Survey Methodology

December 2007
How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-800-263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website at www.statcan.ca.

National inquiries line 1-800-263-1136
National telecommunications device for the hearing impaired 1-800-363-7629
Depository Services Program inquiries 1-800-700-1033
Fax line for Depository Services Program 1-800-889-9734
E-mail inquiries infostats@statcan.ca
Website www.statcan.ca

Accessing and ordering information

This product, catalogue no. 12-001-XIE, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

This product, catalogue no. 12-001-XPB, is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

<table>
<thead>
<tr>
<th></th>
<th>Single issue</th>
<th>Annual subscription</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>CAN$6.00</td>
<td>CAN$12.00</td>
</tr>
<tr>
<td>Other countries</td>
<td>CAN$10.00</td>
<td>CAN$20.00</td>
</tr>
</tbody>
</table>

All prices exclude sales taxes.

The printed version of this publication can be ordered by

- Phone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.ca
- Mail Statistics Canada
  Finance Division
  R.H. Coats Bldg., 6th Floor
  100 Tunney’s Pasture Driveway
  Ottawa (Ontario) K1A 0T6
- In person from authorised agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1-800-263-1136. The service standards are also published on www.statcan.ca under About us > Providing services to Canadians.
Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.
Survey Methodology
A Journal Published by Statistics Canada
Volume 33, Number 2, December 2007

Contents

In This Issue.................................................................................................................................................................................... 95

Waksberg Invited Paper Series

Carl-Erik Särndal
The calibration approach in survey theory and practice ................................................................. 99

Regular Papers

Seppo Laaksonen
Weighting for two-phase surveyed data ................................................................................................. 121

Pascal Ardilly and Pierre Lavallée
Weighting in rotating samples: The SILC survey in France ................................................................. 131

Jay J. Kim, Jianzhu Li and Richard Valliant
Cell collapsing in poststratification ................................................................................................. 139

Fulvia Mecatti
A single frame multiplicity estimator for multiple frame surveys .......................................................... 151

David Haziza
Variance estimation for a ratio in the presence of imputed data ............................................................ 159

James Chipperfield and John Preston
Efficient bootstrap for business surveys .......................................................................................... 167

Jacob J. Oleson, Chong Z. He, Dongchu Sun and Steven L. Sheriff
Bayesian estimation in small areas when the sampling design strata differ from the study domains .......... 173

Enrico Fabrizi, Maria Rosaria Ferrante and Silvia Pacci
Small area estimation of average household income based on unit level models for panel data.............. 187

Anne Renaud
Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results .......... 199

Marc N. Elliott and Amelia Haviland
Use of a web-based convenience sample to supplement a probability sample .................................. 211

Acknowledgements...................................................................................................................................................................... 217

In This Issue

This issue of *Survey Methodology* opens with the seventh paper in the annual invited paper series in honour of Joseph Waksberg. The editorial board would like to thank the members of the selection committee - Gordon Brackstone, chair, Bob Groves, Sharon Lohr and Wayne Fuller – for having selected Carl-Erik Särndal as the author of this year’s Waksberg Award paper. For this occasion, a special Workshop on Calibration and Estimation in Surveys (WCES) was organised on October 31st and November 1st at Statistics Canada. Professor Carl-Erik Särndal was the keynote speaker and presented his Waksberg paper. During the two days, 12 other speakers presented a paper and paid their tribute to Carl-Erik Särndal.

In his paper entitled “The Calibration Approach in Survey Theory and Practice” Särndal discusses the development and application of calibration in survey sampling. He describes the concept of calibration in some detail and contrasts it with generalized regression. He then describes different approaches to calibration including the minimum distance method, instrumental variables, and model calibration. Several examples of calibration and alternatives are considered.

Laaksonen discusses weighting in two phase sampling in which respondents to the first phase are asked if they are willing to participate in the second phase. The weighting thus has to deal with non-response at both phases of the survey, and also account for first-phase respondents who were unwilling to participate in the second phase. Using data from a Finnish survey on leisure-time activities, he empirically evaluates variations on a weighting method that uses response propensity modeling and calibration.

The article by Ardilly and Lavallée discusses the weighting problem for the SILC (Statistics on Income and Living Conditions) survey in France. This survey uses a rotating sample plan with nine panels. To obtain approximate estimators without bias, the authors relied on the weight-share method. Longitudinal weighting is discussed first, and then cross-sectional weighting is also discussed.

The paper by Kim, Li and Valliant deals with the problem of small cells or large weight adjustments when poststratification is used. The authors first describe several standard estimators and then introduce two alternative estimators based on cell collapsing. They study the performance of these estimators in terms of their effectiveness in controlling the coverage bias and the design variance. These properties are evaluated theoretically and also through a simulation study using a population based on the 2003 National Health Interview Survey.

Mecatti proposes a simple multiplicity estimator in the context of multi-frame surveys. She first shows that the proposed estimator is design-unbiased. Then, she proposes an unbiased estimator of the variance of the multiplicity estimator. Using 29 simulated populations, she compares the multiplicity estimator with alternative estimators proposed in the literature.

Haziza studies the problem of variance estimation for a ratio of two totals when marginal random hot deck imputation has been used to fill in missing data. Two approaches to inference are considered, one using an imputation model and a second one using a nonresponse model. Variance estimators are derived under two frameworks: the reverse approach of Shao and Steel (1999) and the traditional two-phase approach.

In their paper, Chipperfield and Preston describe the without replacement scaled bootstrap variance estimator that was implemented in the Australian Bureau of Statistics’ generalized estimation system ABSEST. The without replacement scaled bootstrap estimator is shown to be more efficient than the with replacement scaled bootstrap estimator for stratified samples when the stratum sizes are small. In addition, the without replacement scaled bootstrap estimator was shown to require fewer replicates to achieve the same replication error as the with replacement estimator. For the ABSEST system, bootstrap variance estimators were chosen over other variance estimation methods for their computational efficiency and the without replacement bootstrap was selected for the reasons above.
Oleson, He and Sun describe a Bayesian modelling approach for situations where the sampling design is stratified and the estimation procedure requires post-stratification. The method is illustrated with data from the 1998 Missouri Turkey Hunting Survey for which the strata were defined by the hunter’s place of residence but estimates were required at the county level.

Fabrizi, Ferrante and Pacei discuss a methodology which is increasingly important in modern sample survey applications. They investigate the effect of borrowing strength from additional panel information for cross-sectional household income estimates for small areas in Italy. The proposed methods seem to tackle a problem which may have further relevance for European Official Statistics, and possibly also in the area of small area statistics for indicators which may be used for policy research.

Renaud presents an interesting application of a post-enumeration survey to estimate net undercoverage in the 2000 census in Switzerland. The objective of this survey was slightly different from that of other countries in that it was not designed to adjust the Census counts for net undercoverage, but rather to gather information to improve the quality of subsequent censuses.

In the final paper, Elliot and Haviland consider combining a convenience sample with a probability based sample to obtain an estimate with a smaller MSE. The resulting estimator is a linear combination of the convenience and probability sample estimates with weights that are a function of the bias. By looking at the maximum incremental contribution of the convenience sample, they show that improvement to the MSE may be attainable only in certain circumstances.

Harold Mantel, Deputy Editor
Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work. The author receives a cash award made possible by a grant from Westat, in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially by the American Statistical Association. Previous winners are listed below. Their papers in the series have already appeared in *Survey Methodology*.

**Previous Waksberg Award Winners:**

Gad Nathan (2001)
Wayne A. Fuller (2002)
Norman Bradburn (2004)
Alastair Scott (2006)
Carl-Erik Särndal (2007)

**Nominations:**

The author of the 2009 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, Robert Groves, by email to bgroves@isr.umich.edu. Nominations and suggestions for topics must be received by February 29, 2008.

**2007 Waksberg Invited Paper**

**Author: Carl-Erik Särndal**

Carl-Erik Särndal, retired professor in the Université de Montréal, is a consultant and expert who has been associated with several national statistical institutes, in particular Statistics Canada and Statistics Sweden as well as Statistics Finland, INSEE and Eurostat. His list of publications comprises three books, including the very well known Model Assisted Survey Sampling book that has had a major impact. He is also the author of numerous scientific articles, in sole authorship or in collaboration with researchers from many countries. His research interest in survey sampling has been very diversified, but most often revolved around ways to best using auxiliary information in sampling and estimation.
Members of the Waskberg Paper Selection Committee (2007-2008)

Robert Groves, (Chair)
Wayne A. Fuller, Iowa State University
Daniel Kasprzyk, Mathematica Policy Research
Leyla Mojadji, Westat

Past Chairs:

Graham Kalton (1999 - 2001)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
The calibration approach in survey theory and practice

Carl-Erik Särndal

Abstract

Calibration is the principal theme in many recent articles on estimation in survey sampling. Words such as “calibration approach” and “calibration estimators” are frequently used. As article authors like to point out, calibration provides a systematic way to incorporate auxiliary information in the procedure.

Calibration has established itself as an important methodological instrument in large-scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources.

This paper presents a review of the calibration approach, with an emphasis on progress achieved in the past decade or so. The literature on calibration is growing rapidly; selected issues are discussed in this paper.

The paper starts with a definition of the calibration approach. Its important features are reviewed. The calibration approach is contrasted with (generalized) regression estimation, which is an alternative but conceptually different way to take auxiliary information into account. The computational aspects of calibration are discussed, including methods for avoiding extreme weights. In the early sections of the paper, simple applications of calibration are examined: The estimation of a population total in direct, single phase sampling. Generalization to more complex parameters and more complex sampling designs are then considered. A common feature of more complex designs (sampling in two or more phases or stages) is that the available auxiliary information may consist of several components or layers. The uses of calibration in such cases of composite information are reviewed. Later in the paper, examples are given to illustrate how the results of the calibration thinking may contrast with answers given by earlier established approaches. Finally, applications of calibration in the presence of nonsampling error are discussed, in particular methods for nonresponse bias adjustment.

Key Words: Auxiliary information; Weighting; Consistency; Design-based inference; Regression estimator; Models; Nonresponse; Complex sampling design.

1. Introduction

1.1 Calibration defined

It is useful in this paper to refer to a definition of the calibration approach. I propose the following formulation.

Definition. The calibration approach to estimation for finite populations consists of

(a) a computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s),

(b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units,

(c) an objective to obtain nearly design unbiased estimates as long as nonresponse and other nonsampling errors are absent.

In the literature, “calibration” frequently refers to (a) alone; I shall often use the term for (a) to (c) together. Earlier definitions, although less extensive, agree essentially with mine. Ardilly (2006) defines calibration (or, more precisely, “calage généralisé”) as a method of re-weighing used when one has access to several variables, qualitative or quantitative, on which one wishes to carry out, jointly, an adjustment.

Kott (2006) defines calibration weights as a set of weights, for units in the sample, that satisfy a calibration to known population totals, and such that the resulting estimator is randomization consistent (design consistent), or, more rigorously, that the design bias is, under mild conditions, an asymptotically insignificant contribution to the estimator’s mean squared error. This is the property I call “nearly design unbiased”.

The Quality Guidelines (fourth edition) of Statistics Canada (2003) say: “Calibration is a procedure than can be used to incorporate auxiliary data. This procedure adjusts the sampling weights by multipliers known as calibration factors that make the estimates agree with known totals. The resulting weights are called calibration weights or final estimation weights. These calibration weights will generally result in estimates that are design consistent, and that have a smaller variance than the Horvitz-Thompson estimator.”

Part (c) of the definition merits a comment. Nothing prevents producing weights calibrated to given auxiliary information without requiring (c). But most published work on calibration is in the spirit of (c), so it makes good sense to include it. When non-sampling errors are present, bias in the estimates is unavoidable, whether they are made by calibration or by any other method. In line with (c), I
consider design-based inference to be the standard in this paper. The randomization-based variance of an estimator is thus important. However, the paper focuses on “motivations behind (point) estimation”; for reasons of space, the important question of variance estimation is not addressed.

1.2 Comments arising

The definition in Section 1.1 prompts some comments and references to earlier literature:

(1) **Calibration as a linear weighting method.** Calibration has an intimate link to practice. The fixation on weighting methods on the part of the leading national statistical agencies is a powerful driving force behind calibration. To assign an appropriate weight to an observed variable value, and to sum the weighted variable values to form appropriate aggregates, is firmly rooted procedure. It is used in statistical agencies for estimating various descriptive finite population parameters: totals, means, and functions of totals. Weighting is easy to explain to users and other stakeholders of the statistical agencies.

Weighting of units by the inverse of their inclusion probability found firm scientific backing long ago in papers such as Hansen and Hurwitz (1943), Horwitz and Thompson (1952). Weighting became widely accepted. Later, post-stratification weighting achieved the same status.

Calibration weighting extends both of these ideas. Calibration weighting is outcome dependent; the weights depend on the observed sample. Inverse inclusion probability weights are, by definition, greater than or equal to unity. A commonly heard interpretation is that “an observed unit represents itself and a number of others, not observed”. Calibrated weights, on the other hand, are not necessarily greater than or equal to unity, unless special care is taken in the computation to obtain this property.

Calibration is new as a term in survey sampling - about 15 years old - but not as a technique for producing weights. Those who maintain “I practiced calibration long before it was called calibration” have a point. The last 15 years widened the scope and the appeal of the technique. Weighting akin to calibration has long been used by private survey institutes, for example, in connection with quota sampling, a form of non-probability sampling outside the scope of this paper.

Weighting of observed variable values was an important topic before calibration became a popular term. Some authors derived the weights via the argument that they should differ as little as possible from the unbiased sampling design weights (the inverse of the inclusion probabilities). Others found the weights by recognizing that a linear regression estimator can be written as a linearly weighted sum of the observed study variable values. Terms such as “survey sample weighting” and “regression weighting” and “case weighting” are used. Among such “early papers” are Alexander (1987), Bankier, Rathwell and Majkowski (1992), Bethlehem and Keller (1987), Chambers (1996), Fuller, Loughlin and Baker (1994), Kalton and Flores-Cervantes (1998), Lemaître and Dufour (1987), Särndal (1982) and Zieschang (1990). I comment later on the technique “repeated weighting”, promoted by the Dutch national statistical agency, CBS. The newer term “calibration” conveys a more specific message and a more definite direction than the older “weighting”.

(2) **Calibration as a systematic way to use auxiliary information.** Calibration provides a systematic way to take auxiliary information into account. As Rueda, Martínez, Martínez and Arcos (2007) point out, “in many standard settings, the calibration provides a simple and practical approach to incorporating auxiliary information into the estimation”.

Auxiliary information was used to improve the accuracy of survey estimates long before calibration became popular. Numerous papers were written with this goal in mind, for more or less specialized situations. Today, calibration does offer a systematic outlook on the uses of auxiliary information. For example, calibration can deal effectively with surveys where auxiliary information exists at different levels. In two-stage sampling information may exist for the first stage sampling units (the clusters), and other information for the second stage sampling units. In surveys with nonresponse (that is, essentially all surveys), information may exist “at the population level” (known population totals), and other information “at the sample level” (auxiliary variable values for all those sampled, responding and non-responding). Calibration with “composite information” is reviewed in Sections 8 and 9.

