Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística – IBGE
Diretoria de Pesquisas
Coordenação de Métodos e Qualidade

Textos para discussão
Diretoria de Pesquisas
número 15

Calibration Estimation: When and Why, How Much and How

Pedro Luis do Nascimento Silva

Rio de Janeiro

2004

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro, RJ - Brasil

ISSN 1518-675X **Textos para discussão. Diretoria de Pesquisas**
Divulga estudos e outros trabalhos técnicos desenvolvidos pelo IBGE ou em conjunto com outras instituições, bem como resultantes de consultorias técnicas e traduções consideradas relevantes para disseminação pelo Instituto. A série está subdividida por unidade organizacional e os textos são de responsabilidade de cada área específica.

ISBN 85-240-3714-8

Impressão
Gráfica Digital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2000.

Capa

Gerência de Criação/CDDI

Silva, Pedro Luis do Nascimento
Calibration estimation : when and why, how much and how / Pedro Luis do Nascimento Silva. - Rio de Janeiro : IBGE, Coordenação de Métodos e Qualidade, 2004.
35p. - (Textos para discussão. Diretoria de Pesquisas, ISSN 1518-675X ; n. 15)

Inclui bibliografia.
ISBN 85-240-3714-8

1. Calibração. 2. Amostragem (Estatística). 3. Teoria da estimativa. I. IBGE. Coordenação de Métodos e Qualidade. II. Título. III. Série.

Gerência de Biblioteca e Acervos Especiais CDU 519.24
RJ/2004-09 EST

Impresso no Brasil / Printed in Brazil

Sumário

Apresentação

Estimação com Calibração: Quando e Por quê, Quanto e Como.

Esse assunto foi objeto de estudo realizado pelo autor durante Programa de Pós Doutorado realizado na Universidade de Southampton, Inglaterra, nos 4 meses compreendidos entre novembro de 2002 e fevereiro de 2003. O relatório que ora está sendo disponibilizado foi originalmente produzido em inglês. Porém, a importância de seu conteúdo, que é fortemente relacionado com as práticas que vêm sendo adotadas no processo de expansão das amostras das pesquisas realizadas por amostragem no IBGE, justifica não só essa divulgação, como estimula a preparação de uma versão em português.

O documento apresenta uma revisão da literatura sobre métodos de calibração usados na ponderação e estimação de pesquisas por amostra, aponta referências bibliográficas relevantes, discute questões importantes que surgem quando os métodos de calibração são usados em situações reais de pesquisa e aponta critérios que podem ser usados para avaliar se sua utilização foi bem sucedida ou se ocorreram dificuldades que demandem revisão dos resultados.

Sonia Albieri
Coordenadora da Coordenação de Métodos e Qualidade

1. Introduction

The main purpose of this report is to provide a literature review of calibration methods used for sample survey weighting and estimation, pointing to the most relevant references, as well as to discuss key issues arising when calibration methods are applied to real survey situations.

The report is structured as follows. Section 2 introduces the basic framework and a definition of calibration that we shall adopt throughout. Section 3 discusses reasons for calibration and situations when calibration is worthwhile. Section 4 discusses practical problems which one may face when performing calibration estimation. It also includes a review of several alternative methods for calibration, developed in response to the challenges posed by practical problems. Chapter 5 discusses some criteria that can be used to assess the success of calibration at any particular survey application. Chapter 6 provides some concluding remarks.

2. CALIBRATION ESTIMATION: A FRAMEWORK

Let {*1,…,k,...,N*} be the set of labels that uniquely identify the *N* distinct elements of a target finite population **U**. Without loss of generality, let **U** = {*1,…,k,...,N*}. A survey is carried out to measure the values of *J* survey variables. Denote by $\mathbf{y}_k = (y_{k1}, \ldots, y_{kJ})'$ the *J×1* vector of values of the survey variables for the *k*th population element.

We assume that the primary purpose of the survey is to estimate the population vector of totals $\mathbf{T}_y = \sum_{k \in \mathbf{U}} \mathbf{y}_k = \mathbf{Y}_\mathbf{U}' \mathbf{1}_N$ where $\mathbf{Y}_\mathbf{U}$ denotes the *N×J* population matrix of *y* values given by $\mathbf{Y}_\mathbf{U} = \left[\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\right]'$, and $\mathbf{1}_N$ denotes the *N×1* vector of ones.

Although censuses are sometimes carried out to collect data about certain populations, the vast majority of surveys are sample surveys, in which only a sample of the population elements (usually a small portion) are investigated. We assume that n distinct elements in $\mathbf{U}$ are included in a sample s, $s = \{k_1, \ldots, k_n\} \subset \mathbf{U}$, which is selected for observation in the survey.

Hence the purpose of the survey is to estimate $\mathbf{T}_y$ on the basis of the available survey data $\{\mathbf{y}_k ; k \in s\}$. The "standard" estimator for totals when these are the only data available from the sample is the Horvitz-Thompson estimator defined as

$$\hat{\mathbf{T}}_y = \sum_{k \in s} d_k \mathbf{y}_k \tag{2.1}$$

where $d_k = 1/\pi_k$ is the design weight for unit k, and π_k is the sample inclusion probability for unit k. Denoting by π_{ki} the joint sample inclusion probability for elements k and i, here we assume that all the first and second order inclusion probabilities are strictly positive, i.e. $\pi_k > 0$ and $\pi_{ki} > 0 \ \forall \ k,i \in \mathbf{U}$. The assumption of positive π_{ki} is satisfied by the designs considered in this report, and is adopted throughout because it simplifies the presentation of expressions for design variances and their estimators. However, it is not a crucial assumption, since for many of the designs for which it is not satisfied reasonable approximations and estimators for the design variance of estimators of totals (means) are readily available (see e.g. Berger, 2002).

In most survey applications, however, the survey data may also include information on some auxiliary variables $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})'$, which may often be useful towards estimating the unknown population totals of the survey variables $\mathbf{T}_y$. Assuming for now full response to the selected sample, there are two scenarios for availability of information about the auxiliary variables that one may consider.

a) The full "population auxiliary data matrix" $\mathbf{X}_\mathbf{U} = \left[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\right]'$ is available from the survey frame, in which case the finite population totals $\mathbf{T}_x = \sum_{k \in \mathbf{U}} \mathbf{x}_k = \mathbf{X}_\mathbf{U}' \mathbf{1}_N$ of these p auxiliary variables will also be known, together with the "sample auxiliary data matrix" $\mathbf{X}_s = \left[\mathbf{x}_{k1}, \mathbf{x}_{k_2}, \cdots, \mathbf{x}_{kn}\right]'$, namely the submatrix of $\mathbf{X}_\mathbf{U}$ obtained by keeping only the rows corresponding to the units selected for the sample ($k \in s$).

b) Only the "sample auxiliary data matrix" $\mathbf{X}_s$ is available, together with the vector of auxiliary population totals $\mathbf{T}_x$. For now we shall assume that $\mathbf{T}_x$ is known exactly.

In both scenarios where there is auxiliary population information available for some *x* variables we may ask the question whether this information may be used for improving the estimation of the target parameter $\mathbf{T}_y$. The answer to this question is yes: quite often we can do better when estimating $\mathbf{T}_y$ taking account of the available information about the *x* variables than by using the standard Horvitz-Thompson estimator (2.1).

