y_i}{\pi_i \pi_{ai}} \right).$$

The underlying idea is that we have two concurrent models for y, namely:

(1) $y_i = x_{i1}' b_1 + u_{i1}$ with $E(u_{i1}) = 0$ and $V(u_{i1}) = \sigma_1^2$

and

(2) $y_i = x_{i2}' b_2 + u_{i2}$ with $E(u_{i2}) = 0$ and $V(u_{i2}) = \sigma_2^2$

the second of which we believe is *a priori* better for predicting the value of y_i. Thus in this model-based perspective, x_1 functions as a proxy of x_2. A situation of this type corresponds, for example, to a case in which x_2 represents the update - that is, the update to the date of the survey - of the variable retrieved from the x_1 sampling frame. In other words, if x_2 were available at the level of the entire population, the estimator used would be

$$\sum_{i \in U} x_{i2}' \hat{b}_1 + \sum_{i \in s} \frac{(y_i - x_{i2}' \hat{b}_2)}{\pi_i \pi_{ai}}.$$

Let us now imagine the case of a two-phase sampling survey of households. Assume that the sampling frame is made up of dwellings for which we have information consisting of dwelling size, denoted as x_1, which is therefore known for all individuals in the target population. If all the individuals obtained from the first sampling phase are questioned on the composition of the household, denoted as x_2, in particular on the number of children in the household, the two bodies of information appear to be complementary rather than substitutable for purposes of studying the household budget. This is further reinforced if instead of household composition, the information collected is the age or occupation of the head of household.

In a model-based perspective, the alternative situation, in which the information contained in x_1 is considered complementary to that contained in x_2 for estimating y, thus naturally suggests itself.

3.2 The Information Contained in x_1 is Considered to be Complementary to the Information Contained in x_2 for Estimating y

3.2.1 Decomposition $y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$

The underlying model is then:

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i \text{ with } E(u_i) = 0 \text{ and } V(u_i) = \sigma_1^2.$$

The estimator to be used is then:

$$\hat{Y}_2 = \sum_{i \in U} x_{i1}' \hat{a}_1 + \sum_{i \in s_a} \frac{x_{i2}' \hat{a}_2}{\pi_{ai}} + \sum_{i \in s} \frac{(y_i - x_{i1}' \hat{a}_1 - x_{i2}' \hat{a}_2)}{\pi_i \pi_{ai}}$$

with:

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i \in s} \frac{x_{i1} x_{i1}'}{\pi_{ai} \pi_i} & \sum_{i \in s} \frac{x_{i1} x_{i2}'}{\pi_{ai} \pi_i} \\ \sum_{i \in s} \frac{x_{i2} x_{i1}'}{\pi_{ai} \pi_i} & \sum_{i \in s} \frac{x_{i2} x_{i2}'}{\pi_{ai} \pi_i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in s} \frac{x_{i1} y_i}{\pi_{ai} \pi_i} \\ \sum_{i \in s} \frac{x_{i2} y_i}{\pi_{ai} \pi_i} \end{pmatrix}.$$

The variable here is broken down into three components $y_i = x_{i1}' \hat{a}_1 + x_{i2}' \hat{a}_2 + \hat{u}_i$. The total of y is thus broken down into three components, each of which is estimated at the highest level, that is, with the greatest precision possible:

- U for $x_{i1}' \hat{a}_1$,
- s_a for $x_{i2}' \hat{a}_2$, and
- s for $\hat{u}_i$.

3.2.2 Decomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$

If we wish to make maximum use of the information contained in x_1 available on U, it is natural to introduce another formulation of the same model $y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$ which isolates everything which in y can be taken into account through x_1. It is written as follows:

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i \quad \text{with}$$

$$E(u_i) = 0 \quad \text{and} \quad V(u_i) = \sigma_1^2,$$

where M_{x_1} represents the orthogonal projection, in the metric associated with the weights $1/\pi_{ai}$, on the orthogonal of the vector space generated in s_a (similar to $\Re^n$) by the group of variables x_1.

$M_{x_1}(x_{i2})$ is defined by:

$$M_{x_1}(x_{i2}) = x_{i2}' - \left(\sum_{i \in s_a} \frac{x_{i2} x_{i1}'}{\pi_{ai}} \right) \left(\sum_{i \in s_a} \frac{x_{i1} x_{i1}'}{\pi_{ai}} \right)^{-1} x_{i1}.$$

The associated natural estimator is then:

$$\hat{Y}_3 = \sum_{i \in U} x_{i1}' \hat{c}_1 + \sum_{i \in s_a} \frac{(M_{x_1} x_{i2}' \hat{c}_2)}{\pi_{ai}} + \sum_{i \in s} \frac{(y_i - x_{i1}' \hat{c}_1 - M_{x_1} x_{i2}' \hat{c}_2)}{\pi_i \pi_{ai}},$$

where $\hat{c} = \begin{pmatrix} c_1 \\ \hat{c}_2 \end{pmatrix}$ is the regression coefficient $y = x' c_1 + (M_{x_1} x_2)' c_2 + u$ estimated over s with weights $1/\pi_{ai} \pi_i$ (which differs slightly from $\begin{pmatrix} \hat{b}_1 \\ \hat{a}_2 \end{pmatrix}$).