Regression estimation, or generalized regression (GREG) estimation, competes with calibration as a systematic way to incorporating auxiliary information. It is therefore important to contrast GREG estimation (described in Section 3) with calibration estimation (described in Section 4). The two approaches are different.

(3) **Calibration to achieve consistency.** Calibration is often described as “a way to get consistent estimates”. (Here “consistent” refers not to “randomization consistent” but to “consistent with known aggregates”.) The calibration equations impose consistency on the weight system, so that, when applied to the auxiliary variables, it will confirm (be consistent with) known aggregates for those same auxiliary variables. A desire to promote credibility in published statistics is an often cited reason for demanding consistency. Some users of statistics dislike finding the same population...
Consistency among tables estimated from different surveys is the motive behind repeated weighting, the technique developed at the Dutch national statistical agency CBS in several articles: Renssen and Nieuwenbroek (1997); Nieuwenbroek, Renssen, and Hofman (2000); Renssen, Kroese, and Willeboordse (2001); Knottnerus and van Duin (2006). The stated objective is to accommodate user demands to produce numerically consistent outputs. As the last mentioned paper points out, repeated weighting can be seen as an additional calibration step for a new adjustment of already calibrated weights. The final weights realize consistency with given margins.

Consistency with known or estimated totals may bring the extra benefit of improved accuracy (lower variance and/or reduced nonresponse bias). However, in some articles, especially those authored in statistical agencies, consistency for user satisfaction seems a more imperative motivation than the prospect of increased accuracy.

When the primary motivation for calibration is not so much an agreement with other statistics as rather to reduce variance and/or nonresponse bias, then “balanced weight system” is a more appropriate description than “consistent weight system”, because the objective is then to balance the weights to reflect the outcome of the sampling, the response to the survey, and the information available.

(4) Calibration for convenience and transparency. As Harms and Duchesne (2006) point out, “The calibration approach has gained popularity in real applications because the resulting estimates are easy to interpret and to motivate, relying, as they do, on design weights and natural calibration constraints.” Calibration on known totals strikes the typical user as transparent and natural. Users who understand sample weighting appreciate that calibration leaves the design weights “slightly modified only”, while respecting the controls. The unbiasedness is only negligibly disturbed. The simpler forms of calibration invoke no assumptions, only “natural constraints”. Yet another advantage is appreciated by users: In many applications, calibration gives a unique weighting system, applicable to all study variables, of which there are usually many in large government surveys.

(5) Calibration in combination with other terms. Some authors use the word “calibration” in combination with other terms, to describe various directions of thought. Examples of this proliferation of terms are: Model-calibration (Wu and Sitter 2001); g-calibration (Vanderhoeft, Waeytens and Museux 2000); Harmonized calibration (Webber, Latouche and Rancourt 2000), Higher level calibration (Singh, Horn and Yu 1998); Regression calibration (Demmati and Rao 2004); Non-linear calibration (Plikusas 2006); Super generalized calibration (Calage super généralisé; Ardilly 2006); Neural network model-calibration estimator and Local polynomial model-calibration estimator (Montanari and Ranalli 2003, 2005), Model-calibrated pseudo empirical maximum likelihood estimator (Wu 2003), and yet others. Also, calibration plays a significant role in the indirect sampling methods proposed in Lavallée (2006). In a somewhat different spirit, not reviewed here, are concepts such as calibrated imputation (Beaumont 2005a), and bias calibration (Chambers, Dorfman and Wehrly 1993), Zheng and Little (2003)). The following review pages do not give justice to all the innovations within the sphere of calibration, but the names alone do suggest directions that have been explored.

(6) Calibration as a new direction for thought. If calibration represents “a new approach” with clear differences compared with predecessors, we must examine such questions as: Does calibration generalize earlier theories or approaches? Does calibration give better, more satisfactory answers on questions of importance, as compared with earlier recognized approaches? Sections 4.5 and 7.1 in this paper illustrate how the answers provided by calibration compare with, or contrast with, those obtained in earlier modes of reasoning.

The practice of survey sampling encounters “nuisances” such as nonresponse, frame deficiencies and measurement errors. It is true that imputation and reweighing for nonresponse are widely practiced, through a host of techniques. But they are somehow “separate issues”, still waiting to be more fully embedded into a comprehensive, more satisfactory theory of inference in sample surveys. Many theory papers deal with estimation for an imagined ideal survey, nonexistent in practice, where nonresponse and other non-sampling errors are absent. This is not a criticism of the many excellent but idealized theory papers. The foundations need to be explored, too.

Sections 9 and 10 indicate that calibration can provide a more systematic outlook on inference in surveys even in the presence of the various non-sampling errors. Future fruitful developments are expected in that regard.
2. Basic conditions for design-based estimation in sample surveys

This section sets the background for Sections 3 to 7. By “basic conditions” I will mean single phase probability sampling of elements and full response. In practice, survey conditions are not that simple and perfect, but many theory papers nevertheless address this situation.

A probability sample s is drawn from the finite population \( U = \{1, 2, ..., k, ..., N\} \). The probability sampling design generates for element \( k \) a known inclusion probability, \( \pi_k > 0 \), and a corresponding sampling design weight \( d_k = 1/\pi_k \). The value of the study variable \( y \) is recorded for all \( k \in s \) (complete response). The objective is to estimate a population total \( Y = \sum_{U} y_k \) with the use of auxiliary information. The study variable \( y \) may be continuous or, as in many government surveys, categorical. For example, if \( y \) is dichotomous with value \( y_k = 0 \) or \( y_k = 1 \) according as person \( k \) is employed or unemployed, then the parameter \( Y = \sum_{U} y_k \) to be estimated is the population count of unemployed people. (If \( A \subset U \) is a set of elements, I write \( \sum_A \) for \( \sum_{k \in A} \)). The basic design unbiased estimator of \( Y \) is \( \hat{Y}_{HT} = \sum d_k y_k \), the Horwitz-Thompson estimator. It is, however, inefficient when powerful auxiliary information is available for use at the estimation phase.

The general notation for the auxiliary vector will be \( \mathbf{x}_k \).

In some countries, for some surveys, the sources of auxiliary data permit extensive vectors \( \mathbf{x}_k \) to be built. But some examples of simple vectors are: (1) \( \mathbf{x}_k = (1, x_k)' \), where \( x_k \) is the value for element \( k \) of a continuous auxiliary variable \( x \); (2) the classification vector used to code membership in one of \( P \) mutually exclusive and exhaustive groups, \( \mathbf{x}_k = \gamma_k = (\gamma_{k1}, ..., \gamma_{kp})' \), so that, for \( p = 1, 2, ..., P \), \( \gamma_{kp} = 1 \) if \( k \) belongs to group \( p \), and \( \gamma_{kp} = 0 \) if not; (3) the combination of (1) and (2), \( \mathbf{x}_k = (\gamma_k', x_k)' \); (4) the vector \( \mathbf{x}_k \) that codifies two classifications stringed out ‘side-by-side’, the dimension of \( \mathbf{x}_k \) being \( P + Q - 1 \), where \( P \) and \( Q \) are the respective number of categories, and the ‘minus-one’ is to avoid a singular matrix in the computation of weights calibrated “to the margins”; (5) the extension of (4) to more than two ‘side-by-side’ categorical classifications. Cases 4 and 5 are particularly important for production in national statistical agencies.

In calibration reasoning it is crucially important to specify exactly the auxiliary information. Under the basic conditions we need to distinguish two different cases relative to \( \mathbf{x}_k \):

(i) \( \mathbf{x}_k \) is a known vector value for every \( k \in U \) (complete auxiliary information)

(ii) \( \sum_U \mathbf{x}_k \) is known (imported) total, and \( \mathbf{x}_k \) is known (observed) for every \( k \in s \)

It is often the survey environment that dictates whether (i) or (ii) prevails. Case (i), complete auxiliary information, occurs when \( \mathbf{x}_k \) is specified in the sampling frame for every \( k \in U \) (and thus known for every \( k \in s \)). This environment is typical of surveys on individuals and households in Scandinavia and other North European countries equipped with high quality administrative registries that can be matched with the frame to provide a large number of potential auxiliary variables. The population total \( \sum_U \mathbf{x}_k \) is obtained simply by adding the \( \mathbf{x}_k \).

Case (i) gives considerable freedom in structuring the auxiliary vector \( \mathbf{x}_k \). For example, if \( \mathbf{x}_k \) is a continuous variable value specified for every \( k \in U \), then we are invited to consider \( x_k^2 \) and other functions of \( x_k \) for inclusion in \( \mathbf{x}_k \), because totals such as \( \sum_U x_k^2 \) and \( \sum_U \log x_k \) are readily computed. If the relationship to the study variable \( y \) is curved, it may be a serious omission not to take into account known totals such as the quadratic one or the logarithmic one.

Case (ii) prevails in surveys where (i) is not met, but where \( \sum_U \mathbf{x}_k \) is imported from an outside source considered accurate enough, and the individual value \( x_k \) is available (observed in data collection) for every \( k \in s \). Then \( \sum_U \mathbf{x}_k \) is sometimes called an “independent control total”, to mark its origin from outside the survey itself. Case (ii) is less flexible: if \( x_k \) is a variable with a total \( \sum_U x_k \) imported from a reliable source, then \( \sum_U x_k^2 \) may be unavailable, barring \( x_k^2 \) from inclusion into \( \mathbf{x}_k \).

3. Generalized regression estimation under the basic conditions

3.1 The GREG concept

Before examining calibration, let us consider generalized regression (GREG) estimation (or just regression estimation), for two good reasons: (1) GREG estimation can also be claimed to be a systematic way to take auxiliary information into account; (2) some (but not all) GREG estimators are calibration estimators, in that they can be expressed in terms of a calibrated linear weighting.

GREG estimators and calibration estimators have been extensively studied in the last two decades. The terms alone, “GREG estimation” and “calibration estimation”, reflect a clear difference in thinking. Statisticians who work in the area are of two types: Those dedicated to “GREG thinking” and those dedicated to “calibration thinking”. The distinction may not be completely clear-cut, but it helps structuring this review paper, so I will use it. I am not venturing to say that the latter thinking is more prevalent in national statistical agencies and the former more prevalent in the academic circles, but perhaps there is such a tendency.
The GREG estimator concept evolved gradually since the mid-1970’s. The simple (linear) GREG is explained in Särndal, Swenson, and Wretman (1992); a thorough review of regression estimation is given in Fuller (2002). The central idea is that predicted \( y \)-values \( \hat{y}_k \) can be produced for all \( N \) population elements, via the fit of an assisting model and the use of the auxiliary vector values \( x_k \) known for all \( k \in U \). The predicted values serve to build a nearly design unbiased estimator of the population total \( Y = \sum_U y_k \) as

\[
\hat{Y}_{\text{GREG}} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k)
= \sum_s d_k y_k + (\sum_U \hat{y}_k - \sum_s d_k \hat{y}_k). \tag{3.1}
\]

The obvious motivation behind this construction is the prospect of a highly accurate estimate \( \hat{Y}_{\text{GREG}} \) through a close fitting assisting model that leaves small residuals \( y_k - \hat{y}_k \). That modeling is the corner stone of GREG thinking. Some authors use the (also justifiable) name general difference estimator for the construction (3.1).

The great variety of possible assisting models generates a wide family of GREG estimators of the form (3.1). The assisting model, an imagined relationship between \( x \) and \( y \), can have many forms: linear, non-linear, generalized linear, mixed (model with some fixed, some random effects), and so on. Whatever the choice, the model is “assisting only”; even though it may be short of “true”, (3.1) is nearly unbiased under mild conditions on the assisting model and on the sampling design, so that \( (\hat{Y}_{\text{GREG}} - Y)/N = O_p(n^{-1/2}) \) and \( (\hat{Y}_{\text{GREG}} - Y)/N = (\hat{Y}_{\text{GREG,lin}} - Y)/N + O_p(n^{-1}) \), where the statistic \( \hat{Y}_{\text{GREG,lin}} \), the result of linearizing \( \hat{Y}_{\text{GREG}} \), is unbiased for \( Y \).

\[3.2 \quad \text{Linear GREG}\]

By linear GREG I mean one that is generated by a linear fixed effects assisting model. The predictions are \( \hat{y}_k = x_k' B_{s, dq} \) with

\[
B_{s, dq} = \left( \sum_s d_k x_k x_k' \right)^{-1} \left( \sum_s d_k q_k x_k y_k \right)
\]

so (3.1) becomes

\[
\hat{Y}_{\text{GREG}} = \left( \sum_U x_k \right)' B_{s, dq} + \sum_s d_k (y_k - x_k' B_{s, dq}). \tag{3.2}
\]

The \( q_k \) are scale factors, chosen by the statistician. The standard choice is \( q_k = 1 \) for all \( k \). The choice of the \( q_k \) has some (but often limited) impact on the accuracy of \( \hat{Y}_{\text{GREG}} \). Near-unbiasedness holds for any specification (barring outrageous choices) for the \( q_k \). Although the model is simple, the linear GREG (3.2) contains many estimators, considering the many possible choices of the auxiliary vector \( x_k \) and the scale factors \( q_k \). Under general conditions, \( \hat{Y}_{\text{GREG}} - Y)/N = (\sum_s d_k E_k - \sum_U E_k)/N + O_p(n^{-1}) \)

where \( \sum_s d_k E_k \) is the Horvitz-Thompson estimator in the residuals \( E_k = y_k - x_k' B_{U, dq} \) with \( B_{U, dq} = (\sum_U q_k x_k x_k')^{-1} \). Hence, the design-based properties \( E(\hat{Y}_{\text{GREG}}) = Y \) and \( \text{Var}(\hat{Y}_{\text{GREG}}) = \text{Var}(\sum_U d_k E_k) \). A close fitting linear regression of \( y \) on \( x \) holds the key to a small variance for \( \hat{Y}_{\text{GREG}} \) (and this is very different from claiming that “a linear regression is the true regression”).

The linear GREG in Särndal, Swenson, and Wretman (1992) was motivated via the linear assisting model \( \xi \) stating that \( E_k(y_k) = \beta x_k \) and \( V_k(y_k) = \sigma_k^2 \). Generalized least squares fit gives the estimator (3.2) with \( q_k = 1/\sigma_k^2 \). In that context, an educated guess about the variation of the residuals \( y_k - \beta x_k \) determines the \( q_k \). When the vector \( x_k \) is fixed, the modeling effort boils down to an opinion about the residual pattern. The choice \( \sigma_k^2 = \sigma^2 x_k \) gives the classical ratio estimator. If \( q_k = \mu' x_k \) for all \( k \in U \) and a constant vector \( \mu \), then (3.2) reduces to “the cosmetic form” \( (\sum_k x_k)' B_{s, dq} \).

As Beaumont and Alavi (2004) and others have pointed out, the linear GREG estimator is bias-robust (nearly unbiased although the assisting model falls short of “correct”), but it can be considerably less efficient (have larger mean squared error) than model dependent alternatives which, although biased, may have a considerably smaller variance. Thus one may claim that linear GREG is not variance robust; nevertheless, it is a basic concept in design-based survey theory.