One way to do this is by calibration. The key idea behind calibration estimation is as follows. Although we know the population totals for the *x* variables, suppose we would try to estimate them from the sample, using the Horvitz-Thompson estimator (2.1). This would lead to the estimation of $\mathbf{T}_x$ by $\hat{\mathbf{T}}_x = \sum_{k\in s} d_k \mathbf{x}_k$. However, these estimates $\hat{\mathbf{T}}_x$ often would not match the corresponding population totals $\mathbf{T}_x$ exactly, leading to the so-called "calibration error" $\hat{\mathbf{T}}_x - \mathbf{T}_x$. To avoid this "error", we can try and modify the estimator in such a way that there would be no calibration error. This may be accomplished by using a "calibrated" estimator where the design weights d_k are modified, giving way to new weights w_k to be used in the calibrated estimator

$$\hat{\mathbf{T}}_{xC} = \sum_{k\in s} w_k \mathbf{x}_k \tag{2.2}$$

where $\{w_k, k\in s\}$ are case weights such that there is no calibration error, i.e., satisfying

$$\hat{\mathbf{T}}_{xC} - \mathbf{T}_x = \sum_{k\in s} w_k \mathbf{x}_k - \mathbf{T}_x = \mathbf{0} \tag{2.3}$$

The conditions (2.3) are called the "calibration constraints". The idea is that if the "calibrated" weights $\{w_k, k\in s\}$ succeed in reducing or avoiding error when "estimating" the *x* totals, they may also reduce the error when estimating the *y* totals, using the calibration estimator:

$$\hat{\mathbf{T}}_{yC} = \sum_{k\in s} w_k \mathbf{y}_k \tag{2.4}$$

The "calibrated" case weights $\{w_k, k\in s\}$ may depend on all the information available about the auxiliary variables *x*, but not on the survey variables *y*. If this is the case, then (2.4) is a linear estimator of $\mathbf{T}_y$.

In this report, we concentrate on "calibration to total" estimators of the form (2.4), i.e., linear estimators defined by sets of weights $\{w_k,\ k \in s\}$ satisfying the "calibration to totals constraints" (2.3). Other forms of calibration constraints may be considered, such as calibration to higher-order moments or even to the finite population distribution function of the auxiliary variables (see the discussion in section 10 of Chambers, 1997). However, these other forms of calibration estimators shall not be considered here, and for simplicity, we shall follow the prevailing simple denomination of the estimators defined by (2.4) with weights satisfying (2.3) as "calibration estimators".

A large number of sets of weights $\{w_k,\ k \in s\}$ may satisfy the calibration constraints given the sample data $\mathbf{X}_s$, the design weights $\{d_k,\ k \in s\}$ and the population totals $\mathbf{T}_x$. One way of selecting those that lead to "reasonable" sets of weights to be used to estimate totals for the y variables is to think of calibration weights w_k as modifications to the design weights d_k that change them the least. This is justified because using the design weights d_k provides the corresponding Horvitz-Thompson estimator (2.1) with desirable properties such as design-unbiasedness and consistency (in the sense that as the sample size increases, the estimator converges in probability towards the right target $\mathbf{T}_y$).

Deville and Särndal (1992) defined a family of calibration estimators for $\mathbf{T}_y$ where the weights w_k are chosen such that specified distance functions measuring how far the w_k are from the d_k are minimised. Their idea is to minimize

$$E_P\left(\sum_{k\in s} G_k\left(w_k,d_k\right)\right) \tag{2.5}$$

or equivalently minimize, for every sample s,

$$\sum_{k\in s} G_k\left(w_k,d_k\right) \tag{2.6}$$

subject to (2.3), where $G_k\left(w_k,d_k\right)$ is a measure of the distance between w_k and d_k satisfying some regularity conditions to be specified later, and E_P denotes the expectation with respect to the probability distribution induced by the sampling design used to select the sample s.

One popular choice for the distance function is to take

$$G_k\left(w_k,d_k\right)=\frac{\left(w_k-d_k\right)^2}{q_k d_k} \quad k \in s \tag{2.7}$$

for some known constants $q_k > 0$, $k \in s$, to be specified. In this case, the solution is given by

$$w_k = d_k \times g_k \tag{2.8}$$

where

$$g_k = 1 + q_k \left(\mathbf{T}_x - \hat{\mathbf{T}}_x\right)' \left(\sum_{i \in s} q_i d_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \mathbf{x}_k . \tag{2.9}$$

With the weights (2.8), the resulting calibration estimator for the total of a survey variable y_j can be written as

$$\hat{T}_{y_j C} = \sum_{k \in s} w_k y_{kj} = \hat{T}_{y_j} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_x\right)' \hat{\mathbf{B}}_j \tag{2.10}$$

where $\hat{T}_{y_j} = \sum_{k \in s} d_k y_{kj}$ is the Horvitz-Thompson estimator for $T_{y_j} = \sum_{k \in \mathbf{U}} y_{kj}$ and $\hat{\mathbf{B}}_j$ is defined in (2.13) below. Note that (2.10) is a generalized regression (GREG) estimator (see Särndal, Swensson & Wretman, 1992), motivated by the working population model

$$\begin{aligned} y_{kj} &= \mathbf{x}_k' \mathbf{B}_j + E_{kj} \\ V\left(E_{kj}\right) &= \sigma_j^2 / q_k \end{aligned} \tag{2.11}$$

with the population regression coefficients $\mathbf{B}_j$ defined by

$$\mathbf{B}_j = \left(\sum_{k \in \mathbf{U}} q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{k \in \mathbf{U}} q_k \mathbf{x}_k y_{kj}\right) \tag{2.12}$$

and corresponding sample estimators given by

$$\hat{\mathbf{B}}_j = \left(\sum_{k \in s} q_k d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_{k \in s} q_k d_k \mathbf{x}_k y_{kj}\right). \tag{2.13}$$

If a single set of calibrated weights w_k is to be used for all survey (y) variables then (2.9) means that the same set of constants q_k will be used for all survey variables as well. In many applications, this would not be a problem, since a common choice is to make all these constants the same, i.e. $q_k = 1 \ \forall k \in s$.

However, in some cases different y variables might have residuals of the population linear regression on the auxiliary variables that display different heteroskedasticity patterns.

In such cases, the different sets of values needed for the constants q_k to represent such patterns adequately might lead to different sets of calibrated weights, each set specific to one or more survey variables. On one hand this might be justified on the grounds of improved efficiency for estimating the total of each *y* variable. On the other hand this would lead to potential problems of coherence. For example, weighted estimates of parts of a sum might not match the weighted estimate of total for the sum of the parts. Hence the idea of using different sets of weights for different *y* variables is unattractive in practice.

Although this in fact is not a requirement of calibration, we assume from now on that the derivation of the calibration weights is made with the goal of using a single set of calibrated weights $\{w_k, k \in s\}$ for estimation with all survey variables.

3. REASONS FOR CALIBRATION

Calibration estimators have some nice properties. First, calibration weights satisfying (2.3) provide sample "estimates" for the totals of the auxiliary variables that match exactly the known population totals for these variables. Hence, if the population totals of the auxiliary variables have been published before the survey results are to be produced, then calibration would guarantee that the survey estimates are coherent with those already on the public domain. This property, although not essential, is one of the dominant reasons why calibration is so often used in survey practice. It appeals to survey practitioners in many instances as a way of enforcing agreement between their survey and some public domain totals for key variables.

The second property is their simplicity, namely the fact that calibration estimates are linear. This means that each survey record can carry a single weight to be used for estimation for all survey variables. Calculation of the estimates for totals, means, ratios and many other parameters is straightforward using standard statistical software. In the case of the distance functions defined by (2.6) and (2.7), the calibrated weights are given in a closed form expression and are easy to compute using standard statistical software.

The third property of such calibration estimators is their flexibility to incorporate auxiliary information that can include continuous, discrete or both types of variables at the same time. If the auxiliary totals represent counts of the numbers of population units in certain classes of categorical (discrete) variables, then the values of the corresponding *x* variables are simply indicators of the units being members of the corresponding classes. Cross-classification of two or more categorical variables can also be easily accommodated by defining indicator variables for the corresponding combinations of categories.

Calibration estimators also yield some degree of integration in the sense that some widely used estimators are special cases, e.g. ratio, regression and poststratification estimators (see chapter 7 of Särndal, Swensson & Wretman, 1992), as well as incomplete multiway poststratification (see Bethlehem & Keller, 1987).