3.3 The Three Estimators Derived from the Model-based Approach

The modelling approach has enabled us to construct 3 estimators that can be rewritten synthetically by introducing

new notations. Throughout what follows, for a vector of a given variable z, the following notation will be used:

$$\hat{\hat{Z}} = \sum_{i \in s} \frac{1}{\pi_{ai}\pi_i} z$$

$$\hat{Z} = \sum_{i \in s_a} \frac{1}{\pi_{ai}} z .$$

With these notations, the three estimators are rewritten as follows:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2' \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2' \hat{b}_2]$$

associated with the models:

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1}$$

and

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$$\hat{Y}_2 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$$

associated with the model

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$$\hat{Y}_3 = [X_1' \hat{c}_1] + [M_{x_1} \hat{X}_2' \hat{c}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{\hat{X}}_2' \hat{c}_2]$$

associated with the model

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i .$$

In the same manner as the regression estimator is generalized by calibration estimators, the problem of the use of auxiliary information at several levels may be dealt with through calibration theory, be attempting to construct calibration strategies adapted to the auxiliary information configuration examined in this article.

4. CALIBRATION APPROACH

4.1 Different Strategies Possible

When we try to generalize the calibration estimate proposed in a context in which auxiliary information is present at a single level - that of the entire population - several strategies naturally suggest themselves:

Strategy 1

a) calibrate the structure of the 1st-phase sample s_a on that of the total population U in terms of variable x_1, then,

b) calibrate the structure of the 2nd-phase sample s on that of the 1st phase sample s_a in terms of variable x_2.

Note: For the latter operation, it is better to take account of the preceding calibration in terms of x_1 in order to determine the reference value in the calibration in terms of x_2 on s_a. If the preceding calibration is not taken into account, only the estimates made at the level of s_a will benefit from the improvement made by stage a. A good way to convince oneself of this is to consider the specific extreme case where $x_1 = x_2$.

This strategy corresponds to the following calibration equations:

Stage a:

$$\sum_{i \in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i \in U} x_{i1} = X_1$$

which determines β_1, then

Stage b:

$$\sum_{i \in s} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} F(x_{i2}'\beta_2) x_{i2} = \sum_{i \in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*$$

which determines β_2,

where F designates, as throughout this article, the function which is used in the calibration and which may be linear, exponential, truncated linear or logit (see Deville, Särndal 1993).

Strategy 2

Calibrate the structure of the 2nd-phase sample s simultaneously in terms of variables x_1 and x_2, that is,

- on the structure of the total population U as regards x_1
- on the structure of s_a for x_2.

This second strategy leads us to the following calibration equations:

$$\sum_{i \in s} \frac{F(x_{i1}\alpha_1 + x_{i2}\alpha_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i \in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i \in s} \frac{F(x_{i1}'\alpha_1 + x_{i2}'\alpha_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i \in s_a} \frac{x_{i2}}{\pi_{ai}} = \hat{X}_2,$$

which determines α_1 and α_2.

The first strategy offers the advantage of correcting the 1st phase weightings, that is, of incorporating the auxiliary information at the highest level. The second strategy, for its part, makes it possible to correct the weightings that will actually be used in the estimation, and in particular to obtain a perfect estimate of the total of x_1.

A third strategy may be proposed; it combines the advantages of the above two strategies and would therefore seem preferable to them:

Strategy 3

a) calibrate the structure of the 1st phase sample s_a on that of the total population U in terms of variable x_1, then

b) calibrate the structure of the 2nd phase sample s simultaneously in terms of variables x_1 and x_2, that is,
   - on the structure of the total population U as regards x_1
   - on the structure of s_a modified by taking account of the preceding calibration for x_2.

This strategy leads to the following calibration equations:

Stage a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1}$$

which determines β_1, then

Stage b:

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

which determines γ_1 and γ_2.

Lastly, a fourth strategy may be proposed; it may be seen as a variant of the preceding strategy:

Strategy 4

a) calibrate the structure of the 1st phase sample s_a on that of the total population U in terms of variable x_1, then

b) calibrate the structure of the 2nd phase sample s simultaneously in terms of variables x_1 and x_2, on the basis of the weights modified by the preceding calibration, that is,
   - on the structure of the total population U as regards x_1
   - on the structure of s_a modified taking account of the preceding calibration for x_2.