The specification of \( x_k \) should include variables (with known population totals) that served already in defining the sampling design. Design stage information should not be relinquished at the estimation stage; instead, a “repeated usage” is recommended. For example, in stratified simple (STS) random sampling, the vector \( x_k \) in estimator (3.2) should include, along with other available variables, the dummy coded stratum identifier, \( \gamma_k = (\gamma_{k1}, \gamma_{k2}, ..., \gamma_{kH})' \), where \( \gamma_{kh} = 1 \) if element \( k \) belongs to stratum \( h \), and \( \gamma_{kh} = 0 \) if not; \( h = 1, ..., H \).

We can write the linear GREG (3.2) as a weighted sample sum, \( \hat{Y}_{\text{GREG}} = \sum_k w_k y_k \), with

\[
(\hat{Y}_{\text{GREG}} - Y)/N = (\sum_s d_k E_k - \sum_U E_k)/N + O_p(n^{-1})
\]

\[
w_k = d_k g_k; \quad g_k = 1 + q_k \lambda' x_k; \quad \lambda' = (\sum_U x_k - \sum_s d_k x_k')^{-1} (\sum_s d_k q_k x_k x_k')^{-1}. \tag{3.3}
\]

The weights \( w_k \) happen to be calibrated to (consistent with) the known population \( x \)-total: \( \sum_k w_k x_k = \sum_U x_k \). That \( \hat{Y}_{\text{GREG}} \) is expressible as a linearly weighted sum with calibrated weights is a fortuitous by-product. It is not part of GREG thinking, whose central idea formulated in (3.1) is the fit of an assisting model. A few other GREG’s than the

Statistics Canada, Catalogue No. 12-001-X
The calibration approach in survey theory and practice

3.3 Non-linear GREG

Two features of the linear GREG (3.2) make it a favourite choice for routine production in statistical agencies: (i) the auxiliary population total \( \sum U_i x_k \) becomes factored out, so the estimation can proceed as long as an accurate value for that total can be computed or imported, and (ii) when written as the linearly weighted sum \( \hat{Y}_{GREG} = \sum w_i y_i \), the weight system (3.3) is independent of the \( y \)-variable and can thereby be applied to all \( y \)-variables in the survey. We need not know \( x_k \) individually for all \( k \in U \); knowing \( \sum U_i x_k \) suffices. Needless to say, if we do not know all \( x_k \), more efficient (still nearly design unbiased) members of the GREG family (3.1) can be sought. This will also counter another criticism of the linear GREG, namely that a linear model is unrealistic for some types of data. For example, for a dichotomous \( y \)-variable, a logistic assisting model may be both more realistic and yield a more precise GREG estimator.

By a non-linear GREG estimator I mean one generated as in (3.1) by an assisting model of other type than “linear in \( x_k \) with fixed effects”. Among the first to extend the GREG concept in this direction are Firth and Bennett (1998) and Lehtonen and Veijanen (1998); see also Chambers et al. (1993). In the last few years, several authors have studied model-assisted non-linear GREG’s.

Non-linear GREG is a versatile idea; a variety of estimators become possible via assisting models \( \xi \) of the following type:

\[
E_{\xi}(y_k | x_k) = \mu_k \quad \text{for} \quad k \in U
\]

where the mean model \( \mu_k \) and the model variance \( V_{\xi}(y_k | x_k) \) are given appropriate formulations.

One application of (3.4) is when \( \mu_k = \mu(x_k, \theta) \) is a specified non-linear function in \( x_k \). Having estimated \( \theta \) by \( \hat{\theta} \), the fitted values needed for \( \hat{Y}_{GREG} \) in (3.1) are \( \hat{y}_k = \mu(x_k, \hat{\theta}) \) for \( k \in U \). For example, if the modeler specifies \( \log \mu_k = \alpha + \beta x_k \), the predictions for use in (3.1) are, following parameter estimation, \( \hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k) \).

Other applications of (3.4) include generalized linear models such that \( g(\mu_k) = x_k' \theta, \) for a specified link function \( g(\cdot) \), and \( V_{\xi}(y_k | x_k) = \nu(\mu_k) \) is given an appropriate structure. We estimate \( \theta \) by \( \hat{\theta} \), the fitted values needed for the non-linear GREG estimator (3.1) are \( \hat{y}_k = \hat{\mu}_k = g^{-1}(x_k' \hat{\theta}) \). For example, using a logistic assisting model, \( x_k' \theta = \logit(\mu_k) = \log(\mu_k / (1 - \mu_k)) \), and \( \hat{y}_k = \hat{\mu}_k = \exp(x_k' \hat{\theta}) / (1 + \exp(x_k' \hat{\theta})) \).

Lehtonen and Veijanen (1998) examine the case of a categorical study variable with \( I \) classes, \( i = 1, 2, ..., I \), \( y_{ik} = 1 \) if element \( k \) belongs to category \( i \), and \( y_{ik} = 0 \) if not. For example, in a Labour Force Survey with \( I = 3 \) categories, “employed”, “not employed” and “not in the labour force”, an objective is to estimate the respective population counts \( Y_i = \sum U_i y_{ik} \), \( i = 1, 2, 3 \). These authors use the logistic assisting model

\[
E_{\xi}(y_{ik} | x_k, \mu) = \mu_{ik}; \quad \mu_{ik} = \exp(x_k' \theta) \left( 1 + \sum_{i=2}^I \exp(x_k' \theta) \right)^{-1}.
\]

Estimates \( \hat{\theta}_j \) of the \( \theta_j \) are obtained by maximizing design-weighted log-likelihood. The resulting predictions \( \hat{y}_{ik} = \hat{\mu}_{ik} \) are used to form \( \hat{Y}_{GREG} = \sum U_i \hat{y}_{ik} + \sum d_i (y_{ik} - \hat{y}_{ik}) \), for \( i = 1, 2, ..., I \).

Another development is the application of GREG reasoning to estimation for domains, as in Lehtonen, Särndal and Veijanen (2003, 2005) and Myrskylä (2007). Mixed models are used in the first two of these papers to assist the non-linear GREG. Let \( U_a \) be a domain, \( U_a \subset U \), whose total \( Y_{ia} = \sum U_i y_{ik} \) we wish to estimate, \( i = 1, 2, ..., I \). The 2005 paper derives the predictions for the non-linear GREG from the logistic mixed model stating that for \( k \in U_a \)

\[
E_{\xi}(y_{ik} | x_k; \theta_a) = \exp(x_k' \theta_a) \left( 1 + \sum_{i=2}^I \exp(x_k' \theta_a) \right)^{-1}.
\]

with \( \theta_a = \beta + u_a \), where \( u_a \) is a vector of domain specific random deviations from the fixed effects vector \( \beta \).

Non-linear GREG’s assisted by models such as (3.5) and (3.6) require model fitting for every \( y \)-variable separately; there is no uniformly applicable weight system. However, the question arises: Are there examples of non-linear GREG’s such that the practical advantages of linear GREG are preserved, that is, a linearly weighted form with calibrated weights independent the \( y \)-variable. The answer is in the affirmative. Two directions in recent literature are of interest in this regard:

Breidt and Opsomer (2000), Montanari and Ranalli (2005) consider model-assisted local polynomial GREG estimators, for the case of a single continuous auxiliary variable with values \( x_k \) known for all \( k \in U \). Several choices have to be made in the process: (1) the order \( q \) of the local polynomial expression, (2) the specification of the kernel function, and (3) the value of the bandwidth. The resulting estimator can be expressed in terms of weights calibrated with respect to population totals of the powers of \( x_k \), so that \( \sum w_i x_{ik} = \sum U_i x_{ik} \) for \( j = 0, 1, ..., q \).

Breidt, Claeskens and Opsomer (2005) develop a penalized spline GREG estimator for a single \( x \)-variable; the assisiting model is \( m(x; \beta) = \beta_0 + \beta_1 x + ... + \beta_q x^q + \sum_{j=1}^{q} \beta_j (x - k_j)^2 \), where \((t)^q = t^q \) if \( t > 0 \) and 0 otherwise, \( q \) is the degree of the spline, and the \( k_j \) are suitably spaced knots, for example, uniformly spaced
sample quantiles of the $x_k$-values. After estimation of the $\beta$-parameters, they obtain the predictions $\hat{y}_k = m(x_k; \hat{\beta})$ needed for the general GREG formula (3.1). The authors point out that the resulting GREG estimator is calibrated for the parametric portion of the model, that is, $\sum_j w_j x_j = \sum_j x_j$ for $j = 0, 1, \ldots, q$, and also for the truncated polynomial terms in the model as long a they are left unpenalized.

We can summarize GREG estimation as follows. The linear GREG has practical advantages for large scale statistics production: It can be expressed as a linearly weighted sum of $y_k$-values with weights calibrated to $\sum_j x_j$, the weights are independent of the $y_k$-values and may be applied to all $y$-variables in the survey. It is sufficient to know a population auxiliary total $\sum x_j$, imported from a reliable source. Non-linear GREG may give a considerably reduced variance, as a result of the more refined models that can be considered when there is complete auxiliary information (known $x_j$ for all $k \in U$); near design unbiasedness is preserved. Certain non-linear GREG’s can be written as linearly weighted sums.

In academic exercises with artificially created populations and relationships, one can provoke situations where a nonlinear GREG has a large variance advantage over a linear GREG. Such experiments are important for illustration. However, to meet the daily production needs in national statistical agencies; “farfetched” nonlinear GREG’s seem to be of fairly remote interest at this point in time; the assisting models for GREG must meet requirements of robustness and practicality. The attraction of a minor reduction of the sampling variance is swept away by worries about other (non-sampling) errors and troubles in the daily production process.

The progression from linear to non-linear GREG creates opportunities and generates questions. What is the most appropriate formulation of the model expectation $\mu_k$? How sensitive are the results to the specification of the variance part of the assisting model? To what extent is computational efficiency an issue? Further research will respond more fully to these questions.

4. The calibration approach to estimation

4.1 Calibration under basic conditions

A crucial step in the GREG approach reviewed in the previous section is to produce predicted values $\hat{y}_k$ through the fit of an assisting model. By contrast, the calibration approach, as defined in Section 1.1, does not refer explicitly to any model. It emphasizes instead the information on which one can calibrate. A key element of “calibration thinking” is the linear weighting of the observed $y$-values, with weights made to confirm computable aggregates. This conceptual difference will sometimes lead to different estimators in the two approaches.

The calibration approach has considerable generality; it can deal with a variety of conditions: complex sampling designs, adjustments for nonresponse and frame errors. This section, however, focuses on the basic conditions in Section 2: single phase sampling and full response. The notation remains as in Section 2. The material available for estimating the population total $Y = \sum y_k$ is: (i) the study variable values $y_k$ observed for $k \in s$, (ii) the known design weights $d_k = 1/\pi_k$ for $k \in U$, and (iii) the known vector values $x_k$ for $k \in U$ (or an imported total $\sum x_k$).

These simple conditions prevail in Deville and Särndal (1992) and Deville, Särndal and Sautory (1993), papers which gave the approach a name and inspired further work. Even though the background is simple, calibration raises several issues, some of them computational, as reviewed in Section 5.

The objective in Sections 4.2 and 4.3 is to determine weights $w_k$ to satisfy the calibration equation $\sum w_k x_k = \sum x_k$, then use them to form the calibration estimator of $Y$ as $\hat{Y}_{CAL} = \sum w_k y_k$, which we can confront with the unbiased Horvitz-Thompson estimator by writing $\hat{Y}_{CAL} = \hat{Y}_{HT} = \sum (w_k - d_k) y_k$. It follows that the bias of $\hat{Y}_{CAL}$ is $E(\hat{Y}_{CAL}) - Y = E(\sum (w_k - d_k) y_k)$. Meeting the objective of near design unbiasedness requires $E(\sum (w_k - d_k) y_k) \approx 0$, whatever the $y$-variable. Evidently, the calibration should strive for small deviations $w_k - d_k$.

The objective “calibration for consistency with known population auxiliary totals” can be realized in many ways. We can construct many sets of weights calibrated to the known $\sum x_k$. This section examines this proliferation from two perspectives noted in the literature: the minimum distance method and the instrumental vector method. Yet another construction of a variety of calibrated weights is proposed in Demnati and Rao (2004).

4.2 The minimum distance method

In this method, the calibration sets out to modify the initial weights $d_k = 1/\pi_k$ into new weights $w_k$, determined to “be close to” $d_k$. To this end, consider the distance function $G_k(w, d)$, defined for every $w > 0$, such that $G_k(w, d) \geq 0$, $G_k(d, d) = 0$, differentiable with respect to $w$, strictly convex, with continuous derivative $g_k(w, d) = \partial G_k(w, d)/\partial w$ such that $g_k(d, d) = 0$. Usually the distance function is chosen such that $g_k(w, d) = g(w/d)/q_k$, where the $q_k$ are suitably chosen positive scale factors, $g()$ is a function of a single argument, continuous, strictly increasing, with $g(1) = 0$, $g'(1) = 1$. Let $F(u) = g^{-1}(u)$ be the inverse function of $g()$. Minimizing the total distance
The weights have an optimality property, because a duly specified objective function is minimized, but it is a “weak optimality” in the sense that there are many possible specifications of the distance function and the scale factors \( q_k \).

Much attention has focused on the distance function

\[
G_k(w_k, d_k) = (w_k - d_k)^2 / 2d_k q_k.
\]

It gives \( g_k(w_k, d_k) = (w_k^2 - d_k^2) / q_k^2; g(w/d) = w^2/d^2 - 1; F(u) = g^{-1}(u) = 1 + u. \)

The term “the linear case” is thus appropriate. The task is then to minimize the “chi-square distance” \( \sum (w_k - d_k)^2 / 2d_k q_k \), subject to \( \sum w_k x_k = \sum x_k \).

Equation (4.1) reads \( \sum d_k x_k (1 + q_k x_k) = \sum x_k \), which is easily solved for \( \lambda \). The resulting estimator of \( Y = \sum y_k \) is \( \hat{Y}_{\text{CAL}} = \sum w_k y_k \) with weights \( w_k = d_k g_k \) given by (3.3).

That is, \( \hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}} \) as given by (3.2), and the residuals that determine the asymptotic variance are \( E_k = y_k - x_k B_{\text{CAL}} \) as given in Section 3.2. Some negative weights \( w_k \) may occur.

The linear GREG estimator implies weights that happen to be calibrated (to \( \sum x_k \)), and the opposite side of the same coin says that the linear case for calibration (with chi-square distance) brings the linear GREG estimator.

The tendency in some articles and applications to intertwine GREG thinking and calibration thinking stems from this fact. Many successful applications of the use of auxiliary information stem, in any case, from this linearity on both sides of the coin. The Canadian Labour Force Survey is an example, and an interesting recent development for that survey is the use of composite estimators, with part of the information coming from the survey results in previous months, as described in Fuller and Rao (2001).

The calibration equation is satisfied for any choice of the positive scale factors \( q_k \) in (4.1). A simple choice is \( q_k = 1 \) for all \( k \). But it is not always the preferred choice. For example, if there is a single, always positive auxiliary variable, and \( x_k = x \), then many will intuitively expect \( \hat{Y}_{\text{CAL}} = \sum w_k y_k \) to deliver the usual ratio estimator \( \sum x_k (\sum d_k y_k)/(\sum d_k x_k) \), and it does, but by taking \( q_k = x_k \), not \( q_k = 1 \).