Calibration estimators may also offer some protection against nonresponse bias. Poststratification and regression estimation are widely used to attempt to reduce nonresponse bias in sample surveys. The regression (calibration) estimator will be approximately unbiased when the regression model (2.11) holds and the combined sampling and response mechanism is ignorable given the set of *x* variables for which auxiliary population information is available (e.g. see Bethlehem, 1988, Lundström & Särndal, 1999, and also chapter 15 of Särndal, Swensson & Wretman, 1992).

All these reasons are powerful arguments for using calibration. However, when doing so, users must be aware of some problems or difficulties that may be encountered as well. First, we note that calibration estimators are not exactly design unbiased. In fact, the design bias of the calibration estimator is given by

$$Design\ Bias\left(\hat{\mathbf{T}}_{yC}\right) = E_P\left(\hat{\mathbf{T}}_{yC} - \mathbf{T}_y\right) = E_P\left[\sum_{k\in s}\left(w_k - d_k\right)\mathbf{y}_k\right] \tag{3.1}$$

If the calibrated weights are “close” to the design weights for all samples, then the design bias will be negligible or close to zero. This supports the criterion used to define the calibration weights w_k, which requires that their distance to the d_k be minimized. However, for small or moderate sample sizes one has to be aware of the possibility of facing some amount of design bias.

For large samples, the calibration estimator defined by the regression weights (2.8) and (2.9) is asymptotic design unbiased and has approximate design variance (see Särndal, Swensson & Wretman, 1992, p. 235) given by

$$AV_P\left(\hat{T}_{y_jC}\right) = \sum_{k\in U}\sum_{i\in U}\left(\pi_{ki} - \pi_k\pi_i\right)\left(d_k E_{kj}\right)\left(d_i E_{ij}\right) \tag{3.2}$$

where E_{kj} is the residual of the population regression model (2.11) for the survey variable y_j. If the bias is negligible, we can then compare this approximate variance to that of the standard Horvitz-Thompson estimator for $\hat{T}_{y_j}$, given by:

$$V_P\left(\hat{T}_{y_j}\right) = \sum_{k\in U}\sum_{i\in U}\left(\pi_{ki} - \pi_k\pi_i\right)\left(d_k y_{kj}\right)\left(d_i y_{ij}\right). \tag{3.3}$$

Under simple random sampling without replacement and assuming that $q_k = 1$, the above expressions simplify to

$$AV_{SRS}\left(\hat{T}_{y_jC}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2_{E_j} \tag{3.4}$$

and

$$V_{SRS}\left(\hat{T}_{y_j}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2_{y_j} \tag{3.5}$$

respectively, where $\sigma^2_{E_j}$ is the variance of the residuals E_j and $\sigma^2_{y_j}$ is the variance of the survey variable y_j.

We can then observe that the regression (calibration) estimator will be expected to perform well in terms of precision when the variance of the residuals of the regression model defined by (2.11) is small compared to that of the original *y* variable. This will be the case when the linear relationship is a good approximation for the regression of *y* on **x** and the *x* variables in the regression estimator have good predictive power for *y*. The two plots in Figure 1 illustrate this idea. In this example, the residuals of the regression estimator for the model $y = Bx$ have smaller variance than the original *y* variable (model $y = B$), thus leading to the regression estimator having smaller approximate variance than the variance of the Horvitz-Thompson estimator for samples of the same size.

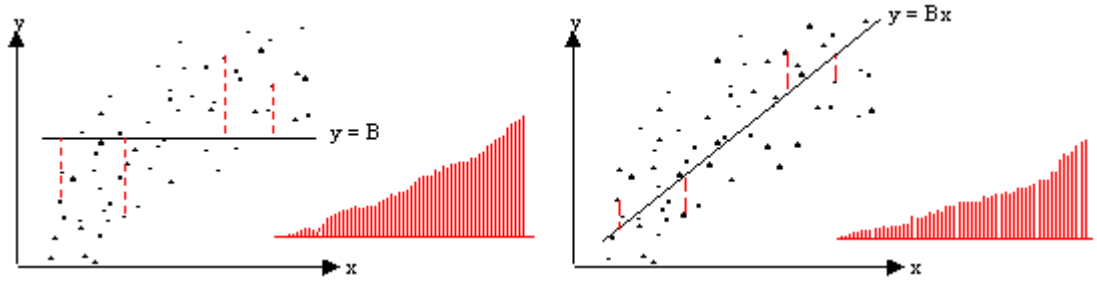

Figure 1 - Population residuals for Horvitz-Thompson (left) and regression (right) estimators

In addition, two relatively straightforward variance estimators are available for use with the regression estimator. Särndal, Swensson & Wretman (1992, p. 235) recommend using

$$\hat{V}_g\left(\hat{T}_{y_jC}\right)=\sum_{k\in s}\sum_{i\in s}\left(1-\pi_k\pi_i/\pi_{ki}\right)\left(g_k d_k e_{kj}\right)\left(g_i d_i e_{ij}\right) \tag{3.6}$$

where $e_{kj} = y_{kj} - \mathbf{x}_k'\hat{\mathbf{B}}_j$. An even simpler variance estimator not requiring the g-weights is given by

$$\hat{V}_s\left(\hat{T}_{y_jC}\right)=\sum_{k\in s}\sum_{i\in s}\left(1-\pi_k\pi_i/\pi_{ki}\right)\left(d_k e_{kj}\right)\left(d_i e_{ij}\right) \quad . \tag{3.7}$$

Both variance estimators are first order asymptotically design unbiased for the approximate variance of the regression estimator, but (3.6) is also approximately model unbiased (Särndal, Swensson & Wretman, 1989). In addition, Silva (1996, p. 48) demonstrated that under simple random sampling without replacement and assuming that the regression model (2.11) holds, the bias of (3.6) is $\mathbf{O}(n^{-5/2})$, whereas the bias of (3.7) is $\mathbf{O}(n^{-2})$. Hence (3.6) should be preferred to (3.7). Holmes and Skinner (2000) support this view based on results of an empirical study carried out to compare alternative variance estimators for the UK Labour Force Survey (UK-LFS).

4. PRACTICAL PROBLEMS WITH CALIBRATION ESTIMATION

While calibration estimators possess a number of attractive properties, they are not problem-free when it comes to practical applications. In this section, we review some of the problems affecting calibration estimators and some of the approaches that have been developed to tackle them. Before detailed discussion, however, it may be helpful to have a quick list of the issues that should be of concern when performing calibration estimation in practice:

A. Samples are finite, often small in certain strata;

B. Large numbers of “model groups” and/or *survey* variables;

C. Negative, small (less than 1) or extreme (large) weights;

D. Large number of *auxiliary* variables;

E. Nonresponse;

F. Measurement error.

The last issue in this list (measurement errors and their effect on calibration), despite its importance, is not going to be discussed here. Readers can find some discussion in Skinner (1999). All the other issues will be dealt with in the following sections.

4.1 - Small Sample Sizes

The problem with calibration estimators when sample sizes are small comes from the fact that their design bias may become important, relative to their variance. It is well known, for example, that ratio estimators are design biased, and that the bias is $O(n^{-1})$ (see Cochran, 1977, p. 160-162). The ratio estimator is a special case of the calibration estimator when the **x** vector includes a single continuous variable x, no intercept term is included, and the q_k constants are set to $q_k = 1/x_k$ (assuming that $x_k > 0 \ \forall \ k$). It is recommended that ratio estimators be used only for samples of sizes large enough so that the bias is negligible. Särndal, Swensson & Wretman (1992, p. 251) suggest that samples of size 20 or more should be sufficient for this to happen. Cochran (1977, p. 162) suggests that the coefficient of variation of the Horvitz-Thompson estimator of the total of the x variable ($CV\left(\hat{T}_x\right)$) should be less than 0.1 (10%) before the bias of the ratio estimator can be ignored or considered small in comparison to its standard error.