This strategy leads to the following calibration equations:

Stage a:

$$\sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i1} = \sum_{i\in U} x_{i1},$$

which determines β_1, then

Stage b:

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i1} = \sum_{i\in U} x_{i1} = X_1, \quad \text{and}$$

$$\sum_{i\in s} \frac{F(x_{i1}'\beta_1)F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i} x_{i2} = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} x_{i2} = \hat{X}_2^*,$$

which determines δ_1 and δ_2.

When the calibration function is exponential, it is clear that strategies 3 and 4 coincide.

In this calibration-based approach, the viewpoint adopted is that of reduction of variance based on the characteristics of the sampling plan, without consideration of the model. Two questions then naturally arise:

- Can each of these four strategies be linked to a model-based approach?
- Can these four strategies be compared in terms of variance?

We will first examine the link between the three strategies defined by a calibration approach and the strategies defined by a model-based or regression approach, after which we will focus on calculating the variances of the estimators associated with each of the strategies.

4.2 Link Between the Different Possible Strategies and the Regression Approach

When F is linear, each of the estimators associated with the four strategies may be rewritten simply.

Notations

Throughout the rest of this article we will use the following notations for a vector of any variable z:

$$\hat{\hat{Z}}^* = \sum_{i\in s} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} z_i \qquad \hat{Z}^* = \sum_{i\in s_a} \frac{F(x_{i1}'\beta_1)}{\pi_{ai}} z_i.$$

We will also omit the i indexes in order to lighten the presentation when there is no ambiguity.

Strategy 1

The weightings are of the form

$$w_i^4 = \frac{F(x_{i1}'\beta_1)}{\pi_{ai}\pi_i} F(x_{i2}'\beta_2),$$

the associated estimator $\hat{Y}_4$ may be rewritten by translating the effect of the second calibration on x_2:

$$\hat{Y}_4 = \hat{\hat{Y}}^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{B}_2 \quad \text{with}$$

$$\hat{B}_2 = \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a \pi} x_2 x_2' \right)^{-1} \left(\sum_s \frac{F(x_1'\beta_1)}{\pi_a \pi} x_2 y \right),$$

then by translating the effect of the first calibration on x_1:

$$\hat{Y}_4 = \hat{\hat{Y}} + [X_1 - \hat{X}_1]' \hat{B}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{B}_2,$$

or:

$$\hat{Y}_4 = [X_1' \hat{B}_1] + [\hat{X}_2^{*\prime} \hat{B}_2 - \hat{X}_1' \hat{B}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime} \hat{B}_2],$$

$$\text{with} \quad \hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a \pi} \right).$$

Now, $\hat{Y}_1$ is rewritten:

$$\hat{Y}_1 = [X_1' \hat{b}_1] + [\hat{X}_2^{*\prime} \hat{b}_2 - \hat{X}_1' \hat{b}_1] + [\hat{\hat{Y}} - \hat{\hat{X}}_2^{*\prime} \hat{b}_2].$$

We thus obtain an estimator similar to the estimator $\hat{Y}_1$ that is obtained from the model-based approach in cases where the information contained in x_1 is considered to be substitutable for the information contained in x_2 for estimating y and also to be of lesser quality. The differences between $\hat{Y}_1$ and $\hat{Y}_4$ concern the following points:

1. $\hat{B}_2$ is estimated by incorporating the changes from the calibration on x_1, unlike $\hat{b}_2$.
2. The estimate $\hat{B}_1 = \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_s \frac{x_1 y}{\pi_a \pi} \right)$ of B_1, is made in part on s_a, unlike $\hat{b}_1$.
3. Lastly, we use the adjusted weights $F(x_1\beta_1)/\pi_a\pi$ in the sums in x_2 on s and on S_a in $\hat{Y}_4$ in unlike what was done for $\hat{Y}_1$: the estimation on x_2 is improved by the knowledge of x_1.

Thus the underlying modelling here is indeed: (1) $y_i = x_{i1}' b_1 + u_{i1}$ and (2) $y_i = x_{i2}' b_2 + u_{i2}$, the second of which we think is *a priori* better for predicting the value of y_i.

Strategy 2

We obtain weights

$$w_i^5 = \frac{F(x_{i1}'\alpha_1 + x_{i2}'\alpha_2)}{\pi_{ai}\pi_i},$$

the associated estimator is rewritten as follows:

$$\hat{Y}_5 = [X_1' \hat{a}_1] + [\hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

We thus obtain exactly the estimator $\hat{Y}_2$ proposed in the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y. The underlying model here is indeed $y_i = x_{i1}a_1 + x_{i2}a_2 + u_i$.