Another distance function of considerable interest is

\[
G_k(w_k, d_k) = \{w_k \log(w_k/d_k) - w_k + d_k\} / q_k.
\]

It leads to \( F(u) = g^{-1}(u) = \exp(u) \), “the exponential case”. Then (4.1) reads \( \sum d_k x_k \exp(q_k x_k) = \sum x_k \). Numeric methods are required to solve for \( \lambda \) to obtain the weights \( w_k = d_k \exp(q_k x_k) \). No negative weights \( w_k \) will occur.

Deville and Särndal (1992) show that a variety of distance functions satisfying mild conditions will generate asymptotically equivalent calibration estimators. Alternative distance functions are compared in Deville, Särndal and Säutoy (1993), Singh and Mohl (1996), Stukel, Hidiroglou and Särndal (1996). Some distance functions will guarantee weights falling within specified bounds, so as to rule out too large or too small (negative) weights. Changes in the distance function will often have minor effect only on the variance of the calibration estimator \( \hat{Y}_{\text{CAL}} = \sum w_k y_k \), even if the sample size is rather small. Questions about the existence of a solution to the calibration equation are discussed in Théberge (2000).

4.3 The instrumental vector method

An alternative to distance minimization is the instrumental vector method, considered in Deville (1998), Estevao and Särndal (2000, 2006) and Kott (2006). It can also generate many alternative sets of weights calibrated to the same information.

We can consider weights of the form \( w_k = d_k F(\lambda' z_k) \), where \( z_k \) is a vector with values defined for \( k \in s \) and sharing the dimension of the specified auxiliary vector \( x_k \), and the vector \( \lambda \) is determined from the calibration equation \( \sum w_k x_k = \sum x_k \). The function \( F() \) plays the same role as in the distance minimization method; several choices \( F() \) are of interest, for example, \( F(u) = 1 + u \) and \( F(u) = \exp(u) \).

Opting for the linear function \( F(u) = 1 + u \), we have \( w_k = d_k (1 + \lambda' z_k) \). It is an easy exercise to determine \( \lambda \) to satisfy the calibration equation \( \sum w_k x_k = \sum x_k \). The resulting calibration estimator is

\[
\hat{Y}_{\text{CAL}} = \sum w_k y_k; w_k = d_k (1 + \lambda' z_k), \lambda' = \left( \sum x_k - \sum d_k x_k \right) \left( \sum d_k z_k x_k \right)^{-1}.
\]

Whatever the choice of \( z_k \), the weights \( w_k = d_k (1 + \lambda' z_k) \) satisfy the calibration equation. The standard choice is \( z_k = x_k \). In particular, setting \( z_k = q_k x_k \), for specified \( q_k \), gives the weights (3.3).

Even “deliberately awkward choices” for \( z_k \) give surprisingly good results. For example, let \( x_k \) be a single continuous auxiliary variable, and \( z_k = c_k x_k^{-1} \). Suppose \( p = 3 \), and \( c_k = 1 \) for 4 elements only, chosen at random from \( n = 100 \) elements in a realized sample \( s \), and \( c_k = 0 \) for the remaining 96. The near-ubiquiasedness of \( \hat{Y}_{\text{CAL}} = \sum d_k (1 + \lambda' z_k) y_k \) is still present. Even with such a sparse \( z \)-vector, the increase in variance, relative to better choices of \( z_k \), may not be excessive.

When both sampling design and \( x \)-vector are fixed, Estevao and Särndal (2004) and Kott (2004) note that there is an asymptotically optimal \( z \)-vector given by

\[
z_k = z_{0k} = d_k^{-1} \sum_{i \in s} (d_i d_i - d_{ki}) x_i
\]
where $d_{kl}$ is the inverse of the second order inclusion probability $\pi_{kl} = P(k & \ell \in s)$, assumed strictly positive. The resulting calibration estimator, $\hat{Y}_{\text{CAL}} = \sum_k d_{kl}(1 + \lambda z_{kl})y_k$, is essentially the “randomization-optimal estimator” due originally to Montanari (1987) and discussed by many since then.

Andersson and Thorburn (2005) view the question from the opposite direction and ask: In the minimum distance method, can a distance function be specified such that its minimization will deliver the randomization-optimal estimator? They do find this distance; not entirely surprisingly, it is related to (but not identical to) the chi-square distance.

4.4 Does calibration need an explicitly stated model?

The calibration approach as presented in Sections 4.2 and 4.3 proceeds by simply computing the weights that reproduce the specified auxiliary totals. There is no explicit assisting model, unless one were to insist that picking certain variables for inclusion in the vector $x_k$ amounts to a serious modeling effort. Instead, the weights are justified primarily by their consistency with the stated controls. Early contributions reflect this attitude, from Deming (1943), and continuing with Alexander (1987), Zieschang (1990) and others. This begs the question: Is it nevertheless important to motivate such “model-free calibration” with an explicit model statement? It is true that statisticians are trained to think in terms of models, and they feel more or less compelled to always have a statistical procedure accompanied by a model statement. It may indeed have some pedagogical merit, also in explaining calibration, to state the associated relationship of $y$ to $x$, even if it is as simple as a standard linear model.

But will a stated model help the users and practitioners better understand the calibration approach? To most of them the approach is perfectly clear and transparent anyway. They need no other justification than the consistency with stated controls. Will a search for “the true model with the true variance structure” bring significantly better accuracy for the bulk of the many estimates produced in a large government survey? It is unlikely.

The next section deals with model-calibration. For that variety, proposed by Wu and Sitter (2001), modeling has indeed an explicit and prominent role. These authors call the linear calibration estimator, $\hat{Y}_{\text{CAL}} = \sum_k w_k y_k$ with weights $w_k$ given by (3.3), “a routine application without modeling”. The description is appropriate in that all that is necessary is to identify the $x$-variables with their known population totals.

4.5 Model-calibration

The idea of model-calibration is proposed in Wu and Sitter (2001) and pursued further in Wu (2003) and Montanari and Ranalli (2003, 2005). The motivating factor is that complete auxiliary information allows a more effective use of the $x_k$ known for every $k \in U$ than what is possible in model-free calibration, where a known total $\sum x_k$ is sufficient. The weights are required to be consistent with the computable population total of the predictions $\hat{y}_k$, derived via an appropriate model formulation. Thus the weight system may not be consistent with the known population total of each auxiliary variable, unless there is special provision to retain this property. Model-calibration still satisfies all three parts, (a) to (c), of the definition of calibration in Section 1.1; in particular, the estimators are nearly design unbiased.

Consider a non-linear assisting model of the type (3.4). We estimate the unknown parameter $\theta$ by $\hat{\theta}$, leading to fitted values $\hat{y}_k = \hat{\mu}_k = \mu(x_k, \hat{\theta})$, computed with the aid of the $x_k$ known for all $k \in U$. It follows that the population size $N$ is known and should be brought to play a significant role in the calibration. If minimum chi-square distance is used, we find the weights of the model-calibration estimator $\hat{Y}_{\text{MCAL}} = \sum_k w_k y_k$ by minimizing $\sum_k (w_k - d_k)^2/(2d_k q_k)$, for specified $q_k$, and $d_k = 1/\pi_k$, subject to the calibration equations

$$\sum_k w_k = N; \sum_k w_k \hat{y}_k = \sum_k \hat{y}_k. \quad (4.3)$$

For simplicity, let us take $q_k = 1$ for all $k$; we derive the calibrated weights, rearrange terms and find that the model-calibration estimator can be written as

$$\hat{Y}_{\text{MCAL}} = N\{\bar{y}_{x,d} - \bar{y}_{x,d} B_{x,d}\} \quad (4.4)$$

where $\bar{y}_{x,d} = \sum d_k y_k / \sum d_k$; $\bar{y}_{x,d} = \sum d_k \hat{y}_k / \sum d_k$, and

$$B_{x,d} = (\sum d_k (\hat{y}_k - \bar{y}_{x,d}) y_k) / (\sum d_k (\hat{y}_k - \bar{y}_{x,d})^2).$$

The regression implied by $B_{x,d}$ is one of observed $y$-values on predicted $y$-values. The idea of this regression would hardly occur to the modeler if his/her attempts to structure the relation between $y_k$ and $x_k$, but it proves effective in building the calibration estimator. Wu and Sitter (2001) present evidence that

$$(\hat{Y}_{\text{MCAL}} - Y) / N = (\sum d_k \hat{E}_k - \sum U \hat{E}_k) / N + O_p(n^{-1})$$

with $\hat{E}_k = y_k - \bar{y}_U - (\mu_k - \bar{\mu}_U) \hat{B}_U$, where $\hat{B}_U = (\sum (\mu_k - \bar{\mu}_U) y_k) / (\sum (\mu_k - \bar{\mu}_U)^2)$, and $\bar{\mu}_U = \sum \mu_k / N$. The coefficient $B_{x,d}$ may not be near one even in large samples. It expresses a regression of $y_k$ on its assisting model mean $\mu_k = (x_k, \hat{\theta})$. That is, $\hat{Y}_{\text{MCAL}}$ can be viewed as a regression estimator that uses the model expectation $\mu_k$ as the auxiliary variable, leaving $\hat{E}_k$ as the residuals that determine the asymptotic variance of $\hat{Y}_{\text{MCAL}}$. 

Statistics Canada, Catalogue No. 12-001-X
How does this asymptotic variance compare with that of the non-linear GREG construction (3.1) for the same non-linear assisting model and the same \( \hat{y}_k = \hat{\mu}_k \)? Formula (3.1) implies a slope equal to unity in the regression between \( y_k \) and \( \hat{y}_k = \hat{\mu}_k \); viewed in that light, \( \hat{Y}_{GREG} \) is a difference estimator rather than a regression estimator and hence less sensitive to the pattern in the data. The non-linear GREG \( \hat{Y}_{GREG} \) is in general less efficient than \( \hat{Y}_{MCAL} \). (It is of course possible to modify \( \hat{Y}_{GREG} \) to also account for the information contained in the known population size \( N \).

On the other hand, compared with the linear (model-free) calibration estimator \( \hat{Y}_{MCAL} = \sum w_k y_k \) with weights as in (3.3), the model-calibration estimator \( \hat{Y}_{MCAL} \) given by (4.4) may have a considerable variance advantage but implies a loss of the practical advantages of a consistency with the known values \( Y_k \) for all \( k \in U \). The non-linear GREG \( \hat{Y}_{GREG} \) is a normally distributed error. The model-calibration estimator \( \hat{Y}_{MCAL} \) has a variance advantage over the non-linear \( \hat{Y}_{GREG} \). They created a finite population \( U \) of size \( N = 2,000 \) with values \( (y_k, x_k), k = 1, \ldots, 2,000 \), such that \( \log(y_k) = 1 + x_k + e_k \); the 2,000 values \( x_k \) are realizations of the Gamma(1,1) random variable, and \( e_k \) is a normally distributed error. The auxiliary information consists of the population size \( N \) and the known values \( x_k \) for \( k = 1, \ldots, 2,000 \). Repeated simple random samples of size \( n = 100 \) were taken; the assisting model for both estimators was the log-linear

\[
E(y_k | x_k) = \mu_k \text{ with } \log(\mu_k) = \alpha + \beta x_k.
\]

This model was fit for each sample, using pseudo-maximum quasi-likelihood estimation. The fitted values \( \hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k) \) were used to form both \( \hat{Y}_{MCAL} \) and \( \hat{Y}_{GREG} \). The simulation variance was markedly lower for \( \hat{Y}_{MCAL} \). (The linear GREG (3.2), identical to the model-free calibration estimator, was also included in the Wu and Sitter study; not surprisingly, it is even less efficient than the non-linear GREG, under the strongly non-linear relationship imposed in their experiment.)

Montanari and Ranalli (2005) provide further evidence, for several artificially created populations, on the comparison between \( \hat{Y}_{MCAL} \) and the non-linear \( \hat{Y}_{GREG} \). Their assisting model, \( y_k = \mu_k + e_k \), is fitted via nonparametric regression (local polynomial smoothing), yielding predictions \( \hat{y}_k = \hat{\mu}_k \) for \( k \in U \). With this type of model fit, the predictions \( \hat{y}_k = \hat{\mu}_k \) are highly accurate. Not surprisingly, the model-calibration estimator \( \hat{Y}_{MCAL} \) achieves only marginal improvement over the non-linear \( \hat{Y}_{GREG} \).

We can summarize the calibration approach as follows: The estimator of \( Y = \sum y_k \) has the linearly weighted form \( \hat{Y} = \sum w_k y_k \). In linear (model-free) calibration, the calibration equation reads \( \sum w_k x_k = \sum x_k \); a known population auxiliary total \( \sum x_k \) is required, but complete auxiliary information (known \( x_k \) for all \( k \in U \)) is not required; the same weights can be applied to all \( y \)-values (multi-purpose weighting); the estimator is identical to the linear GREG estimator (but derived by different reasoning). In model-calibration, the assisting model mean \( \mu_k \) is non-linear in \( x_k \); complete auxiliary information is usually required; the calibration constraints include the equation \( \sum w_k \hat{y}_k = \sum y_k \); the weights \( w_k \) depend on the \( y_k \)-values, implying a loss of the multi-purpose property.

5. Computational aspects, extreme weights and outliers

The computation of calibrated weights raises important practical issues, discussed in a number of papers. All computation must proceed smoothly and routinely in the large scale statistics production of a national statistical agency. Undesirable (or unduly variable) weights should be avoided. Many practitioners support the reasonable requirement that all weights be positive (even greater than unity) and that very large weights should be avoided.

A few of the weights computed according to (3.2) can turn out to be quite large or negative. Huang and Fuller (1978) and Park and Fuller (2005) proposed methods to avoid undesirable weights.

In the distance minimization method, the distance function can be formulated so that negative weights are excluded, while still satisfying the given calibration equations. The software CALMAR (Deville, Särndal and Sautory 1993) allows several distance functions of this kind. An expanded version, CALMAR2, is described in LeGuennec and Sautory (2002). Other statistical agencies have developed their own software for weight computation. Among those are GES (Statistics Canada), CLAN97 (Statistics Sweden), Bascula 4.0 (Central Bureau of Statistics, The Netherlands), g-CALIB-S (Statistics Belgium). These strive, in different ways, to resolve the computational issues arising. The user needs to consult the users’ guide in each particular case to see exactly how the computational issues, including an avoidance of undesirable weights, are handled.

GES uses mathematical programming to minimize the chi-square distance, subject to the calibration constraints as well as to individual bounds on the weights, so that they will satisfy \( A_k \leq w_k \leq B_k \) for specified \( A_k, B_k \). Bascula 4.0 is
described in Nieuwenbroek and Boonstra (2002). The software g-CALIB-S, described in Vanderhoeft, Waetens and Museux (2001), Vanderhoeft (2001), uses generalized inverse (the Moore–Penrose) for the weight computation; consequently one need not be concerned about a possible redundancy in the auxiliary information.