Despite these well-known "rules of thumb" or limitations that should prevent ratio estimators from being used with very small samples, modern software makes it easy for ratio and other calibration estimators to be computed for samples of any size, often without any warnings that sample sizes may be insufficient to warrant safe utilization. This leaves room for applications where not even minimal precautions are taken, like this one of checking whether the sample size is adequate. In cases where sample sizes are too small, calibration estimates may be subject not only to large variance (as expected due to the small sample size) but also to noticeable bias. Users of calibration estimators are urged to avoid applying the technique when sample sizes are too small. As yet, no simple safety rules regarding minimal sample sizes were developed for the general family of calibration estimators. However, one could at least suggest that the same rules applicable to simple ratio estimation should be satisfied before applying some other form of calibration estimation.

4.2 - Large number of “model groups” and/or survey variables

Another source of difficulties for using regression (or calibration) estimators is the fact that these are often applied separately for a number of “model groups”, defined as groups of units for which both sample membership and auxiliary population information is available. These model groups may coincide with pre-defined sampling strata, or they may be formed after the sample was selected, in which case they will play the role of post-strata. When such model groups are numerous, small sample sizes may result for some (or indeed many) of them.

The problem is often compounded by the fact that the number of survey variables may also be large. In this case, although the computation of the calibration weights is carried out just once (the weights do not depend on the *y* variables), the suitability of the underlying models that provide the conditions for calibration estimators to perform well (in the sense of providing residuals with small variance) should be verified. In some cases, this task may become just too large to be feasible within tight production schedules that typical surveys have to adhere to. For this reason, users are thus cautioned against attempting to perform calibration at levels that are too detailed in the sense of involving too many model groups. The more model groups are considered for the calibration, the more resources should be devoted to model validation and analysis of the resulting calibration estimates.

This discussion resembles that for comparing separate and combined ratio estimators. Separate ratio estimators are calibration estimators where calibration is performed to totals known at the stratum (model group) level. Combined ratio estimators involve calibration only at the aggregate level (for the sample as a whole or for some broader groups formed from sets of pooled strata). Cochran (1977, p. 167) argues “the use of a separate ratio estimate in each stratum is likely to be more precise if the sample in each stratum is large enough so that the approximate formula for the variance of the separate ratio estimator is valid, and the cumulative bias that can affect the separate ratio estimator is negligible. With only a small sample in each stratum, the combined ratio estimate is to be recommended unless there is good empirical evidence to the contrary.” I could not find better words to put it myself, and would suggest that this advice should also be applied to calibration estimators in general.

4.3 - Negative, small or extreme weights

The third set of problems comprises those situations arising when the calibration weights are in some sense considered extreme or unrepresentative. One important case

occurs when calibration results in negative weights, that is, in having some weights $w_k < 0$ (or $g_k < 0$). This situation represents no problem from a strictly theoretical point of view, but it leads to two difficulties from a practical perspective. First, the usual interpretation of case weights as the number of population units represented by the corresponding sample unit is lost for these cases, and release of such weights would be a very uncomfortable decision for many statistical agencies. The second problem is that negative weights might eventually yield negative estimates for some domains with small sample sizes, which is not an acceptable outcome for most practical survey applications when the survey variables are intrinsically non-negative. We also note that negative weights may provide an indication of some problem with the attempted calibration that needs attention from the statistician in charge.

To tackle this problem of the possibility of negative weights, a number of approaches have been developed. One approach that is implemented in software packages developed by some statistical agencies is to compute the calibration adjustment weights g_k that minimize

$$\sum_{k\in s}(w_k - d_k)^2/q_k d_k = \sum_{k\in s}(d_k g_k - d_k)^2/q_k d_k = \sum_{k\in s} d_k(g_k - 1)^2/q_k \qquad (4.1)$$

subject to the calibration constraints

$$\hat{\mathbf{T}}_{xC} - \mathbf{T}_x = \sum_{k\in s} g_k d_k \mathbf{x}_k - \mathbf{T}_x = \mathbf{0} \qquad (4.2)$$

and also to the additional boundary constraints

$$L \leq g_k \leq U \text{ for } k \in s \qquad (4.3)$$

where $0 < L < 1 < U$.

This is the approach adopted in the development of GES (Generalized Estimation System) by Statistics Canada (Estevao et al., 1995). This problem corresponds to minimization of a quadratic function (4.1) under linear (4.2) and (non-linear) boundary (4.3) constraints. GES attempts to solve this problem using an efficient algorithm, but a solution is not always guaranteed to exist. GES includes, in addition to determination of the calibration weights, efficient estimation of totals, means and ratios for populations and by domains, together with corresponding variances for stratified element or single stage cluster sampling designs. Statisticians looking for a computational tool to implement calibration should give this package due consideration. One drawback is its dependency on SAS statistical

software, which makes this a reasonably pricey option. If SAS is already available, site licensing of GES from Statistics Canada is not prohibitive for most large scale statistical agencies, and would cost a lot less than developing an equivalent software.

Another implementation of the above approach is available in BASCULA (see Nieuwenbroek & Boonstra, 2002). The main difference between GES and BASCULA is the algorithm used to compute the calibration weights. BASCULA adopts an algorithm proposed by Huang & Fuller (1978) to compute calibrated weights satisfying the boundary constraints. As is the case with GES, BASCULA is also not always guaranteed to find a solution satisfying all the specified constraints. BASCULA is a stand-alone program, and thus may be cheaper to obtain than GES if the organization is not yet a user of SAS.

Another approach that was proposed to solve the problem of negative weights is due to Deville & Särndal (1992), who defined the family of calibration estimators. In the previous approach, the standard distance function leading to regression weights was maintained and boundary conditions were imposed as additional constraints. The approach proposed by Deville & Särndal consists of modifying the distance function to be used when computing the calibrated weights, in such a way as to avoid the possibility of negative weights from the start. Hence the idea is to define calibration weights that minimize

$$\sum_{k \in s} G_k\left(w_k, d_k\right) \tag{4.4}$$

for every sample *s*, subject to the calibration constraints (4.2), where the distance functions G_k can be one of the choices in Table 1. Note that the standard distance function (case 1) is also included for completeness, because it is a member of the family, but it can yield negative weights. All the distance functions considered satisfy some regularity conditions, namely, for every fixed $d>0$:

a) $G_k\left(w,d\right) \geq 0 \quad \text{and} \quad G_k\left(d,d\right) = 0$;

b) $G_k\left(w,d\right)$ is defined in an interval D_k containing d;

c) $G_k\left(w,d\right)$is strictly convex and differentiable twice in w;

d) $\partial G_k\left(w,d\right)/\partial w$ is continuous and maps D_k onto an interval $Im_k(d)$ in a one to one fashion.

Table 1 – Distance functions for calibration estimation proposed in Deville & Särndal (1992)

Case	Distance Functions $q_k \times G_k(w_k, d_k)$
1	$(w_k - d_k)^2 / 2d_k$
2	$w_k [\log(w_k / d_k) - 1] - d_k$
3	$2\left(\sqrt{w_k} - \sqrt{d_k}\right)^2$
4	$w_k - d_k [\log(w_k / d_k) + 1]$
5	$(w_k - d_k)^2 / 2w_k$
6	$(g_k - L)\log\left(\frac{g_k - L}{1 - L}\right) + (U - g_k)\log\left(\frac{U - g_k}{U - 1}\right)$ $g_k = \frac{w_k}{d_k}, 0 < L < 1 < U$

The solution for the minimization problem can be obtained using the method of Lagrange multipliers. Using this method, the w_k that minimize (4.4) subject to (2.3) are obtained as a solution to

$$\partial G_k(w_k, d_k)/\partial w_k - \mathbf{x}_k'\boldsymbol{\lambda} = 0 \quad \forall k \in s\,. \tag{4.5}$$

If a solution exists, considering the regularity assumptions adopted, it will be unique, and given by

$$w_k = d_k F(q_k \mathbf{x}_k'\boldsymbol{\lambda}) = d_k g_k \tag{4.6}$$

where $F(\cdot)$ is the reciprocal mapping of $\partial G_k(w,d)/\partial w$ (see Table 2), $g_k = F(q_k \mathbf{x}_k'\boldsymbol{\lambda})$ and $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers, that solves

$$\sum_{k \in s} d_k [F(q_k \mathbf{x}_k'\boldsymbol{\lambda}) - 1]\mathbf{x}_k = \mathbf{T}_x - \hat{\mathbf{T}}_x \tag{4.7}$$

The resulting calibration estimator is then given by

$$\hat{\mathbf{T}}_{yC} = \sum_{k \in s} g_k d_k \mathbf{y}_k \tag{4.8}$$

with the calibration adjustment factors g_k defined by one of the calibration functions $F(\cdot)$ in table 2.