Strategy 3

We obtain weights

$$w_i^6 = \frac{F(x_{i1}'\gamma_1 + x_{i2}'\gamma_2)}{\pi_{ai}\pi_i},$$

the associated estimator is rewritten as:

$$\hat{Y}_6 = \hat{\hat{Y}} + [X_1 - \hat{\hat{X}}_1]' \hat{a}_1 + [\hat{X}_2^* - \hat{\hat{X}}_2]' \hat{a}_2$$

thus:

$$\hat{Y}_6 = [X_1' \hat{a}_1] + [\hat{X}_2^{*\prime} \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2].$$

Now,

$$\hat{X}_2^* = \sum_{s_a} \frac{x_2}{\pi_a} + \left(\sum_{s_a} \frac{x_2 x_1'}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right) \left[X - \sum_{s_a} \frac{x_1}{\pi_a} \right].$$

From this it can be deduced by replacing in $\hat{Y}_6$ that:

$$\hat{Y}_6 = [X_1' \hat{C}_1] + [M_{x_1} \hat{X}_2' \hat{a}_2] + [\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2],$$

with

$$\hat{C}_1 = \hat{a}_1 + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a} \right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a} \right) \hat{a}_2.$$

We thus obtain an estimator that is close to the estimator $\hat{Y}_3$ proposed in the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y. The underlying model here is $y_i = x_{i1}'c_1 + M_{x_1}(x_{i2})'c_2 + u_i$. The differences between $\hat{Y}_3$ and $\hat{Y}_6$ concern the estimated coefficients: $\begin{pmatrix} \hat{C}_1 \\ \hat{a}_2 \end{pmatrix}$ differs slightly from $\begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix}$ and $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{a}_1 - \hat{\hat{X}}_2' \hat{a}_2]$ differs slightly from $[\hat{\hat{Y}} - \hat{\hat{X}}_1' \hat{c}_1 - M_{x_1} \hat{\hat{X}}_2' \hat{c}_2]$. On the other hand, these quantities are asymptotically equivalent.

Strategy 4

We obtain weights

$$w_i^7 = \frac{F(x_{i1}'\beta_1) F(x_{i1}'\delta_1 + x_{i2}'\delta_2)}{\pi_{ai}\pi_i},$$

the associated estimator is rewritten as follows:

$$\hat{Y}_7 = \hat{\hat{Y}}^* + [X_1 - \hat{\hat{X}}_1^*]' \hat{a}_1^* + [\hat{X}_2^* - \hat{\hat{X}}_2^*]' \hat{a}_2^*.$$

By changing the initial weights in $d_i = F(x_{i1}\beta_1)/\pi_{ai}\pi_i$ we obtain in the same manner:

$$\hat{Y}_7 = [X_1' \hat{a}_1^*] + [\hat{X}_2^{*\prime} \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1^{*\prime} \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*].$$

By replacing $\hat{X}_2^*$ by its expression found above, we obtain:

$$\hat{Y}_7 = [X_1' \hat{C}_1^*] + [M_{x_1} \hat{X}_2' \hat{a}_2^*] + [\hat{\hat{Y}}^* - \hat{\hat{X}}_1'^* \hat{a}_1^* - \hat{\hat{X}}_2^{*\prime} \hat{a}_2^*],$$

with

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} \left(\sum_{s_a} \frac{x_1 x_2'}{\pi_a}\right) \hat{a}_2^*.$$

Finally, $\hat{Y}_7$ and $\hat{Y}_6$ are asymptotically equivalent.

Say that $w = y - x_1' \hat{a}_1^* - x_2' \hat{a}_2^*$. Then $\hat{Y}_7 = \hat{Y}_6 + [\hat{\hat{W}}^* - \hat{\hat{W}}]$. Now, asymptotically $[\hat{\hat{W}}^* - \hat{\hat{W}}]$ is an infinitely small negligible before $\hat{Y}_6$:

$$[\hat{\hat{W}}^* - \hat{\hat{W}}] = \left(\sum_s \frac{w x_1'}{\pi_a \pi}\right) \left(\sum_{s_a} \frac{x_1 x_1'}{\pi_a}\right)^{-1} [X_1 - \hat{\hat{X}}_1],$$

and

$$\left(\sum_s \frac{w x_1'}{\pi_a \pi}\right) \quad \begin{matrix}\text{tends toward} \\ \text{zero and}\end{matrix} \quad [X_1 - \hat{\hat{X}}_1] = O\left(\frac{1}{\sqrt{m}}\right).$$

Ultimately we obtain $\hat{Y}_7 \cong \hat{Y}_6$.