In Bankier, Houle and Luc (1997) the objective is two-fold: to keep the computed weights within desirable bounds, and to drop some x-variables to remove near-linear dependencies. Isaki, Tsay and Fuller (2004) consider quadratic programming to obtain both household weights and person weights that lie within specified bounds.

An intervention with the weights (so as to get rid of undesirable weight values) raises the question how far one can deviate from the design weights \(d_k\) without compromising the desirable feature of nearly design unbiased estimation. An idea that has been tried is to modify the set of constraints so that tolerances are respected for the difference between the estimator for the auxiliary variables and the corresponding known population totals. Hence, Duchesne (1999) minimizes a “cost-ridged loss function”.

Outlying values in the auxiliary variables may be a cause of extreme weights. Calibration in the presence of outliers is discussed in Duchesne (1999). His technique of “robust calibration” may introduce a certain bias in the estimates; it may, however, be more than offset by a reduction in variance.

When the set of constraints is extended to make the weights restricted to specified intervals, a solution to the optimization problem is not guaranteed. The existence of a solution is considered in Théberge (2000), who also proposes methods for dealing with outliers.

### 6. Calibration estimation for more complex parameters

The calibration approach adapts itself to the estimation of more complex parameters than a population total. Examples are reviewed in this section. Single phase sampling and full response continue to be assumed; the notation remains as in Section 2. One example is the estimation of population quantiles (Section 6.1), another is the estimation of functions of totals (Section 6.2). Other examples in this category, not reviewed here, are Théberge (1999), for the estimation of bilinear parameters, and Tracy, Singh and Arnab (2003), for calibration with respect to second order moments.

#### 6.1 Calibration for estimation of quantiles

The median and other quantiles of the finite population are important descriptive measures, especially in economic surveys. To estimate quantiles, the finite population distribution function must first be estimated. Before calibration became popular, several papers considered the estimation of quantiles, with or without the use of auxiliary information. More recent articles have turned to the calibration approach for the same purpose, including Kovačević (1997), Wu and Sitter (2001), Ren (2002), Tillé (2002), Harms (2003), Harms and Duchesne (2006) and Rueda et al. (2007). As these papers illustrate, there is more than one way to implement the calibration approach. The non-smooth character of the finite population distribution function causes certain complexities; these are resolved by different authors in different ways.

Let \(\Delta(\cdot)\) denote the Heaviside function, defined for all real \(z\) so that \(\Delta(z) = 1\) if \(z \geq 0\) and \(\Delta(z) = 0\) if \(z < 0\). The unknown distribution function of the study variable \(y\) is

\[
F_y(t) = \frac{1}{N} \sum_{u} \Delta(t - y_u).
\]

The \(\alpha\)-quantile of the finite population is defined as \(Q_{y,\alpha} = \inf\{t | F_y(t) \geq \alpha\}\). The auxiliary variable \(x_j\), taking values \(x_{jk}\), has the distribution function \(F_{x_j}(t) = (1/N)\sum_{u} \Delta(t - x_{jk})\) with \(\alpha\)-quantile denoted \(Q_{x_j,\alpha}\), \(j = 1, 2, \ldots, J\). A natural estimator of \(F_y(t)\) based on the design weights \(d_k = 1/\pi_k\) is

\[
\hat{F}_y(t) = \frac{1}{\sum_{j} d_k} \sum_{j} d_k \Delta(t - y_{jk}).
\]

A calibration estimator \(\hat{F}_{y,\text{CAL}}(t)\) of takes the form

\[
\hat{F}_{y,\text{CAL}}(t) = \frac{1}{\sum_{j} w_k} \sum_{j} w_k \Delta(t - y_{jk})
\]

where the weights \(w_k\) are suitably calibrated to a specified auxiliary information; then from \(\hat{F}_{y,\text{CAL}}(t)\) we obtain the \(\alpha\)-quantile estimator as \(\hat{Q}_{y,\alpha} = \inf\{t | \hat{F}_{y,\text{CAL}}(t) \geq \alpha\}\). A formula analogous to (6.2) holds for \(\hat{F}_{x_j,\text{CAL}}(t)\).

Without explicit reference to any model, Harms and Duchesne (2006) specify the information available for calibration as a known population size, \(N\), and known population quantiles \(Q_{x_j,\alpha}\) for \(j = 1, 2, \ldots, J\). The complete auxiliary information, with values \(x_k = (x_{k1}, \ldots, x_{kJ})'\), known for \(k \in U\), is not required. (But in practice, the complete information would usually be necessary, because accurate quantiles of several x-variables are not likely to be importable from outside sources.) They determine the \(w_k\) to minimize the chi-square distance \(\sum_s (w_k - d_k)^2 / 2d_k q_{ks}\), for specified \(q_{ks}\), subject to the calibration equations

\[
\sum_j w_k = N; \hat{Q}_{x_j,\text{CAL},\alpha} = Q_{x_j,\alpha}, j = 1, 2, \ldots, J
\]

for suitably defined estimates \(\hat{Q}_{x_j,\text{CAL},\alpha}\). Now, if we were to specify \(\hat{Q}_{x_j,\text{CAL},\alpha} = \inf\{t | \hat{F}_{x_j,\text{CAL}}(t) \geq \alpha\}\), then it is in general not possible to find an exact solution of the calibration

Statistics Canada, Catalogue No. 12-001-X
problem as stated. Instead, Harms and Duchesne substitute smoothed estimators, called “interpolated distribution estimators”, of the distribution functions $F_j(t), j = 1, 2, ..., J$. They replace $\Delta(t)$ by a slightly modified function. Weights $w_k$ can now be obtained, as well as a corresponding estimated distribution function $\hat{F}_{j,\text{CAL}}(t)$; finally, $Q_k$ is estimated as $\hat{Q}_k = \hat{F}_{j,\text{CAL}}^{-1}(\alpha)$.

The resulting calibrated weights $w_k$ allow us to retrieve the known population quantiles of the auxiliary variables. This is reassuring; one would expect such weights to produce reasonable estimators for the quantiles of the study variable $y$. Moreover, in the case of a single scalar auxiliary variable $x$, the resulting calibration estimator delivers exact population quantiles for $y$ when the relationship between $y$ and $x$ is exactly linear, that is, when $y_k = \beta_k x_k$ for all $k \in U$. An idea involving smoothed distribution functions is also used in Tillé (2002).

The computationally simpler method of Rueda et al. (2007) is an application of model-calibration, in that they calibrate with respect to a population total of predicted $y$-values. Complete auxiliary information is required. Using the known $x_k$, compute first the linear predictions $\hat{y}_k = \hat{\beta}^T x_k$ for $k \in U$, with $\hat{\beta} = (\Sigma_j d_j q_j x_j x_j^T)^{-1} (\Sigma_j d_j q_j x_j y_j)$, where $d_j = 1/\pi_j$ and the $q_j$ are specified scale factors. The weights $w_k$ are obtained by minimizing the chi-square distance subject to calibration equations stated in terms of the predictions, so as to have consistency at $J$ arbitrarily chosen points $t_j, j = 1, ..., J$:

$$\frac{1}{N} \sum_k w_k \Delta(t_j - \hat{y}_k) = F_j(t_j), j = 1, ..., J$$

where $F_j(t_j)$ is the finite population distribution function of the predictions $\hat{y}_j$, evaluated at $t_j$. It is suggested that a fairly small number of arbitrarily selected points $t_j$ may suffice, say less than 10. Once the $w_k$ are determined, the $\alpha$-quantile estimate is obtained from $\hat{F}_{j,\text{CAL}}(t) = (1/N) \sum_k w_k \Delta(t - \hat{y}_k)$.

Quantile estimation provides a good illustration that the calibration approach can be carried out in more than one way when somewhat more complex parameters are being estimated. Both methods mentioned give nearly design unbiased estimation. The Harms and Duchesne (2006) weights are multi-purpose, independent of the $y$-variable; by contrast, the method of Rueda et al. (2007) requires a new set of weights for every new $y$-variable. Empirical evidence, by simulation, suggests that both methods compare favourably with the earlier quantile estimation methods, not based explicitly on calibration thinking (but on the same auxiliary information).

An extension of the calibration approach to the estimation of other complex parameters, such as the Gini coefficient, is sketched in Harms and Duchesne (2006).

6.2 Calibration for other complex parameters

Plikusas (2006), and Krapavickaitė and Plikusas (2005) examine calibration estimation of certain functions of population totals. (Their term “non-linear calibration” signifies “non-linear function of totals”; I do not use it here.) A simple example is the estimation of a ratio of two totals, $R = \sum_U y_{1k} / \sum_U y_{2k}$, where $y_{1k}$ and $y_{2k}$ are the values for element $k$ of the variables $y_1$ and $y_2$, respectively. (The distribution function (6.1) is in effect also of ratio type, with $y_{2k} = 1$, and $N = \sum_U 1$ as the denominator total.) These authors examine the calibration estimator $\hat{R}_{\text{CAL}} = \sum_k w_k y_{1k} / \sum_k w_k y_{2k}$. Its weights $w_k$, common to the numerator and the denominator, are determined by calibration to auxiliary information stated as follows: There is one auxiliary variable, $x_{1k}$, for $y_{1k}$, and another, $x_{2k}$, for $y_{2k}$; the ratio of totals $R_0 = \sum_U x_{1k} / \sum_U x_{2k}$ is a known value, by a complete enumeration at a previous occasion or from some other accurate source. The proposed calibration equation is $\sum_k w_k e_k = 0$, where $e_k = x_{ik} - R_0 x_{ik}$. Because $\sum_k e_k = 0$, the weights, by minimum chi-square distance, are

$$w_k = d_k \left(1 - \left(\sum_j d_j e_j^2\right)^{-1} e_k^2 \right).$$

These weights correctly retrieve the known ratio value $R_0$; setting $y_{1k} = x_{1k}$ and $y_{2k} = x_{2k}$ in $\hat{R}_{\text{CAL}}$, we have

$$\sum_k w_k x_{1k} - R_0 = \sum_k w_k x_{2k} = 0.$$

The empirical evidence in Plikusas (2006), and Krapavickaitė and Plikusas (2005) suggests that their calibration estimator compares favourably (lower variance, while maintaining near design unbiasedness) with other estimators, derived through other arguments than calibration, while relying on the same auxiliary information.

7. Calibration contrasted with other approaches

As many have noted, users view calibration as a simple and convincing way to incorporating auxiliary information, for simple parameters (Section 4), as for more complex parameters such as quantiles, ratios and others (Section 6). Simplicity and practicality are undeniable advantages, but aside from that, is calibration also “theoretically superior”? Are there instances where calibration can be shown to give more accurate and/or more satisfactory answers on questions of importance, when contrasted with other design-based approaches?

Section 4.5 gave one indication that calibration thinking may have an advantage over GREG thinking, in that model-calibration may give more precise estimates than the
non-linear GREG, for the same assisting model. The following Section 7.1 gives another example where calibration reasoning and GREG reasoning give diverging answers, with an advantage for the calibration method.

7.1 An example in domain estimation

The example in this section, from Estevao and Särndal (2004), shows, for a simple practical situation, a conflict between the results of GREG thinking and calibration thinking. The context is the estimation of the total for a larger group (a domain).

A probability sample is drawn from \( U = \{1, 2, ..., k, ..., N\} \); the known design weights are \( d_k = 1/\pi_k \). Let \( U_a \) be a domain; \( U_a \subset U \). The domain indicator is \( \delta_{ak} \) with \( \delta_{ak} = 1 \) if \( k \in U_a \) and \( \delta_{ak} = 0 \) if not. The target of estimation is the domain total \( \gamma_a = \sum_{y \in a} y \), which \( y \) is observed for each \( k \in s \). The Horvitz-Thompson estimator \( \hat{Y}_{HT} = \sum d_k y_{ak} \), although design unbiased, has low precision, especially if the domain is small; the use of auxiliary information will bring improvement. An auxiliary vector value \( x_k \) is specified for every \( k \in U \).

As is frequently the case in practice, the elements belonging to a domain of interest are not identified in the sampling frame. (If they are, some very powerful information is available from the start, but frequently real world conditions are not that favourable.) But suppose elements in a larger group \( U_C \) are identifiable; \( U_a \subset U_C \subset U \). For example, suppose \( y \) “income” and \( U_C \), a professional group specified for the persons listed in the frame, while \( U_a \) is a professional sub-group not identified in the frame. We can identify the sample subsets \( s_C = s \cap U_C \) and \( s_a = s \cap U_a \), and we can benefit from knowing the total \( \sum_{U} x_{Ck} \), estimable without bias by \( \sum d_k x_{ak} \), where \( x_{ak} = \delta_{ak} x_k \), and \( \delta_{ak} \) is the information group indicator: \( \delta_{ak} = 1 \) if \( k \in U_C \) and \( \delta_{ak} = 0 \) if not. The domain auxiliary total \( \sum d_k x_{ak} \) is unavailable, because \( U_a \) is not identified. Calibration to satisfy \( \sum w_k x_{ak} = \sum x_{ak} \) gives the nearly design unbiased estimator \( \hat{Y}_{CAL} = \sum w_k y_{ak} \), where \( w_k = \frac{d_k}{(1 + \lambda_k z_k)} \), with \( \lambda_k^* = \sum d_k x_{ak} - \sum d_k x_{ak} y_k \). The asymptotically optimal instrument for the given vector \( x_k \) is (see Section 4.3) \( z_k = z_{0ck} = d_k z_{0ck} \), \( d_k \).

By contrast, regression thinking for the same auxiliary information leads to \( \hat{Y}_{GREG} = \sum d_k y_{ak} + \sum d_k x_{ak} y_k \). The result of a weighted least squares fit at a suitable level, using all \( (\hat{Y}, x_k) \) or part \( (\hat{Y}, x_k) \) of the data points \( (y_k, x_k) \) available for \( k \in s \).

For example, the modeller may opt for a regression fit “extending beyond the domain” (so that \( \bar{s} \subset s \subset U \), in an attempt to borrow strength for \( \hat{Y}_{GREG} \) by letting it depend also on \( y \)-data from outside the domain. By contrast, \( \hat{Y}_{CAL} \) relies exclusively on \( y \)-data in the domain, and this is in effect better. Estevao and Särndal (2004) show that \( \hat{Y}_{CAL} \) with \( z_k = z_{0ck} \) has smaller (asymptotic) variance than \( \hat{Y}_{GREG} \), no matter how \( \bar{s} \) is chosen. Bringing in \( y \)-data from the outside does not help; calibration thinking and regression thinking do not agree.

8. Calibration estimation in the presence of composite information

As the preceding sections have shown, many papers choose to study estimation for direct, single phase sampling of elements, without any nonresponse. The information available for calibration is simple; the \( k \)-th element of the finite population \( U = \{1, 2, ..., k, ..., N\} \) has an associated auxiliary vector value \( x_k \).

However, in an important category of situations, the auxiliary information has composite structure. The complexity of the information increases with that of the sampling design. In designs with two or more phases, or in two or more stages, the information is typically composed of more than one component, reflecting the features of the design. The information is stated in terms of more than one auxiliary vector. For example, in two-stage sampling, some information may be available about the first stage sampling units (the clusters), other information about the second stage units (the elements).