Table 2 – Calibration functions corresponding to various distance functions proposed by Deville & Särndal (1992)

Case	"Calibration" Functions $F(q_k u)$
1	$1+q_k u$
2	$\exp(q_k u)$
3	$(1-q_k u/2)^{-2}$
4	$(1-q_k u)^{-1}$
5	$(1-2q_k u)^{-1/2}$
6	$\frac{L(U-1)+U(1-L)\exp(Aq_k u)}{(U-1)+(1-L)\exp(Aq_k u)}$, $A=\frac{U-L}{(1-L)(U-1)}$, $0<\mathrm{L}<1<\mathrm{U}$
7	L if $u < (L\text{-}1)/q_k$ $1+q_k u$ if $(L\text{-}1)/q_k \le u \le (U\text{-}1)/q_k$ U if $u > (U\text{-}1)/q_k$

(1) Note that the calibration function 7 corresponds to the distance function number 1 of table 1, but with bounds specified for the calibration weights.

Hence an algorithm for computing the calibration weights may be specified as the following sequence of steps.

Step 1: Compute the calibration error for the Horvitz-Thompson estimator of the totals of the auxiliary variables: $\mathbf{T}_x - \hat{\mathbf{T}}_x$.

Step 2: For the chosen calibration function $F(\cdot)$, solve the *calibration equations* needed to determine $\boldsymbol{\lambda}$, namely

$$\sum_{k\in s} d_k \left[F(q_k \mathbf{x}_k'\boldsymbol{\lambda})-1\right]\mathbf{x}_k = \mathbf{T}_x - \hat{\mathbf{T}}_x \qquad (4.9)$$

This may be accomplished by using Newton's method. First, define

$$\mathbf{H}_s(\boldsymbol{\lambda}) = \sum_{k \in s} d_k \left[F\left(q_k \mathbf{x}_k' \boldsymbol{\lambda}\right) - 1\right] \mathbf{x}_k \,. \tag{4.10}$$

Then the step 2 of the algorithm requires finding the value $\boldsymbol{\lambda}$ that solves $\mathbf{H}_s(\boldsymbol{\lambda}) = \mathbf{T}_x - \hat{\mathbf{T}}_x$. First we compute an initial value for $\boldsymbol{\lambda}$ as

$$\boldsymbol{\lambda}_1 = \left(\sum_{k \in s} q_k d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left[\mathbf{T}_x - \hat{\mathbf{T}}_x\right] \tag{4.11}$$

Then perform iterations of Newton's method computing, at each iteration r=1,2,…, the updated value

$$\boldsymbol{\lambda}_{r+1} = \boldsymbol{\lambda}_r + \left[\mathbf{H}_s'(\boldsymbol{\lambda}_r)\right]^{-1} \left[\mathbf{T}_x - \hat{\mathbf{T}}_x - \mathbf{H}_s(\boldsymbol{\lambda}_r)\right] \tag{4.12}$$

where

$$\mathbf{H}_s'(\boldsymbol{\lambda}_r) = \partial \mathbf{H}_s(\boldsymbol{\lambda}) / \partial \boldsymbol{\lambda} \Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_r} \,. \tag{4.13}$$

Iterations proceed until convergence (given specified tolerance limits) or until the maximum number of iterations allowed is reached without achieving convergence, in which case an alert should be issued that a solution was not found.

Step 3: Once the solution for $\boldsymbol{\lambda}$ was obtained, compute the *calibration weights*

$$w_k = d_k F\left(q_k \mathbf{x}_k' \boldsymbol{\lambda}\right). \tag{4.14}$$

The calibration weights and corresponding estimators obtained as a result of this algorithm preserve all desirable properties that we discussed in connection with regression estimators (sections 2 and 3). In addition, raking ratio estimators such as used for weighting persons in the UK-LFS may also be seen as special cases of the general class of calibration estimators. Deville & Särndal (1992) demonstrated that members of this class have asymptotic properties identical to those of GREG estimators based on the same set of auxiliary variables. Hence, general calibration estimators defined by one of the above distance functions are asymptotically design unbiased, with approximate variance given by (3.2). In addition, their variance can also be estimated by (3.6) or (3.7).

Calibration estimators of this type were implemented in the SAS Macro CALMAR (Sautory, 1993). This program performs weight computation only, but a variant named CALJACK was developed at Statistics Canada (Bernier & Lavallée, 1994) that includes Jackknife variance estimation for totals, means, ratios and differences of these. CALMAR also requires SAS, but a more recent (but limited) implementation of the method is available: g-CALIB-S, developed at Statistics Belgium, runs under SPSS (Vanderhoeft, 2001).

Calibration estimation as now extended provides the tools to try and resolve the problem of negative weights, which can be avoided by choosing calibration functions 2 to 7 in Table 2. It also gives some control over the problem of extreme weights or weights less than 1, which can be avoided by choosing calibration functions 6 or 7 and making $L = 1/\min\{d_k, k \in s\}$ and specifying some suitable *U*. However, several of the problems discussed before remain unsolved.

First, for small and moderate samples, bias may be an issue and now, the choice of distance function may become important in this respect. Second, although the method is geared towards avoiding negative or extreme weights, a solution is not guaranteed. Deville & Särndal (1992) proved that the probability of finding a solution for $\boldsymbol{\lambda}$ in step 2 of the algorithm tends to one as *n* increases. However, it is not one with finite samples. Hence in some applications the method may fail to converge depending on the choices of $F(\cdot)$, *L* and *U*. When this is the case, users of the method should try and investigate the causes behind the failure to find a solution. It may be due to small or "extreme" samples, in the sense that the resulting calibration weights may need to be more extreme than we are prepared to allow for when we specified the boundary constraints *L* and *U*. It may also happen because large numbers of *x* variables are considered for calibration, which may lead to problems of collinearity, an issue that we discuss in the next section.

4. 4 - Large number of auxiliary variables

One problem that the approaches discussed above do not tackle is what to do when a large number of potential *x* variables are available to be considered for calibration. One simplistic option is to consider every one of the potential *x* variables in the calibration. This may seem desirable from a practical point of view, because calibration error would be zero for all known population totals. However, this option may also cause a number of problems. First, it may be more difficult to solve the system of calibration equations required for determining $\boldsymbol{\lambda}$ in step 2 of the algorithm, because its size increases with the number of *x* variables, and computation may be demanding. Second, larger numbers of *x* variables may

lead to collinearity problems that affect solution of the calibration equations. Bankier (1990) and Sautory (1993) proposed discarding linearly dependent auxiliary variables prior to attempting the solution of the calibration equations in step 2 of the algorithm. This solution is rather easy to implement and does not lead to loss of calibration for any *x* variables, since discarded variables are exact linear combinations of variables retained in the calibration problem, and the resulting calibration estimators are linear. An alternative solution using generalized inverses of matrices was implemented in the g-CALIB-S program (Vanderhoeft, 2001).

Bankier (1990) and Bankier, Rathwell & Majkowski (1992) also proposed discarding auxiliary variables to control weight variation while retaining the standard distance function 1. This solution leads to loss of calibration for discarded *x* variables, as well as to loss of control over which *x* variables shall be calibrated upon.