In conclusion, when the calibration function is exponential, the estimator $\hat{Y}_7$ coincides exactly with the preceding. When F is linear, $\hat{Y}_7$ is close to the preceding and thus still corresponds to the regression model approach in the case in which the information contained in x_1 is considered complementary to the information contained in x_2 for estimating y and in which the decomposition $y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_i$ is used.

Conclusion: The Three Classes of Estimators

We have just seen that the four strategies derived from a calibration approach could be associated with regression modelling. We thus obtain three classes of estimators:

$Y_4 \cong Y_1$ associated with the models

$$(1) \quad y_i = x_{i1}' b_1 + u_{i1},$$

and

$$(2) \quad y_i = x_{i2}' b_2 + u_{i2}$$

$\hat{Y}_5 = \hat{Y}_2$ associated with the model

$$y_i = x_{i1}' a_1 + x_{i2}' a_2 + u_i$$

$\hat{Y}_6 \cong \hat{Y}_3$ and $\hat{Y}_7 \cong \hat{Y}_3$ associated with the model

$$y_i = x_{i1}' c_1 + M_{x_1}(x_{i2})' c_2 + u_1.$$

The approximation $\cong$, which indicates that the estimators are attached to the same regression model, takes on its full meaning when we are interested in calculating the variance of these different estimators, since the estimators that are attached to the same regression model have the same asymptotic variance.

5. ESTIMATION OF VARIANCES

Let us consider the variances of the different estimators $\hat{Y}_1, \ldots, \hat{Y}_7$ defined above. AV designates the asymptotic variance of an estimator that is obtained when N, n and m tend toward infinity in a constant relationship.

5.1 Estimator $\hat{Y}_1$ and $\hat{Y}_4$: model $y_i = x_{i1}' b_1 + u_{i1}$ and (2) $y_i = x_{i2}' b_2 + u_{i2}$.

• Estimator $\hat{Y}_1$

The variance of this estimator and its estimate are given in the work of Särndal, Swensson and Wretman (1991). The variance breaks down into two terms that measure the amounts of variance due respectively to the first and the second phase of the sampling.

$$\mathrm{AV}(\hat{Y}_1) = \left(\sum_{i,j \in U} \Delta_{ij}^1 \frac{u_{1i} u_{1j}}{\pi_i \pi_j}\right) + \left(E_{s_a} \sum_{i,j \in s_a} \Delta_{ij}^2 \frac{u_{2i} u_{2j}}{\pi_i \pi_{ai} \pi_j \pi_{aj}}\right),$$

with: $\Delta_{ij}^2 = \pi_{ij} - \pi_i \pi_j$,

$$\Delta_{ij}^1 = \pi_{aij} - \pi_{ai} \pi_{aj},$$

$$u_{1i} = y_i - x_{i1}' b_1,$$

$$u_{2i} = y_i - x_{i2}' b_2,$$

$$b_1 = \left(\sum_{i \in U} x_{i1} x_{i1}'\right)^{-1} \left(\sum_{i \in U} x_{i1} y_i\right),$$

$$b_2 = \left(\sum_{i \in U} x_{i2} x_{i2}'\right)^{-1} \left(\sum_{i \in U} x_{i2} y_i\right).$$

Thus the variance estimator also breaks down into two terms that estimate the amounts of variance relating to each of the sampling phases. We find that by construction

of $\hat{Y}_1$, x_1 serves to reduce the variance brought about by the first phase and x_2 serves to reduce the variance brought about by the second phase.

$$\hat{V}(\hat{Y}_1) = \left(\sum_{i\in s}\sum_{j\in s}\frac{\Delta_{ij}^1}{\pi_{ij}\pi_{aij}}\frac{\hat{u}_{1i}\hat{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta_{ij}^2}{\pi_{aij}}\frac{\hat{u}_{2i}\hat{u}_{2j}}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

1st phase 2nd phase

with: $\hat{u}_{1i} = y_i - x_{i1}'\hat{b}_1,$

$\hat{u}_{2i} = y_i - x_{i2}'\hat{b}_2.$

Such a decomposition is based on the expression $V(\hat{Y}_1) = V(E[\hat{Y}_1 \mid s_a]) + E(V[\hat{Y}_1 \mid s_a])$, which will apply for all the other estimators.