Consequently, estimation by calibration (or by any alternative method) must take the composite structure of the information systematically into account. The total information has several pieces; the calibration can be done in more than one way. All relevant pieces should be taken into account, for best possible accuracy in the estimates. To accomplish this in a general or “optimal” way is not a trivial task. Calibration reasoning offers one way.

Regression reasoning, with a duly formulated assisting model, is an alternative way, but it will strike some users as more roundabout. Hence, surveys that allow composite auxiliary information bring further perspectives on the contrast between calibration thinking and GREG thinking.

Two-phase sampling and two-stage sampling are discussed in this section. Another example of composite information occurs for nonresponse bias adjustment, as discussed in Section 9.

Another aspect of composite information occurs when the objective is to combine information from several surveys. This, too, can be a way to add strength and improve accuracy of the estimates. It is a motivating factor (in addition to the user oriented motive to achieve consistency.
among surveys) in the previously mentioned repeated weighting methodology of the Dutch statistical agency. Combined auxiliary information for GREG estimation is considered in Merkouris (2004).

8.1 Composite information for two-phase sampling designs

Double sampling refers to designs involving two probability samples, \( s_1 \) and \( s_2 \), from the same population \( U = \{1, \ldots, k, \ldots, N\} \). Auxiliary data may be recorded for both \( U \) and \( s_1 \), the study variable values \( y_k \) are recorded only for \( k \in s_2 \) with an objective to estimate \( Y = \sum_U y_k \). Hidiroglou (2001) distinguishes several kinds of double sampling: In the nested case (traditional two phase sampling), the first phase sample \( s_1 \) is drawn from \( U \), the second phase sample \( s_2 \) is a sub-sample from \( s_1 \), so that \( U \supset s_1 \supset s_2 \). Two non-nested cases can be distinguished: In the first of these, \( s_1 \) is drawn from the frame \( U_1 \); \( s_2 \) from the frame \( U_2 \), where \( U_1 \) and \( U_2 \) cover the same population \( U \); the sampling units may be defined differently for the two frames. In the second non-nested case, \( s_1 \) and \( s_2 \) are drawn independently from \( U \).

To illustrate how composite information intervenes in the estimation, consider the nested case. The design weights are \( d_{ik} = 1/\pi_{ik} \) (\( s_1 \) sampled from \( U \)); \( d_{2k} = 1/\pi_{2k} (\pi_{2k} = \pi_{4k} \) in sub-sampling \( s_2 \) from \( s_1 \)). The combined design weight is \( d_k = d_{1k} d_{2k} \). The basic unbiased estimator \( \hat{Y} = \sum_k d_k y_k \) can be improved by a use of auxiliary information, specified here at two levels:

Population level: The vector value \( x_{ik} \) is known (given in the frame) for every \( k \in U \), thus known for every \( k \in s_1 \) and for every \( k \in s_2 \); \( \sum_U x_{ik} \) is a known population vector total;

First sample level: The vector value \( x_{2k} \) is known (observed) for every \( k \in s_1 \), and thereby known for every \( k \in s_2 \); the unknown total \( \sum_U x_{2k} \) is estimated without bias by \( \sum_k d_{ik} x_{2k} \).

How do we best take this composite information into account? In an adaptation of GREG thinking, Särndal and Swensson (1987) formulated two linear assisting models, the first one stated in terms of the \( x_{1k} \)-vector, the other one also brings in the \( x_{2k} \)-vector. The two models are fitted; the resulting predictions, of two kinds, are used to create an appropriate GREG estimator \( \hat{Y}_{GREG} \) of \( Y = \sum_U y_k \).

Dupont (1995) makes the important point that the given composite information invites “two different natural approaches”: Besides the GREG approach, there is a calibration approach that will deliver final weights \( w_k \) for a calibration estimator \( \hat{Y}_{CAL} = \sum_k w_k y_k \). It is of interest to compare the results of the two approaches. Both of them allow more than one option: In the GREG approach, there are alternative ways of formulating the linear assisting models with their respective variance structures. In the calibration approach, alternative formulations of the calibration equations are possible.

For example, a two-step calibration option is as follows: First find intermediate weights \( w_{ik} \) to satisfy \( \sum_k w_{ik} x_{ik} = \sum_U x_{ik} \); then use the weights \( w_{ik} \) in the second step to compute the final weights \( w_i \) to satisfy

\[
\sum_j w_i x_{ik} = \sum_k w_{ik} x_{ik} = \left( \sum_k w_{ik} x_{ik} \right) \left( \sum_k w_{ik} x_{2k} \right)
\]

where \( x_k \) is the combined auxiliary vector

\[
x_k = \left( x_{ik} \right)^T = \begin{pmatrix} x_{ik} \\ x_{2k} \end{pmatrix}.
\]

Alternatively, in a single step option, we determine the \( w_i \) directly to satisfy

\[
\sum_i w_i x_{ik} = \left( \sum_k w_{ik} x_{ik} \right) \left( \sum_i d_{1k} x_{2k} \right).
\]

The final weights \( w_i \) are in general not identical in the two options. Suppose that \( \sum_i x_{ik} \) is an imported \( x_i \)-total. At closer look, the two-step option requires more extensive information, because individually known values \( x_{ik} \) are required for \( k \in s_1 \), whereas it is sufficient in the single step option that they be available for \( k \in s_2 \). Some variance advantage may thus be expected from the two-step option, since \( \sum_k w_{ik} x_{2k} \) is often more accurate (as an estimator of \( \sum_i x_{ik} \)) than \( \sum_i d_{1k} x_{2k} \) in the single step procedure. Nevertheless, this anticipation is not always confirmed; the single step method can be better, as when \( x_1 \) and \( x_2 \) are weakly correlated.

Dupont (1995) and Hidiroglou and Särndal (1998) examine links that exist, not surprisingly, between the two approaches. A GREG estimator, derived from assisting models with specific variance structures, may be identical to calibration estimator, if the weights of the latter are calibrated in a certain way. In other cases, differences may be small.

The efficiency of different options depends in rather subtle ways on the pattern of correlation among \( y_k, x_{ik} \) and \( x_{2k} \). For example, to what extent do \( x_1 \) and \( x_2 \) complement each other, to what extent are they substitutes for one another? In the GREG approach, it is difficult or even futile to pinpoint a variance structure that truly captures a “reality” behind the data. The calibration approach is more direct. Some of its possibilities are explored in Estevao and Särndal (2002, 2006).

8.2 Composite information in two-stage sampling designs

The traditional two-stage sampling set-up (clusters sampled at stage one, elements sub-sampled within selected
clusters in stage two) has in common with two-phase sampling that the total information may have more than one component. There may exist (a) information at the cluster level (about the clusters); (b) information at the element level for all clusters; (c) information at the element level for the selected clusters only. Here again, authors are of two different orientations: some exploit the information via calibration thinking, others follow the GREG thinking route.

Estevao and Särndal (2006) develop calibration estimation for the traditional two-stage set-up, with composite information specified as follows: (i) for the cluster population \( U_1 \), there is a known total \( \Sigma_{i \in U_1} x_{(i)} \), where \( x_{(i)} \) is a vector associated with the cluster \( U_1 \), for \( i \in U_1 \); (ii) for the population of elements \( U = \bigcup_{i \in U_1} U_i \), there is a known total \( \Sigma_{i \in U} x_k \), where the vector value \( x_k \) is associated with the element \( k \in U \). Suppose both cluster statistics and element statistics are to be produced in the survey: Both the cluster population total \( Y_i = \sum_{i \in U_1} y_{(i)} \) and the element population total \( Y = \sum_{i \in U} y_k \) are to be estimated.

If no relation is imposed between cluster weights \( w_y \) and element weights \( w_x \), the former are calibrated to satisfy \( \sum_i w_y x_{(i)} = \sum_i x_{(i)} \), the latter to satisfy \( \sum_i w_x x_k = \sum_k x_k \). (Here, \( s_1 \) is the sample of clusters from \( U_1 \); \( s \) is the sample of elements from the cluster \( U_i \); \( s = \bigcup_{i \in s_1} s_i \) is the entire sample of elements.) Then \( \hat{Y}_{\text{CAL}} = \sum_i w_y y_{(i)} \) estimates the cluster population total \( Y_i \), and \( \hat{Y}_{\text{CAL}} = \sum_k w_x y_k \) estimates the element population total \( Y \).

Integrated weighting is often used in practice: A convenient relationship is imposed between the cluster weight \( w_y \) and the weights \( w_x \) for the elements within the selected cluster. Two forms of integrated weighting are discussed in Estevao and Särndal (2006).

One of these is to impose \( w_y = d_{kj} w_x \), where \( d_{kj} \) is the inverse of the probability of selecting element \( k \) within cluster \( i \). (For example, in single stage cluster sampling, when all elements \( k \) in a sampled cluster are selected, then \( d_{kj} = 1 \). Consequently \( w_y = w_x \) is imposed, and all elements in the cluster receive the same weight for computing element statistics, and that same weight is also used for computing cluster statistics.) The calibration equation \( \sum_i w_y x_{(i)} = \sum_k x_k \) then reads \( \sum_i w_y \sum_k d_{kj} x_k = \sum_k x_k \). The cluster weights \( w_y \) are now derived by minimizing \( \sum_k (w_y - d_{kj})^2 / d_{kj} \) subject to the calibration equation that takes both kinds of information into account:

\[
\begin{pmatrix}
\sum_k w_y x_{(i)} \\
\sum_k w_y \sum_k d_{kj} x_k
\end{pmatrix} =
\begin{pmatrix}
\sum_i x_{(i)} \\
\sum_k x_k
\end{pmatrix}.
\]  

(8.1)

Once the \( w_y \) are determined, the element weights \( w_x = d_{kj} w_y \) follow.

Another reasonable integrated weighting is to impose \( \sum_i w_y = N w_x \). For example, for single stage cluster sampling it implies that the cluster weight \( w_y \) is the average of the element weights \( w_x \) in that cluster.

Two-stage sampling is also the topic in Kim, Breidt and Opsomer (2005). They assume auxiliary information for clusters, via a single quantitative cluster variable \( x_{(i)} \), but none for elements. They develop and examine a GREG type estimator of the element total \( Y = \sum_i y_i \), \( \hat{Y} = \sum_{i \in s_1} \hat{\mu}_i + \sum_{i \in s_2} d_i (\hat{t}_i - \hat{\mu}) \), where \( \hat{t}_i \) is design unbiased for the cluster total \( t_i = \sum_{i \in U} y_k \), and \( \hat{\mu} \) is obtained by local polynomial regression fit. The estimator can be expressed on the linearly weighted form, with weights that turn out to be calibrated to the population totals of powers of the cluster variable \( x_{(i)} \).

### 8.3 Household weighting and person weighting

Some important social surveys set the objective to produce both household estimates and person estimates; some study variables are household (cluster) variables, others are person (element) variables. Consequently, a number of papers have addressed the situation with single stage cluster sampling \( (d_{kj} = 1) \) and the integrated weighting that gives all members of a selected household equal weight, a weight also used for producing household statistics. A general solution for this weighting problem, when both household information and person information are specified, is to obtain the household weights \( w_y \) calibrated as in equation (8.1) with \( d_{kj} = 1 \), then take \( w_k = w_y \).

Several articles focus on auxiliary vector values \( x_k \) attributed to persons. Alexander (1987) derives weights by minimizing chi-square distance, whereas Lemaître and Dufour (1987) and Nieuwenbrook (1993) derive the integrated weights via a GREG estimator. The Lemaître and Dufour technique proceeds by an indirect construction of an “equal shares auxiliary vector value” for all persons in a selected household; their result is derivable from the direct procedure in Section 8.2.

The household-weighting/person-weighting question is revisited in more recent papers. Some authors display calibration thinking, others GREG thinking. Isaki, Tsay and Fuller (2004) formulate the problem as one of calibrated weighting; their weights respect both household controls and person controls; no explicit assisting models are formulated. By contrast, Steel and Clark (2007) proceed by the GREG approach, with linear assisting model statements and accompanying variance structures.
9. Calibration for nonresponse adjustment

9.1 Traditional adjustment for nonresponse

The context of many good theory articles is the simple one of Section 2, which includes total absence of nonresponse. It is good theory for conditions that seldom or never occur. (As an author of papers in that stream, I am not without guilt.) Practically all surveys encounter nonresponse; although undesirable, it is a natural feature, and theory should incorporate it, from the outset, via a perspective of selection in two phases.

In many surveys, nonresponse rates are extremely high today, compared with what they were 40 years ago, that is, so low that one could essentially ignore the problem. Today, survey sampling theory needs more and more to address the damaging consequences of nonresponse. In particular, one pressing objective is to examine the bias and to try to reduce it as far as possible.

A probability sample $s$ is drawn from $U = \{1, 2, \ldots, k, \ldots, N\}$; the known design weight of element $k$ is $d_k = 1/\pi_k$. Nonresponse occurs, leaving a response set $r$, a subset of $s$, the study variable value $y_k$ is observed for $k \in r$ only. The unknown response probability of element $k$ is $Pr(k \in r|s) = \theta_k$. The unbiased estimator $\hat{Y} = \sum_k d_k \phi_k y_k$ is ruled out because $\phi_k = 1/\theta_k$ is unknown. To keep the idea of a linearly weighted sum, how do we then construct the weights? Unit nonresponse adjustment by weighting, based on “nonresponse modeling”, has a long history. Calibration offers a newer perspective.

In what we may call “the traditional procedure”, the probability design weights $d_k = 1/\pi_k$ are first adjusted for nonresponse and possibly for other imperfections such as outliers. The information used for this step is often a grouping of the sampled elements. Finally, if reliable population totals are accessible, the adjusted design weights are subjected to a calibration with respect to those totals.

The methodology of the Labour Force Survey of Canada, described in Statistics Canada (1998), exemplifies this widespread practice. A (modified) design weight is first computed for a given household, as the product of three factors. The product of the design weight and a nonresponse adjustment factor is called the sub-weight. The sub-weights are subjected in the final step to a calibration with respect to postcensal, highly accurate estimates of population by age group, sex and sub-provincial regions. The final weights meet the desirable objective of consistency, in regions within a province, with the postcensal estimates. The nonresponse bias remaining in the resulting estimates is unknown but believed to be modest.

The traditional procedure is embodied in the estimator type $\hat{Y} = \sum_k d_k (1/\hat{\theta}_k) y_k$, where $\theta_k$ has been estimated by $\hat{\theta}_k$ in a preliminary step, using response (propensity) modeling. What theory demands of the statistician is not an easy task, namely, to formulate “the true response model”, capable of providing accurate, non-biasing values $\hat{\theta}_k$. But the factors $1/\hat{\theta}_k$ are applied in many surveys in an uncritical and mechanical fashion, for example, by straight expansion within the strata already used for sample selection.

The traditional procedure is apparent for example in Ekholm and Laaksonen (1991) and in Rizzo, Kalton and Brick (1996).