An additional problem encountered when many *x* variables are considered in the calibration is that of potential increase in the mean square error (MSE) of the resulting calibration estimator. Silva (1996, chapter 4) and Silva and Skinner (1997) showed that sometimes large numbers of auxiliary variables may actually reduce efficiency of the calibration (regression) estimator for small to moderate sample sizes. For example, under Simple Random Sampling without replacement (SRS) and assuming the model (2.11) to hold with $q_k = 1$ for every *k*, Silva (1996, p. 45) showed that

$$MSE_{SRS}(N^{-1}\hat{T}_y) = (1 - n/N)\frac{\sigma^2}{n}(1 + p/n) + \mathrm{O}(n^{-5/2}) \qquad (4.15)$$

where σ^2 is the variance of the residuals of the regression of *y* on **x**, and *p* is the number of *x* variables considered. This expression reveals that the MSE of a regression estimator can actually increase as the number of *x* variables increases, if the increase in the second order term *p/n* offsets the decrease in the variance of the residuals σ^2. Of course, this is not a problem if the sample is large, but for small to moderate samples, the number of auxiliary variables may have some noticeable effect on the MSE of the regression estimator.

As an illustration of the problem, Figure 2 plots the MSE of the regression estimator for increasing sets of auxiliary variables, assuming simple random sampling with n=100 from a population of heads of households for which data were collected as part of the test population census of Limeira, São Paulo state, Brazil, 1988.

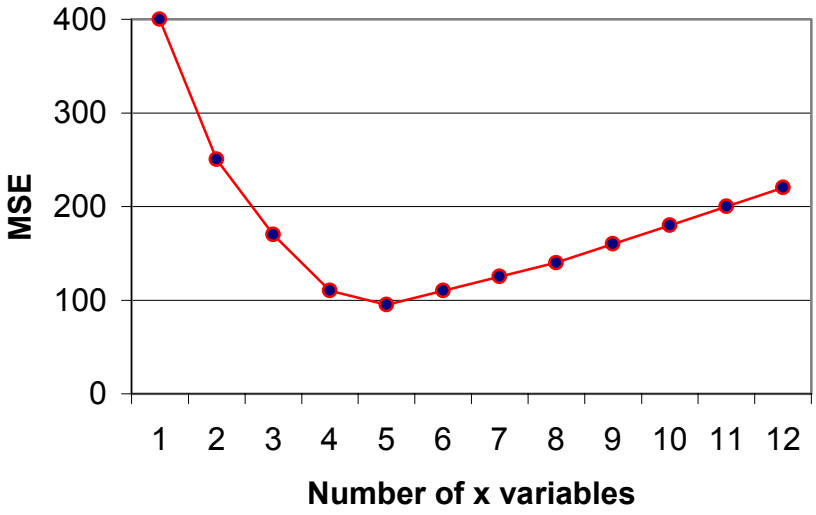

Figure 2 – MSE of regression estimator versus number of x variables

Silva and Skinner (1997) showed, in a limited simulation exercise, that regression estimation after subset selection may be more efficient than saturated regression estimation, for moderate sample sizes (n = 100; J = 5 ;10). Similar findings are reported by Clark (2002) for n = 100; 250 and J = 24; 40. Both sources reported also that the incidence of negative weights was smaller after subset selection than when the calibration used the saturated set of x variables. This suggests that part of the problem with negative weights comes from "excessive" calibration.

Although the idea of applying some form of variable selection procedure to select x variables for calibration may lead to more efficient estimators for some specified y variables, this approach is not problem-free. First, it leads to loss of calibration for discarded x variables. Second, the approach is intrinsically univariate, in the sense that subset selection is y-variable specific, which would imply different sets of weights for different y variables. Also, variance estimation becomes more difficult after subset selection, as noted in Silva and Skinner (1997). However, the message emerging from these studies is that performing calibration or regression estimation "automatically" with all the available auxiliary variables may not be an efficient strategy, particularly for samples with small to moderate sizes or when the number of auxiliary variables is large relative to the sample size. In such cases, devoting some attention to the selection of suitable subsets of the available variables is recommended, even if one is not prepared to use formal model selection procedures like those described by Silva & Skinner (1997). In repeated surveys, for example, one may devote effort in the first few rounds of the survey to establish an adequate set of variables to

calibrate upon, and then use the fixed subset for calibration in subsequent rounds of the survey.

Some other approaches have been proposed to handle the case of negative or extreme weights. Chambers (1997) proposed the so-called "ridge calibration" estimators, where the basic idea is to minimize the modified distance function

$$\sum_{k\in s}(w_k - d_k)^2 / d_k q_k + \frac{1}{\gamma}\left(\mathbf{T}_x - \hat{\mathbf{T}}_{xC}\right)'\boldsymbol{\Delta}\left(\mathbf{T}_x - \hat{\mathbf{T}}_{xC}\right) \tag{4.16}$$

where $\boldsymbol{\Delta}$ is a diagonal matrix of calibration error costs, and γ is a scalar ridging parameter to be specified. In this approach, there are no constraints to be satisfied. The resulting weights are given by

$$w_k = d_k\left[1 + q_k\left(\mathbf{T}_x - \hat{\mathbf{T}}_x\right)'\left(\gamma\boldsymbol{\Delta}^{-1} + \sum_{i\in s} q_i d_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}\mathbf{x}_k\right] \tag{4.17}$$

Note that in this approach a measure of the amount of calibration error is incorporated as the second term of the distance function. However, since there are no binding calibration constraints, the resulting weights and estimator are no longer guaranteed to avoid calibration error entirely. Some choice of γ such that all "ridge-calibrated" weights (4.17) are positive is always possible. One idea may be to choose the smallest γ satisfying this condition. Other approaches for choosing γ are discussed by Chambers (1997). Careful specification of the "calibration error cost matrix" $\boldsymbol{\Delta}$ allows flexible selection of subsets of auxiliary variables for which calibration error must be eliminated. To do so, it is sufficient to use very large diagonal elements in this matrix corresponding to the auxiliary variables for which calibration error must be zero. The approach is an improvement over procedures that discard auxiliary variables altogether, in the sense that some control over the amount of calibration error can be maintained for all *x* variables. Chambers (1997) considered other versions of ridge calibration that have as a starting point weights derived under either a model-based or a non-parametric model-based approach. He also considered outlier robust modifications of this approach that can be of help in cases where outliers on the *y* variables are of concern. These are *y*-variable specific, however, and hence shall not be considered further here.

Rao & Singh (1997) proposed yet another approach along similar lines, which is called "ridge-shrinkage calibration" estimation. Again the idea rests on minimizing a modified distance function, but this time, under range restrictions (boundary constraints). The

procedure is similar to the ridge calibration of Chambers (1997) if the initial solution satisfies the boundary constraints. If not, these boundary constraints are relaxed and the procedure iterates between adaptively modifying the calibration error cost matrix and the desired range restrictions until convergence.

Hedlin et al (2001) also discussed the problem of extreme calibration weights. This paper explored the behaviour of calibration (GREG) estimators when the underlying models were misspecified, and proposed some diagnostics measures to assess the adequacy of the model for a given survey situation. Part of the diagnostics considered the idea that the g-weights defined by (2.9) are functions of well-known measures of influence of a sample unit in the fitting of linear regression models. The diagnostics were used to locate the most extreme g-weights, and to propose remedies that involved, for example, poststratifying the sample and using regression or calibration estimation only for those sample units for which the g-weights are not extreme, and using the simple expansion estimator for the poststrata formed with the units with extreme g-weights.