• **Estimator $\hat{Y}_4$**

The terms of the development to the first order in $1/\sqrt{m}$ of $\hat{Y}_1$ and $\hat{Y}_4$ coincide exactly. We can therefore give a more precise meaning to the expression $\hat{Y}_4 \cong \hat{Y}_1$. We deduce from this that $\mathrm{AV}(\hat{Y}_1) = \mathrm{AV}(\hat{Y}_4)$. Thus:

$$\hat{V}(\hat{Y}_4) = \left(\sum_{i,j\in s}\sum_{j\in s}\frac{\Delta_{ij}^1}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta_{ij}^2}{\pi_{aij}}\frac{\tilde{u}_{2i}\tilde{u}_{2j}}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

with: $\tilde{u}_{1i} = y_i - x_{i1}'\hat{B}_1,$

$\tilde{u}_{2i} = y_i - x_{i2}'\hat{B}_2.$

5.2 Estimators $\hat{Y}_2 = \hat{Y}_5$: model $y_i = x_{i1}'a_1 + x_{i2}'a_2 + u_i$

It is easy to show (see Dupont 1994) that:

$$\mathrm{AV}(\hat{Y}_2) = \mathrm{AV}(\hat{Y}_5) \cong \left(\sum_{i,j\in U}\Delta_{ij}^1\frac{v_iv_j}{\pi_i\pi_j}\right) + \left(E_{s_a}\sum_{i,j\in s_a}\Delta_{ij}^2\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

with: $v_i = y_i - x_{i1}'a_1,$

$u_i = y_i - x_{i1}'a_1 - x_{i2}'a_2$

$$a = \begin{pmatrix}a_1\\a_2\end{pmatrix} = \begin{pmatrix}\sum_{i\in U}x_{i1}x_{i1}' & \sum_{i\in U}x_{i1}x_{i2}'\\ \sum_{i\in U}x_{i2}x_{i1}' & \sum_{i\in U}x_{i2}x_{i2}'\end{pmatrix}\begin{pmatrix}\sum_{i\in U}x_{i1}y_i\\ \sum_{i\in U}x_{i2}y_i\end{pmatrix}$$

From this we deduce that:

$$\hat{V}(\hat{Y}_2) = \hat{V}(\hat{Y}_5) = \left(\sum_{i,j\in s}\frac{\Delta_{ij}^1}{\pi_{ij}\pi_{aij}}\frac{\hat{v}_i\hat{v}_j}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta_{ij}^2}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

with: $\hat{v}_i = y_i - x_{i1}'\hat{a}_1,$

$\hat{u}_i = y_i - x_{i1}'\hat{a}_1 - x_{i2}'\hat{a}_2.$

In this formulation we find that by construction of $\hat{Y}_2 - \hat{Y}_5$, x_1 reduces the variance brought about by the first phase and x_1 and x_2 are used simultaneously to reduce the variance brought about by the second phase.

5.3 Estimators $\hat{Y}_3$, $\hat{Y}_6$ and $\hat{Y}_7$: model $y_i = x_{i1}'c_1 + M_{x_1}(x_{i2})'c_2 + u_i$

We show that $\mathrm{AV}(\hat{Y}_6) = \mathrm{AV}(\hat{Y}_7) = \mathrm{AV}(\hat{Y}_3)$. Thus,

$$\mathrm{AV}(\hat{Y}_3) = \mathrm{AV}(\hat{Y}_6) = \mathrm{AV}(\hat{Y}_7) \cong \left(\sum_{i,j\in U}\Delta_{ij}^1\frac{u_{1i}u_{1j}}{\pi_i\pi_j}\right) + \left(E_{s_a}\sum_{i,j\in s_a}\Delta_{ij}^2\frac{u_iu_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$u_{1i} = y_i - x_{i1}'c_1 = y_i - x_{i1}'b_1,$$

$$u_i = y_i - x_{i1}'c_1 - M_{x_1}x_{i2}'c_2 = y_i - x_{i1}'a_1 - x_{i2}'a_2.$$

From this we deduce the three variance estimators, which differ owing to different estimated coefficients:

$$\hat{V}(\hat{Y}_3) = \left(\sum_{i,j\in s}\frac{\Delta_{ij}^1}{\pi_{ij}\pi_{aij}}\frac{\hat{u}_{1i}\hat{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta_{ij}^2}{\pi_{aij}}\frac{\hat{u}_i\hat{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$\hat{u}_{1i} = y_i - x_{i1}'\hat{c}_1,$$

$$\hat{u}_i = y_i - x_{i1}'\hat{c}_1 - M_{x_1}x_{i2}'\hat{c}_2,$$

$$\hat{V}(\hat{Y}_6) = \left(\sum_{i,j\in s}\frac{\Delta_{ij}^1}{\pi_{ij}\pi_{aij}}\frac{\tilde{u}_{1i}\tilde{u}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s}\frac{\Delta_{ij}^2}{\pi_{aij}}\frac{\tilde{u}_i\tilde{u}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$\tilde{u}_{1i} = y_i - x_{i1}'\hat{C}_1,$$