Practitioners often act as if the resulting $\hat{Y} = \sum_k d_k (1/\hat{\theta}_k) y_k$ (following a more or less probing response modeling trying to get the $\hat{\theta}_k$) is essentially unbiased, something which it is not (unless the ideal model happens to be specified); one acts (for purposes of variance estimation, for example) as if $\pi_k \hat{\theta}_k$ is the true selection probability of element $k$ in a single step of selection, something which it is definitely not. This practice, with roots in the idyllic past, becomes more and more vulnerable as nonresponse rates continue their surreptitious climb.

An unavoidable bias results from the replacement of $\theta_k$ by $\hat{\theta}_k$. Decades ago, when the typical nonresponse was but a few per cent, it was defensible to ignore this bias, but with today’s galloping nonresponse rates, the practice becomes untenable. By first principles, unbiased estimation is the goal, not an estimation where the squared bias is a dominating (and unknown) contributor to the Mean Squared Error. We must resolve to limit the bias as much as possible. Calibration reasoning can help in constructing an auxiliary vector that meets this objective.

9.2 Calibration for nonresponse bias adjustment


Calibration reasoning starts by assessing the total available auxiliary information: information at the sample level (auxiliary variable values observed for respondents and for nonrespondents), information at the population level (known population auxiliary totals). The objective is to make the best of the two sources combined, so as to reduce both bias and variance. The design weights are modified, in one or two calibration steps, to make them reflect (i) the outcome of the response phase, (ii) the individual characteristics of the respondents, and (iii) the specified auxiliary information. The information can be summarized as follows:

Statistics Canada, Catalogue No. 12-001-X
Population level: The vector value \( x_k^* \) is known (specified in the frame) for every \( k \in U \), thus known for every \( k \in s \) and for every \( k \in r \); \( \sum_k x_k^* \) is a known population total.

Sample level: The vector value \( x_k^o \) is known (observed) for every \( k \in s \), and thereby known for every \( k \in r \); the unknown total \( \sum_k x_k^o \) is estimated without bias by \( \sum_k d_k x_k^o \).

Calibration on this composite information can be done in two steps (intermediate weights computed first, then used in the second step to produce final weights) or directly in one single step. Modest differences only are expected in bias and variance of the estimates. In the single step option, the combined auxiliary vector and the corresponding information are

\[
x_k = \begin{pmatrix} x_k^* \\ d_k x_k^o \end{pmatrix}, \quad X = \begin{pmatrix} \sum_k x_k^* \\ \sum_k d_k x_k^o \end{pmatrix}.
\]

Using an extension of the instrument vector method in Section 4.3, we seek calibrated weights \( w_k = d_k v_k \), where \( v_k = F(\lambda z_k) \) is the nonresponse adjustment factor, with a vector \( \lambda \) determined through the calibration equation \( \sum_k w_k x_k = X \); the resulting calibration estimator is \( \tilde{Y}_{CAL} = \sum_k w_k y_k \). It is enough to specify the instrument vector value \( z_k \) for respondents only; \( z_k \) is allowed to differ from \( x_k \). The function \( F(\cdot) \) has the same role as in Sections 4.2 and 4.3. Here, \( F(\lambda z_k) \) implicitly estimates the inverse response probability, \( \phi_k = 1/\theta_k \) as Deville (2002), Dupont (1995), Kott (2006) have noted. In the linear case, \( F(u) = 1 + u \), and \( v_k = 1 + \lambda^* z_k \), with \( \lambda^* = (\sum_k x_k^* - \sum_k d_k x_k^o)/(\sum_k d_k x_k^o) \).

The variables that make up the vector \( x_k^o \), although observed for sampled elements only, can be crucially important for the reduction of nonresponse bias (although less important than the \( x_k^* \) for the reduction of variance). For example, Beaumont (2005b) discusses data collection process variables can be used in building the \( x_k^o \) vector component.

9.3 Building the auxiliary vector

In some surveys, there are many potential auxiliary variables, as pointed out for example by Rizzo, Kalton and Brick (1996), and Särndal and Lundström (2005). For example, for surveys on households and individuals in Scandinavia, a supply of potential auxiliary variables can be derived from a matching of existing high quality administrative registers. A decision then has to be made which of these variables should be selected for inclusion in the auxiliary vector \( x_k \) to make it as effective as possible, for bias reduction in particular. As Rizzo, Kalton and Brick (1996) point out, “the choice of auxiliary variables is probably more important than the choice of the weighting methodology.”

Let us examine the bias, when \( z_k = x_k \). We need to compare alternative \( x_k \)-vectors in order to finally settle one likely to yield the smallest bias. (I assume \( x_k \) to be such that \( \mu^* x_k = 1 \) for all \( k \) and some constant vector \( \mu \), as is the case for many \( x_k \)-vectors, including the examples 1 to 5 at the beginning of Section 2.) A close approximation to the bias of \( \tilde{Y}_{CAL} \) is obtained by Taylor linearization as

\[
\text{nearbias}(\tilde{Y}_{CAL}) = (\sum_k x_k^o)'(B_{U:0} - B_U),
\]

which involves a difference between the weighted regression coefficient \( B_{U:0} = (\sum_k \theta_k x_k x_k^o)'/\sum_k \theta_k x_k y_k \) and the unweighted one, \( B_U = (\sum_k x_k x_k^o)'(\sum_k x_k y_k) \). Unless all \( \theta_k \) are equal, the bias caused by the difference in the two regression vectors may be substantial, even though \( x_k \) is a seemingly “good auxiliary vector”. This expression for nearbias is given in Särndal and Lundström (2005); related bias expressions, under different conditions, are found in Bethlehem (1988) and Fuller et al. (1994). We can write alternatively

\[
\text{nearbias}(\tilde{Y}_{CAL}) = (\sum_k x_k y_k)'M_k - 1)/y_k \text{ and } (\sum_k y_k)'/N,
\]

where \( M_k = (\sum_k x_k y_k)'/(\sum_k x_k y_k)' \). In comparing possible alternatives \( x_k \), a convenient benchmark is the “primary auxiliary vector”, \( x_k = 1 \) for all \( k \in U \), which gives \( \tilde{Y}_{CAL} = N \bar{y}_r = N \bar{y}_s/n_r \), where \( n_r \) is the number or respondents, with \( \text{nearbias}(N \bar{y}_r) = N(\bar{y}_{U:0} - \bar{y}_U) \), where \( \bar{y}_{U:0} = \sum_k \theta_k y_k'/\sum_k \theta_k \) and \( \bar{y}_U = \sum_k y_k/N \). The ratio

\[
\text{relbias}(\tilde{Y}_{CAL}) = \frac{\text{nearbias}(\tilde{Y}_{CAL})}{\text{nearbias}(N \bar{y}_r)} = \frac{\sum_k (\theta_k M_k - 1)y_k}{N(\bar{y}_{U:0} - \bar{y}_U)}
\]

measures how well a candidate vector \( x_k \) succeeds in controlling the bias, when compared with the primitive vector. We seek an \( x_k \) that will give a small bias. But \( \text{relbias}(\tilde{Y}_{CAL}) \) is not a computable bias indicator; it depends on unobserved \( y_k \) and on unobservable \( \theta_k \). We need a computable indicator that approximates \( \text{relbias}(\tilde{Y}_{CAL}) \) and depends on the \( x \)-vector but not on the \( y \)-variables, of which the survey may have many.

It is easy to see that \( \text{relbias}(\tilde{Y}_{CAL}) = 0 \) if an ideal (probably non-existent) \( x \)-vector could be constructed such that \( \phi_k = 1/\theta_k = \lambda x_k \) for all \( k \in U \) and some constant vector \( \lambda \).

For an \( x \)-vector that can actually be formed in the survey, we can at least obtain predictions of the \( \phi_k \). Determine \( \lambda \) to minimize \( \sum_k \theta_k (\phi_k - \lambda x_k)^2 \); we find \( \lambda = \hat{\lambda}_U \), where \( \hat{\lambda}_U = (\sum_k x_k y_k)'(\sum_k \theta_k x_k x_k^o)' \); the predicted value of \( \phi_k \) is \( \hat{\phi}_U = \hat{\lambda}_U x_k = M_k \). The (theta-weighted) first and second moment of the predictions \( \hat{\phi}_U = M_k \) are, respectively, \( \bar{M}_{U:0} = \sum_k \theta_k M_k / \sum_k \theta_k = N/\sum_k \theta_k = 1/\bar{\theta}_U \) and

\[
Q = \frac{1}{\sum_k \theta_k} \sum_k (M_k - \bar{M}_{U:0})^2 = (1/\bar{\theta}_U)(\bar{M}_U - 1)/\bar{\theta}_U.
\]
where $\hat{\Phi}_U = \sum d_k \phi_k / N$. Särndal and Lundström (2007) show that relbias($\hat{Y}_{CAL}$) and $Q$ have under certain conditions an approximately linear relationship,

$$\text{relbias}(\hat{Y}_{CAL}) \approx 1 - \frac{Q}{Q_0}$$

where $\Phi_U = \sum d_k \phi_k / N$ and $Q_0 = (1/\bar{\Theta}_U)(\Phi_U - 1/\bar{\Theta}_U)$ is the maximum value of $Q$. Thus if $Q$ were computable, it could serve as an indicator for comparing the different candidate $x_k$ -vectors. A computable analogue $\hat{Q}$ of $Q$ is instead obtained as the variance of the corresponding sample-based predictions $(\sum r_k x_k x_k')^{-1} x_k = m_k$, so that

$$\hat{Q} = \frac{1}{\sum r_k d_k} \sum r_k d_k (m_k - \bar{m}_{r,d})^2 = \bar{m}_{r,d} (\bar{m}_{r,d} - \bar{m}_{r,d})$$

where

$$\bar{m}_{r,d} = \frac{\sum r_k d_k m_k}{\sum r_k d_k} = \frac{\sum r_k d_k}{\sum r_k d_k} \bar{m}_{r,d} = \frac{\sum r_k d_k m_k}{\sum r_k d_k}.$$

We expect relbias to decrease in a roughly linear fashion as $\hat{Q}$ increases; thus, independently of the $y$ -variables, $\hat{Q}$ may be used as a tool for ranking different $x$ -vectors in regard to their capacity of reduce the bias.

We can use $\hat{Q}$ as a tool to select $x$ -variables for inclusion in the $x_k$ -vector, for example, by stepwise forward selection, so that variables are added to $x_k$ one at a time, the variable to enter in a given step being the one that gives the largest increment in $\hat{Q}$. The method is described in Särndal and Lundström (2007).

10. Calibration to account for other non-sampling error

Nonresponse errors are critical determinants of the quality of published statistics. When we examine how the calibration approach may intervene in the treatment other sources of non-sampling error than the nonresponse, the literature to date is not surprisingly much less extensive. However, several authors sketch a calibration reasoning to also incorporate frame errors, measurement errors, and outliers. Calibration has a potential to provide a more general theory for estimation in surveys, encompassing the various non-sampling errors.

As Deville (2004) points out (my translation from the French): “The concept of calibration lends itself to be applied with ease and efficiency to a great variety of problems in survey sampling. Its scope goes beyond that of regression estimation, an idea to which some seem to wish to reduce the calibration approach”. He provides a brief sketch of how a treatment of several of the nonresponse errors may be accomplished under the caption of calibration thinking.

Folsom and Singh (2000) present a weight calibration method using what they call the generalized exponential model (GEM). It deals with three aspects: extreme value treatment, nonresponse adjustment and calibration through post-stratification. The method provides built-in control for extreme values. Calibration to treat both coverage errors (under- or over-coverage of the frame) and nonresponse is discussed in Särndal and Lundström (2005) and Kott (2006). Skinner (1998) discusses uses of calibration in the presence of nonresponse and measurement error. He notes something which remains a challenge almost ten years later: “More research is needed to investigate the properties of calibration estimates in the presence of non-sampling errors”.

11. Conclusion

If I am to select one issue for a concluding reflection on the contents of this paper, let me focus on the concept of auxiliary information. It is the pivotal concept in the paper. If there is no auxiliary information, there is no calibration approach; there is nothing to calibrate on. I noted on the other hand that regression (GREG) estimation is an alternative but different thought process for putting auxiliary information to work in the estimation.

An objective in this paper has been to give a portrait of the two types of reasoning, and I made a point of noting how the thinking differs. I gave examples where essentially the same estimation objective is tackled by some authors through calibration reasoning, by others through GREG reasoning (or at least primarily by one or the other type). The respective estimators that they end up recommending may or may not agree. Whether or not the difference has significant consequence (for variance, for bias, for practical matters such as consistency and transparency) depends on the situation. This paper may help contributing an awareness of the separation existing between two thought processes that have guided researchers survey sampling.

References


Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. Survey Methodology, 21, 125-135.


Weighting for two-phase surveyed data

Seppo Laaksonen

Abstract

Missingness may occur in various forms. In this paper, we consider unit non-response, and hence make attempts for adjustments by appropriate weighting. Our empirical case concerns two-phase sampling so that first, a large sample survey was conducted using a fairly general questionnaire. At the end of this contact the interviewer asked whether the respondent was willing to participate in the second phase survey with a more detailed questionnaire concentrating on some themes of the first survey. This procedure leads to three missingness mechanisms. Our problem is how to weight the second survey respondents as correctly as possible so that the results from this survey are consistent with those obtained with the first phase survey. The paper first analyses missingness differences in these three steps using a human survey dataset, and then compares different weighting approaches. Our recommendation is that all available auxiliary data should have been used in the best way. This works well with a mixture of the two classic methods that first exploits response propensity weighting and then calibrates these weights to the known population distributions.

Key Words: Calibration; Internal vs. external auxiliary variables; Response propensity modelling method; Selective sub-sample.

1. Introduction

A standard survey is composed of one step or phase. This means that the potential survey units have first been chosen using a certain sampling design, and attempts have then been made to contact and interview these units as well as possible. However, varying amounts of non-response or other forms of missingness or data deficiencies will have occurred. Usually, addressing missingness leads to the application of post-survey adjustment methods of varying degrees of sophistication, which take advantage of available auxiliary variables. The auxiliary variables may be derived from various sources (see e.g., Laaksonen 1999 or an extended version in Laaksonen 2006b, and Lundström and Sämdal 2001), but for weighting purposes these are usually taken from registers, or other administrative sources or surveys. These kinds of auxiliary variables could be called external, if we want to distinguish them from internal auxiliary variables, that is, internal in the sense that the information is derived from the same survey or from its predecessor.

Internal auxiliary variables are especially used for imputations when the values for some items are missing. Such variables are also extensively used in panel surveys if a certain respondent has responded in one wave but not in another. In panel surveys, internal auxiliary information may be used both for weighting adjustments and for imputations.

This paper does not concern a standard survey as described above. It discusses two special characteristics:

(i) A survey consisting of two (or, in some sense, three) steps or phases. The first phase is like a standard survey, in which a certain number of units respond. For the second phase, we only keep in the frame the respondents who are willing to contribute to a more detailed survey. This leads first to having to distinguish such first-phase respondents who say they are willing to participate voluntarily, on one hand, and those of these respondents who actually answer the second questionnaire. This being the case, the latter subgroup will thus respond to both questionnaires.

(ii) When making attempts for post-survey adjustments, we will have the option to exploit both external and internal auxiliary variables in the second phase. The internal variables will thus be available from the first survey.