The message, again, is that mere automatic calculation of the calibration weights is not good practice, and some attention must be devoted to analyse the resulting weights to assess whether the use of calibration or regression estimation is “safe” and efficient. One simple way to do this is to perform data analysis of the g-weights and to try and flag down those that are extreme in some sense. The most obvious cases are the negative or small g-weights (those leading to final calibrated weights less than 1) or the very large g-weights (say, with $g_k > U$). The cut-off point U may be determined arbitrarily (say, make $U = 5$ or 10), or by data-dependent methods ($U > Q_3+1.5(Q_3-Q_1)$), where Q_1 and Q_3 denote the sample quartiles of the g_k, $k \in s$.

4.5 - Calibration and nonresponse

So far we discussed calibration estimation under the assumption of complete observation of the selected sample. However, nonresponse is a pervasive problem. Most real-life surveys will experience some amount of nonresponse, despite incorporating well designed methods and procedures to prevent nonresponse. An important new issue brought about by nonresponse is that of bias. Standard Horvitz-Thompson (π-inverse weighted) estimators will be biased unless the nonresponse is completely at random, and even in this unlikely situation, estimation of totals requires at least some simple adjustment to compensate for the loss of sample units due to nonresponse.

Calibration is an useful approach to try and reduce the bias due to nonresponse. Lundström (1997) and Lundström & Särndal (1999) even suggest "calibration as a standard method for treatment of nonresponse". Calibration estimators are approximately design unbiased if there is complete response, for any fixed choice of auxiliary variables. Under nonresponse, however, calibration estimators may be biased even in large samples. Skinner (1999) examined the impact of nonresponse on calibration estimators. Some of his conclusions included the following:

- "the presence of nonresponse may be expected to lead to negative weights much more frequently";
- "the weights w_k will not converge to the original weights d_k as the sample size increases";
- "the variance of the calibration estimator will be dependent on the G_k functions and revised methods of variance estimation need to be considered".

The intended bias reduction by calibration will only be achieved, however, if the combined nonresponse and sampling mechanisms are ignorable given the *x* variables considered for calibration. This suggests that the choice of the *x* variables to be considered for calibration has to take account of the likely effects of nonresponse, and in particular, should aim to incorporate all *x* variables for which auxiliary population data is available that carry information about the unknown probabilities of responding to the survey. Under a simplified model where δ_k denotes the probability that a unit will respond to the survey given that it is selected into the sample, and response is independent for distinct units, a condition for the calibration estimator to be approximately unbiased under the joint sampling and response distribution is that $\delta_k^{-1} = 1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}$ for every *k* and some vector of constants $\boldsymbol{\lambda}$ (See Lundström, 1997, p. 46). However, because the calibration weights w_k are always of the form $w_k = d_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda})$ (see 4.14), it is easy to see that calibration will lead to approximately unbiased estimates when $w_k = d_k \delta_k^{-1}$, a condition that depends both on the choice of *x* variables and on the form of the distance (or calibration) function used to obtain the calibration weights.

One example where this question may be well illustrated comes from the weighting performed for the UK-LFS (see ONS, 2001, section 9). In this survey, weighting takes into account regional distribution (17 Regions or 454 Local Authorities), age (either 11 or 17 age groups), and sex of the sampled individuals. These are variables for which auxiliary

information is available at the population level from external sources. The number of *x* variables used for the calibration is fairly large (1,002) and the distance function chosen is type 2 in Table 1, corresponding to the weights implied by the raking ratio estimator. However, a study of the incidence of nonresponse in this survey showed that nonresponse is not completely at random, as indicated in the table in page 43 of the UK-LFS User Guide (ONS, 2001). It appears that the probability of responding depends on characteristics such as:

- Household composition;
- Employment status;
- Rent status and type of accommodation;
- Socio-economic status;
- Region of residence;
- Region of origin; e
- Marital status, age and education of head.

Clearly, then, one can see that calibration only on age, sex and region as is currently the case cannot hope to eliminate all bias due to nonresponse. It is not the number of *x* variable that matters, but rather having the right *x* variables! Of course this is easier said than done, and in the case of the UK-LFS, clearly there are difficulties. For example, if nonresponse is dependent on the Employment status, one could be tempted to try and calibrate on external information provided from register based sources such as the claimant count. For many of the other variables, though, reliable auxiliary population information may be unavailable, or hard or costly to obtain.

The message here is that it is not sufficient to calibrate on "all that is available" to be free from bias. Even more, Gambino (1999) suggests that in some cases calibration may even make the matters worse, and argues "it is well known that in many surveys, young males tend to be missed disproportionately. Since demographic estimates by age-sex are typically used in calibration, the effect is to increase the weights of the young males who happen to respond to the survey. If, for some variables of interest, the young males who tend to become respondents differ substantially from the young males that tend to be missed from the sample …, then the effect of calibration is to give more weight to a non-

representative component of the sample". Suppose we knew that the young males that then to be missed more frequently are those living alone, and those that are more likely to respond are those that are living with their parents or family. Hence the weighting should aim to increase the weights of those in the first group (young males living alone) but not of those in the second group (living with family). The crucial bit of information we would need to do that would be population totals by age and sex and household composition (single person households vs. other households). If this information is not available, there is still some limited remedy to be tried. Weighting could be performed at household level and not at the individual level. Hence young males living alone would have weights that depend on which type of household they live in, but this would not be the bias-correction that we would be aiming for, just the next best substitute given the available data.

The above example illustrates the issues one has to address concerning the use of calibration to compensate for nonresponse. If the response mechanism is dependent on household characteristics (apart from regional location), such as its size and composition, as well as those of the head, then perhaps the household should be the unit for which weights are computed, with individual members of the same household then receiving the household weight.

Gambino (1999) warns that for nonresponse adjustment, "poor choice of adjustment variables or classes can make matters worse", and concludes that "it is our duty as statisticians to work with the users to ensure that calibration tools are used wisely". One reason why calibration can make matters worse is because it may mask the effects of nonresponse. For example, using unadjusted sampling weights to estimate population counts by age and sex, one could locate the cells for which the estimates are under the expected level by an amount that is too large to be due to sampling error. These are the cells for which elements are more likely to be missed by the survey. Such estimates could then be used to detect the cells for which the likely effects of nonresponse are higher. But if estimates are computed only with calibrated weights, such deviations from the expected or known counts will not appear. It would take users extra effort to compute the pre-calibration estimates required to analyse the likely effects of nonresponse, if the pi-inverse weights d_k are made available together with the survey data.

5. CRITERIA TO ASSESS SUCCESS OF CALIBRATION

An important component of any statistical analysis or estimation job is the assessment of how well the adopted procedures performed in the application at hand. With calibration estimation applications, in addition to the usual estimation of variance that should

be routinely performed, we suggest that it is also important to assess a number of other aspects of the outcome. This is important to verify that some of the intended goals of calibration have been reached and to check for the potential problems that may affect the outcome.

As a first measure, we suggest examining the amount of calibration error remaining for the complete set of *x* variables that were initially selected for calibration. This should ideally be zero, if calibration error was eliminated entirely, but may be nonzero if some of the *x* variables were discarded during the process of determining the calibration weights, or if some of the approaches that do not lead to exact calibration are adopted. The average relative absolute calibration error for the estimated population totals of the *x* variables is given by

$$\mathrm{M1} = \frac{1}{p}\sum_{j=1}^{p}\left|\hat{T}_{x_jC} - T_{x_j}\right| / T_{x_j} \,. \tag{5.1}$$

A second measure, which is important to allow checking whether we need to be concerned about bias with the calibration estimator, is the average coefficient of variation of the Horvitz-Thompson estimates of the totals of the *x* variables, namely

$$\mathrm{M2} = \frac{1}{p}\sum_{j=1}^{p}\left[\hat{V}\left(\hat{T}_{x_j}\right)\right]^{1/2} / T_{x_j} \tag{5.2}$$

where $\hat{V}\left(\hat{T}_{x_j}\right)$ is an estimate of the variance of the Horvitz-Thompson estimator of the total of the jth auxiliary variable, given by

$$\hat{V}\left(\hat{T}_{x_j}\right) = \sum_{k\in s}\sum_{i\in s}\left(1-\pi_k\pi_i / \pi_{ki}\right)\left(d_k x_{kj}\right)\left(d_i x_{ij}\right). \tag{5.3}$$

Some other measures that may be of interest refer to the distribution of the g-weights. Two of these are the proportions of extreme (small or large) *g*-weights, where some definition of what the acceptable range is needed:

$$\mathrm{M3} = \frac{1}{n}\sum_{k\in s} I\left(g_k < L\right) \tag{5.4}$$

$$\mathrm{M4} = \frac{1}{n}\sum_{k\in s} I\left(g_k > U\right). \tag{5.5}$$

The coefficient of variation is another measure regarding the distribution of the g-weights that may be useful:

$$\mathrm{M5} = \sqrt{\frac{1}{n-1}\sum_{k\in s}\left(g_k - \bar{g}\right)^2} \,/\, \bar{g} \tag{5.6}$$

where $\bar{g} = \frac{1}{n}\sum_{k\in s} g_k$ is the average g-weight.