$$\tilde{u}_i = y_i - x_{i1}'\hat{a}_1 - x_{i2}'\hat{a}_2,$$

$$\hat{V}(\hat{Y}_7) =$$

$$\left(\sum_{i,j\in s} \frac{\Delta_{ij}^1}{\pi_{ij}\pi_{aij}} \frac{\tilde{\tilde{u}}_{1i}\tilde{\tilde{u}}_{1j}}{\pi_i\pi_j}\right) + \left(\sum_{i,j\in s} \frac{\Delta_{ij}^2}{\pi_{aij}} \frac{\tilde{\tilde{u}}_i\tilde{\tilde{u}}_j}{\pi_i\pi_{ai}\pi_j\pi_{aj}}\right),$$

$$\tilde{\tilde{u}}_{1i} = y_i - x_{i1}'\hat{C}_1^*,$$

$$\tilde{\tilde{u}}_i = y_i - x_{i1}'\hat{a}_1^* - x_{i2}'\hat{a}_2^*,$$

$$\hat{C}_1 = \hat{a}_1^* + \left(\sum_{s_a} \frac{x_1x_1'}{\pi_a}\right)^{-1}\left(\sum_{s_a} \frac{x_1x_2'}{\pi_a}\right)\hat{a}_2^*.$$

We find that by construction of $\hat{Y}_3$, $\hat{Y}_6$ and $\hat{Y}_7$, x_1 is used to achieve *maximum* reduction of the variance brought about by the first phase and x_2 serves to reduce the variance brought about by the second phase.

6. CHOICE OF ESTIMATORS WHERE THERE IS SELECTION BIAS

In practice, when a survey is adjusted, it is not unusual to want not only to improve the estimation, but also and more especially to correct the biases introduced by uncontrolled selections of individuals, such as nonresponse.

We shall examine the case of a two-phase sampling in which the second phase is equivalent to total nonresponse. The weights π_i of the second-phase sampling are thus unknown. The calibration of s will enable us to estimate these probabilities, while reducing the variance (*cf.* Deville and Dupont 1993). However, asymptotically, the corrections of bias to be made to the weights are greater than the changes to be made in order to improve the estimators. It is therefore the implicit response model that will guide the choice between the different estimators:

The implicit response model for the first class of estimators is $p_i = 1/F(x_{i2}'B_2)$.

The implicit response model for the second and third classes of estimators is $p_i = 1/F(x_{i2}'A_2 + x_{i1}'A_1)$.

- Whatever the response model, an evaluation of the three classes of estimators on the basis of the sampling plan alone still indicates that the third is preferable, since it is appropriate for all the response models.
- If the strategies are evaluated on the basis of regression modelling, we will use the first class of estimators only if the response mechanism is well explained by x_2, that is, $p_i = 1/F(x_{i2}'B_2)$. Now, we have seen that the modelling associated with the first class of estimators takes on its meaning when the variables x_1 and x_2 are highly correlated. It is therefore fairly probable that in this context, the variable x_2 will be sufficient to explain the response mechanism. Should this not be the case, it will be necessary to turn to the third class of estimators.

The comparison between the three strategies may thus be adapted in a context in which we wish to correct the biases introduced by uncontrolled selections. The conclusions remain largely the same.

According to the same principle, it is of course possible to make comparisons between alternative adjustment strategies in the context of samplings that entail more than two phases and one or more uncontrolled selections.

7. *A PRIORI* AND *A POSTERIORI* USE OF AUXILIARY INFORMATION

The calibration estimator enables us to improve the estimate *a posteriori*, by reducing the variance and correcting the bias, as noted above. However, we may want to incorporate the auxiliary information *a priori*, at the sampling stage rather than *a posteriori* at the estimation stage. We then encounter, in a more complex context, the classical opposition between stratification and poststratification, well known in the case of single-phase sampling, when all the auxiliary variables are qualitative.

It is possible to transpose the terms of the choice between using the information *a priori* and *a posteriori*, in the sampling and auxiliary information configuration studied, when the auxiliary variables are qualitative. When the auxiliary variables are qualitative, a calibration corresponds exactly to poststratification.

We saw earlier that in order to determine the proper adjustment procedure, it was necessary to distinguish two possible modellings of the variable of interest, depending on whether the information in x_1 and the information in x_2 were considered substitutable or complementary. Each of these two modellings then led to one or more different adjustment procedures. Similarly, these two modellings arise when it is a matter of identifying the best sampling strategy for incorporating the auxiliary information:

- When the information in x_1 and the information in x_2 are substitutable, the modelling of the variable of interest is as follows:

  (1) $y_i = x_{i1}b_1 + u_{i1}$ and

  (2) $y_i = x_{i2}b_2 + u_{i2}$ where the second model is better for predicting the value of y_i.

  We have seen that the use of the auxiliary information *a posteriori* leads to calibration strategy No. 1, that is, to the first class of estimators. If we wish to take account of the auxiliary information at the sampling stage, it is natural to propose a sampling stratified on x_1 for the first phase and a sampling stratified on x_2 for the second phase.

  However, the parallel between the adjustment procedure and the sampling procedure is not complete: in a calibration, only the marginal information in x_1 can be used.

This results in incomplete poststratification (Särndal and Deville 1992). On the other hand, in the sampling procedure proposed as an *a priori* alternative, we are obliged to use all the cross-tabulations of the x_1 variables. The *a priori* equivalent of a calibration would accordingly be a sampling balanced on the margins of the vector of variables x_1.

- When the information contained in x_1 and the information contained in x_2 are complementary, the modelling of the variable of interest is y is $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$. We have seen that in this case the use of *a posteriori* auxiliary information led to calibration strategies 2, 3 and 4 in estimator classes 2 and 3. If we wish to take account of the auxiliary information at the sampling stage, it is natural to propose a sampling stratified on $x1$ for the first phase and a sampling stratified on x_1 and x_2 for the second phase.

  As before, there is no exact parallel between the *a priori* and *a posteriori* procedures, since the use of the information *a priori* mobilizes all the cross-tabulations between the variables x_1 and x_2.

Thus it is possible to make a choice between incorporating the information either *a priori* or *a posteriori*, and indeed to optimize the sampling plan, when the auxiliary variables are qualitative. The terms of the choice are the same as in a single-phase sampling with a single level of information. An additional consideration is the multiplicity of strata created by the cross-tabulations of x_1 and x_2 in the case in which the modelling used is $y_i = x_{i1}b_1 + x_{i2}b_2 + u_i$, which reinforces the advantages of using the information *a posteriori*.

When the auxiliary variables are quantitative, the choice depends on their conversion into qualitative variables, it not being possible to generalize correctly except by using the parallel between calibration and balanced sampling (*cf.* Deville 1992).

8. CONCLUSION

In a two-phase sampling, when two different sets of information are available for the total population on the one hand and the sample resulting from the first phase on the other hand, several strategies are possible when one wishes to use the auxiliary information to improve the estimation of totals.

Two different natural approaches have been used to derive estimators: a regression model assisted approach, which seeks to adapt the idea of the regression estimator; and a calibration approach, which attempts to adapt the idea of calibration. The estimators obtained by the two approaches may be linked together. We generated three alternative underlying modellings to which the various estimators obtained may be attached. Thus we obtained three classes of estimators. Several conceivable calibration strategies were eliminated at the outset as irrelevant.

We have shown that the estimators of a given class, that is, the estimators attached to a given model, are asymptotically equivalent; we gave the form of the variances derived in the case of a linear calibration function, but with asymptotic equivalences, these results remain valid for any calibration function.

For purposes of evaluating strategies, the form of the variances indicates, as intuition would suggest, that one of the classes of estimators (estimators 3, 6 and 7 (calibration strategies 3 and 4)) is preferable to the other from the standpoint of variance when the evaluation is based on the sampling plan alone. When it is based on a modelling of the variable of interest, it suggests that the preferable class of estimators is the one associated with the modelling adopted.

In a situation in which the goal is to adjust a survey and to simultaneously correct the biases that would arise from the use of gross weightings and reduce the variance, the conclusions must be adapted. The changes introduced in the weighting to correct the biases are greater than the corrections to reduce variance. Hence the variables will be incorporated into the calibration once it appears that they affect the probability of selection and thus participate in the creation of bias.

When the auxiliary variables are qualitative, the choice between *a priori* and *a posteriori* use of auxiliary information, that is, between using it at the sampling stage or at the adjustment stage, still rests on the distinction between the two modellings of the variable of interest.

These results may easily be generalized to the case of sampling involving more than two phases.

ACKNOWLEDGEMENTS

I am deeply grateful to Jean-Claude Deville, Louis Meuric and Carl-Erik Särndal for their many helpful suggestions regarding this article.

REFERENCES

CASSEL, C.M., SÄRNDAL, C.-E., and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrica*, 63, 615-620.

DEVILLE, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, October 1992, Örebro.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DEVILLE, J.-C., and DUPONT, F. (1993). Calage et redressement de la non-réponse totale. Journées de Méthodologie.

DUPONT, F. (1994). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. Working paper of Direction des Statistiques Démographiques et Sociales, F9409.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā* C, 37, 117-132.

GOURIEROUX, C. (1981). *Théorie des sondages*. Edition Economica Paris.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

SÄRNDAL, C.-E. (1980). On π inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phases sampling and nonresponse. *International Statistical Review*, 55, 279-294.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.