The distance between the g-weights and the d-weights is also an important measure, which we suggest should be considered:

$$\mathrm{M6} = \frac{1}{n}\sum_{k\in s}\left(w_k - d_k\right)^2 / q_k d_k = \frac{1}{n}\sum_{k\in s} d_k (g_k - 1)^2 / q_k \tag{5.7}$$

Note that we normalize this distance by dividing the calibration distance function by the sample size, so that it is easier to compare between samples of different sizes. Yet another possibility would be to divide by the sample size minus the number of *x* variables considered, which would allow for sets of auxiliary variables of different sizes. We suggest that the chi-square distance function be used even when the actual distance function that was minimized to obtain the calibration weights is one of the other functions in Table 1.

Last, but not the least, users should try and access the gains from calibration in comparison to standard Horvitz-Thompson estimators. This can be accomplished by comparing the average efficiency for a specified set of *y* variables, given by:

$$\mathrm{M7} = \frac{1}{J}\sum_{j=1}^{J} \hat{V}_g\left(\hat{T}_{y_j C}\right) / \hat{V}\left(\hat{T}_{y_j}\right) \tag{5.8}$$

where the variances in the numerators are estimated using (3.6) and those in the denominators are estimated using (5.3) with the *x* values replaced by the *y* values.

Using this set of seven measures and any others that might be providing information about the same aspects is strongly recommended for users of calibration estimation.

6. CONCLUDING REMARKS

In this report we reviewed the literature on calibration estimation, and tried to convey the message that it is a flexible, powerful and convenient approach to survey weighting and estimation. At the same time, we pointed out some of the difficulties that users of the

technique may face in practical applications, as well as provided some guidance on how these can be detected and circumvented.

The value of calibration estimation for practical survey situations is clearly demonstrated by the increasing number of software packages developed for its application as well as the number of major surveys in several countries that rely on calibration for their weighting and estimation. In countries like Canada, the United States, the United Kingdom, and France, calibration estimation is used for the labour force surveys. In Brazil and Canada, it is also used for the samples collected as part of the Population Censuses.

While this value is recognised, we would encourage users to be critical of the outcome of calibration weighting, and stress the importance of careful checking of the resulting weights and estimates, to see that they meet the performance criteria and high standards of quality and reliability that is expected from mainstream survey statistics. Calibration should not be used to "disguise" biased survey results, where coverage and nonresponse bias are "covered" by means of simple calibration to known population totals. Users of the technique should, above all, seek to be transparent about the methodology application and of the assessment of its success, in order to provide users of the survey results with the possibility to exercise a critical assessment of the fitness for purpose of the resulting statistics produced with calibration estimators.

Finally, we would encourage data producers that choose to adopt calibration weighting in surveys where microdata files are to be released, to provide secondary users with the information needed for them to make proper use of the data, in the sense of being able to compute point and variance estimates correctly. This is a lot more challenging than when usual Horvitz-Thompson estimators are used, because information on the full set of x variables used for calibration must also be released together with the original d-weights and the final w-weights for each survey record. Research is still needed on how best to accomplish this goal without sacrificing the necessary restrictions required for the protection of confidentiality.

7. REFERENCES

BANKIER, M. D. (1990). Generalized least squares estimation under poststratification. Proceedings of the Section on Survey Research Methods, American Statistical Association, 730-755.

BANKIER, M. D., Rathwell, S. & Majkowski, M. (1992, Statistics Canada, Methodology Branch Working Paper Series SSMD 92-007E) – Two step generalized least squares estimation in the 1991 Canadian Census.

BERGER, Y.G. (2002). A simple variance estimator for unequal probability sampling without replacement. Unpublished manuscript.

BERNIER, N. and Lavallée, P. (1994). The SAS macro: CALJACK. Ottawa: Statistics Canada, Social Survey Methods Division.

BETHLEHEM, J. G. (1988) Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics, 4*, 251-260.

BETHLEHEM, J. G. and Keller, W. J. (1987). Linear weighting of sample survey data. Journal of Official Statistics **3**, 141-153.

CHAMBERS, R. L. (1997). Weighting and calibration in sample survey estimation. In Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth, Birkhäuser Verlag Basel.

CLARK, R. G. (2002). Sample design and estimation for household surveys. Wollongong: University of Wollongong, School of Mathematics and Applied Statistics, unpublished Ph.D. dissertation.

COCHRAN, W.G. (1977). Sampling Techniques, 3rd edition. New York: John Wiley & Sons.

DEVILLE, J.C. & Särndal, C.E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, **87**, 376-382.

ESTEVAO, V., Hidiroglou, M.A. & Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. Journal of Official Statistics, **11**, 181-204.

GAMBINO, J. (1999). Issues in weighting household and business surveys – discussant comments. Proceedings of the International Statistical Institute.

HEDLIN, D. et al. (2001). Does the model matter for GREG estimation? A business survey example. Journal of Official Statistics, **17**, 527-544.

HOLMES, D.J. and Skinner, C.J. (2000). Variance estimation for Labour Force Survey estimates of level and change. GSS Methodology Series no. 21.

HUANG, E. T. & Fuller, W. A. (1978). Nonnegative regression estimation for sample survey data. Proceedings of the Social Statistics Section, American Statistical Association, 300-305.

LUNDSTRÖM, S. (1997). Calibration as a standard method for the treatment of nonresponse. Doctoral dissertation, Department of Statistics, University of Stockholm.

LUNDSTRÖM, S. & Särndal, C.E. (1999). Calibration as a standard method for the treatment of nonresponse. Journal of Official Statistics, **15**, 305-327.

NIEUWENBROEK, N. & Boonstra, H.J. (2002) – Bascula 4.0 for weighting sample survey data with estimation of variances. Survey Statistician, July, 2002.

ONS (2001). Labour Force Survey – User Guide – Volume 1 – LFS Background & Methodology 2001. London: Office for National Statistics.

RAO, J.N.K. & Singh, A.C. (1997) – Range restricted weight calibration for survey data. Unpublished manuscript.

SÄRNDAL, C.E., Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. Biometrika, **76**, 527-537.

SÄRNDAL, C.E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

SAUTORY, O. (1993) – La macro CALMAR: redressement d'un echantillon par calage sur marges, INSEE.

SILVA, P. L. d. N. (1996). Utilizing auxiliary information in sample survey estimation and analysis. Southampton: Univ. of Southampton, Department of Social Statistics, unpublished Ph.D. dissertation.

SILVA, P. L. d. N. and Skinner, C. J. (1997). Variable selection for regression estimation in finite populations. Survey Methodology, **23**, 23-32.

SKINNER, C. J. (1999). Calibration weighting and non-sampling errors. Research in Official Statistics, No. 1, 33-43.

VANDERHOEFT, C. (2001) – Generalised calibration at Statistics Belgium – SPSS module g-CALIB-S and current practices, Statistics Belgium Working Paper N. 3.