We are only considering weighting adjustments although some of our ideas could be used in imputations, too. The approach of the paper has not been much used in cross-sectional surveys although the same problem has been often met. For instance, it is typical that a face-to-face survey is conducted first and that at the end of it the interviewer will request the interviewee to respond to a self-administered additional questionnaire and, if the respondent is willing, the interviewer will hand out the questionnaire immediately for filling in, or submit it later to the volunteer. In both cases, the answers will be received by post or email. A recent example of this type is the European Social Survey (ESS) in which the supplementary questionnaire concerns especially the values of life (see www.Europeansocialsurvey.com).
Naturally, not all face-to-face respondents fill in this questionnaire.

The second-phase questions do not necessarily concern the same topic as those of the core questionnaire. Another usual strategy is to start with a broad questionnaire on a specific subject and then continue in the second phase with more detailed questions about the same subject. There can be some feedback from the first phase to the second questionnaire, and even to a sample that depends on the distribution of the key variables of the first survey (this is an example of adaptive sampling). This is often the case where there is not much experience in this type of a survey. Thus, the first survey also plays the role of a pilot survey. The so-called master samples are also close to this idea whereby the first phase survey (including a sample from administrative sources, like a micro census in some countries) may be conducted for constructing an appropriate sampling frame. In this case, the variables of the master sample are fairly limited, including usually only factual or background information.

In the case of master sampling, the objective is that the constructed sampling frame is a good representation of the target population. Hence, when going to a sample, this frame information can be well used as auxiliary data for editing, imputing and weighting of the real survey (second phase survey). Each real survey is thus a sub-sample of the master sample. We consider here a more complex case as illustrated by Table 1.

### Table 1 Illustration of initial sample and three follow-up datasets

<table>
<thead>
<tr>
<th>Sample with Auxiliary Variables</th>
<th>First Phase Respondents with Variables $Y_1$ (e.g., Health, Outdoor)</th>
<th>Volunteers for the Second Phase with Variables $Y_1$ Weights for 8,481 Units</th>
<th>Second Phase Respondents with Variables $Y_2$ (e.g., Skating, Boating) Weights for 5,480 Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$ (Gender, Age, Region, Season)</td>
<td>First Phase Respondents with Variables $Y_1$ (e.g., Health, Outdoor)</td>
<td>Volunteers for the Second Phase with Variables $Y_1$ Weights for 8,481 Units</td>
<td>Second Phase Respondents with Variables $Y_2$ (e.g., Skating, Boating) Weights for 5,480 Units</td>
</tr>
<tr>
<td>Design Weights for 12,554 Units without Overcoverage</td>
<td>First Phase Respondents with Variables $Y_1$ (e.g., Health, Outdoor)</td>
<td>Volunteers for the Second Phase with Variables $Y_1$ Weights for 8,481 Units</td>
<td>Second Phase Respondents with Variables $Y_2$ (e.g., Skating, Boating) Weights for 5,480 Units</td>
</tr>
</tbody>
</table>

First, there is a standard sampling procedure including some auxiliary variables $X$. A fairly high response rate was obtained (10,666 out of 12,554 units, about 85%) for the first survey. Some attrition occurred due to the fact that all respondents were not willing to participate in the second survey (we now have 68% of the initial sample left). Due to a rather high non-response rate in this second phase (in spite of voluntariness), our remaining sub-sample covers only 44 per cent of the initial sample. We now have the following three datasets available for the analysis:

- **A.** First-phase respondents with survey variables $Y_1$
- **B.** Second-phase respondents with survey variables $Y_2$
- **C.** Both first and second-phase respondents with survey variables $Y_1$ and $Y_2$.

Most users will receive both files, $A$ and $B$, and they can merge these together and obtain file $C$ if a common identifier is available. What does a user expect having received both data files? Naturally, that the estimate for the same parameter from both files is as identical as possible, that is, the results are consistent with each other. The user obviously understands that a certain parameter estimated from the smaller file $C$ is less accurate than that estimated from a larger file. In principle, it is possible to impute the missing values for variables $Y_2$, but we do not believe that it is possible to do this well, hence we approach this question by weighting. Our aim is to attempt to construct adjusted sampling weights for file $B$ so that the analysis over variables $Y_1$ and $Y_2$ is as adequate as possible.

Several strategies can be used for this kind of weighting. Useful general aspects have been presented by, among others, Kalton and Kasprzyk (1986), Little (1986), Särndal, Swensson and Wretman (1992), Fuller, Loughin and Baker (1994), Wu and Sitter (2001), and Lundström and Särndal (2001). If we assume that the missingness only depends on the sampling design, we can construct the weights for files $A$ and $B$ in the respective way. For example, if stratified random sampling has been applied, the same stratification would naturally be applied to both phases. In the case of post-stratification, an analogous strategy may be applied.

In our particular example survey, the sampling frame contained the respondents of the first rotation group of the 12 months of the Finnish LFS. Each monthly sample was drawn randomly. The LFS is based on simple random sampling, but due to nonresponse these weights were adjusted by a standard calibration technique (Deville, Särndal and Sautory 1993) using gender, age group (six categories) and region (five categories) as auxiliary variables. Later, we refer to these as design weights. The basic weights for the first-phase respondents were constructed correspondingly adding variable season (4×2 = 8 categories over two years) to the pattern of auxiliary variables. The ‘season’ variable is rarely used in Finnish surveys but was here considered to be necessary due to the ‘seasonal’ nature of the survey (see Section 2). The present paper does not consider this aspect in detail. The first three variables are usual in Finnish human surveys because such information can be validated well from updates in the population register. This being the case, we now presume
that we will have the best possible estimates for the first-phase respondents when using these adjusted calibrated weights. In any case, we have no further access to other possible useful information from external sources.

It is possible to use estimates obtained from the first-phase respondents as benchmarking information to calibrate the second-phase weights. This strategy is not difficult as such, but all variables $X$ and as many of the variables $Y_i$ as possible should be included in this process. Moreover, as precise aggregate or domain levels as possible should be found for this strategy, which is not an easy job and hence not attempted in this study. Nevertheless, it is not guaranteed that the estimates for other aggregates will be unbiased enough (results in Laaksonen 1999 give some evidence for this conclusion).

My proposed strategy is more straightforward and works without technical problems for all domains although it is not of course guaranteed that a possible bias will be substantially reduced for all domains. Thus, I have not tried any advanced calibration strategy, although this could be workable. I hope that other authors will show its possible benefits. A useful reference for them is the paper by Dupont (1995) that considers calibration of two-phase survey data, however without empirical evidence. It should be noted that I use calibration, but not a very advanced one (see Section 3).

The proposed methodology of this paper is largely based on a response propensity modelling that has been successfully used in other types of situations, see e.g., Ekholm and Laaksonen (1991), Laaksonen (1999), Duncan and Stasny (2001), and Laaksonen and Chambers (2006). The situation of Rizzo, Kalton and Brick (1996) is fairly close to the two-phase case of this paper, although it is concerned with a panel. Their methodology also has some similar features. In addition, a major difference concerns the response mechanism that here occurs in two steps, that is, both due to voluntariness and due to response in the second phase. We analyse these steps separately, too. Naturally, we compare the results obtained with alternative techniques. In Section 2, we briefly further describe our surveys and datasets, and Section 3 details the principles of our methods. Section 4 presents comparison results, and Section 5 draws a conclusion.

2. Principles of the datasets

The data are from a special survey conducted among Finnish citizens aged 15-74 years old (for more information, see Virtanen, Pouta, Sievänen and Laaksonen 2001). The topic concerns their leisure time activities, especially relating to outdoor hobbies and activities. First, a CAPI (computer-assisted personal interview) survey was conducted, covering various leisure time and hobby questions such as cycling, motorcycling, walking, jogging, sailing, swimming, hunting, fishing, nature photography, skiing, skating and riding. In all cases, the reference period was the previous year. Second, at the end of this survey the respondents were asked whether they would be willing to receive a special postal survey questionnaire in which more detailed questions would be asked about some of these activities. This survey would be conducted in a few weeks' time.

The survey was conducted over two years (1998-2000) in order to reduce response and interviewing burden. Another reason for this was that since these activities are seasonal to some extent, the responses were expected to be seasonally influenced (e.g., responses to questions about skiing might be different in summer and winter). The initial sample size after the removal of overcoverage (104 units of overcoverage) was 12,554 individuals.

We chose the following binary variables for our analysis presented in Section 4: Outdoors (person has performed regularly some outdoor activities in the nature), Health (is health good enough for outdoor activities?), Skiing, Fishing, Skating, Boating, Cycling and Jogging. In all cases, value = 1 means that a person has engaged in the activity during the preceding year, and value = 0, respectively, the opposite. All these variables were included in the first-phase questionnaire and we hence knew what to expect after the two consecutive phases. Note, that there are more complex variables in the data set but this simpler choice was made in order to interpret results more easily. The main conclusions are the same in the case of another choice.

In Section 4, we present two types of comparison, (i) those based on known information from the first phase, and (ii) those not based on known information. In both cases, we can fortunately check how well we have succeeded in the reduction of bias since we actually know the ‘true’ (or best possible) estimates. In addition, we analyse some variables only included in the second questionnaire, but we cannot say definitely how well each method works in these cases. We do not present the latest results in detail but these were observed to behave similarly to those of the second approach.

3. Response propensity modelling method and calibration

This study comprises three steps with the following weighting specifications:

First, well-designed calibrated sampling weights for the first phase respondents were created using the variables Region, Gender, Age group and Season (see also Section 1). Let us use symbol $w_i$ for these sampling weights for
respondent $k$. These weights thus are based on calibration, and also called ‘Basic’. Note that before this we have, before naturally, constructed design weights for the dataset, based on the stratified random sampling design. These are thus available for the non-respondents of the first phase, too.

Next, we model voluntariness/response probabilities using the most common link function (Logit is not necessarily the best link function as learned from (2006a), but this is what we use here.), that is, logit $= \log(\pi/(1-\pi))$, in which $\pi$ is the binary response probability (either $1 = \text{volunteer}$ $\text{vs.}$ $0 = \text{non-volunteer}$ or $1 = \text{respondent}$ $\text{vs.}$ $0 = \text{non-respondent}$) and the explanatory variables consist of variables $X$ and some variables $Y_1$ that have been considered to be ‘good.’ The model gives the predicted response probabilities $p_k$ that are now used in the following way when constructing each particular adjusted sampling weight:

$$w_k(\text{res}) = \frac{w_k}{p_k} g_c,$$

Here $g_c = a$ scaling factor which benchmarks the weights to certain known aggregates at level $c$. There are several alternatives for this benchmarking, but some type of calibration could be considered as a standard way. In this study we use post-stratified aggregates calibration could be considered as a standard way. In this Section we also use weights (vol) and the whole number of cells $= 6*2*5 = 60$) using the following straightforward technique

$$g_h = \frac{\sum_i w_k}{\sum_i w_k / p_k}.$$

As already pointed out, the quality is high in Finland for these kinds of post-stratified aggregates but not necessarily for any other aggregates.

Because we have two steps for the second phase, we have the following three model options that were all also used in Section 4:

(a) Model for voluntariness

(b) Model for the response given that the person volunteered (called also ‘TwoStep’).

(c) Model for the response as one step (called also ‘Direct’ and ‘OneStep’).

Note that steps (a) and (b) together give the weights for file B. This leads to the following formulation (vol = volunteer; $p_1 = \text{estimated response probability at step 1}$, $p_2 = \text{estimated response probability at step 2}$; respectively for the scaling factors $g_1$ and $g_2$):

Step 1: $w_k(\text{vol}) = \frac{w_k}{p_{2k}} g_{sh}$,

Step 2: $w_k(\text{res}) = \frac{w_k(\text{vol})}{p_{2k}} g_{2k}$.

The correct sampling weights had to be used in each modelling task. For models (a) and (c) this meant weights $w_k$, but for model (b) weights $w_k(\text{vol})$. In our comparison tests we also modelled the first-phase response and here we used design weights. The use of weights in the modelling gives more correct estimates, since we are trying to make our analysis representative for the target population. In some cases, the influence of weights is substantial, like in business surveys where weights often vary more than in standard household surveys. In this case, the results between weighted and unweighted models were not highly different, although the weighted ones should be used (this is well justified in Laaksonen and Chambers 2006 in which the influence of weights is substantial; Rizzo et al. 1996 also use weights). The empirical results (estimates, their standard errors and response probabilities) for the weighted solutions are presented later in Section 4.

In addition to our above key techniques, in the next Section we also use weights $w_k$ when providing our ‘best possible’ estimates for such parameters that are known, thus based on variables $Y_i$.

Moreover, we compare our specific results using post-stratified calibration only without modelling (we use also symbol ‘cal’ in the remaining sections). The latter could be interpreted as a very standard way of approaching the weighting problem (this was a house style prior the methodology proposed here). Note, however, that if a response model only includes the variables (and the same categories) used in post-stratification, the response propensity-based weights are exactly the same as obtained by post-stratification.

4. Empirical results

This Section presents results from different methods. First, we give results from different response models and then go on to compare different weights with each other, and at the end of the Section compare some parameter estimates based on different techniques.

4.1 Models for voluntariness and response

In order to fully understand the behaviour of missingness (due to non-response and voluntariness) in all three phases of the survey, we present in Table 2 results that are based on such auxiliary variables $X$ that are available in each step (in practice, we also used the variable ‘season’ but do not include its effects in this analysis since it is not a key issue in this paper).
Table 2

Logistic regressions using the three common explanatory variables in the three phases, that is, for the first phase respondents, for the voluntary respondents in the second phase and for the real respondents in the second phase. The estimates are odds ratios; their 95% confidence intervals are presented in parenthesis.

<table>
<thead>
<tr>
<th>Explanatory variables and other statistics</th>
<th>Model 1 First phase response</th>
<th>Model 2a Voluntariness</th>
<th>Model 3a Response for volunteers</th>
<th>Model 4a Second phase response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (ref. Female)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.71 (0.65, 0.78)</td>
<td>0.75 (0.68, 0.83)</td>
<td>0.77 (0.71, 0.83)</td>
<td></td>
</tr>
<tr>
<td>Age group (65+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 and under</td>
<td>1.00 (0.83, 1.21)</td>
<td>0.84 (0.65, 0.68)</td>
<td>0.78 (0.76, 0.93)</td>
<td>1.49 (1.28, 1.73)</td>
</tr>
<tr>
<td>25-34</td>
<td>0.96 (0.79, 1.15)</td>
<td>1.62 (0.90, 1.47)</td>
<td>1.35 (0.71, 1.02)</td>
<td>0.77 (1.49, 2.00)</td>
</tr>
<tr>
<td>35-44</td>
<td>0.85 (0.71, 1.02)</td>
<td>1.73 (0.76, 0.83)</td>
<td>1.05 (3.46, 4.81)</td>
<td>1.75 (1.40, 1.88)</td>
</tr>
<tr>
<td>45-54</td>
<td>0.89 (0.74, 1.07)</td>
<td>0.82 (4.08, 4.81)</td>
<td>1.20 (1.71, 3.69)</td>
<td>1.75 (1.58, 2.10)</td>
</tr>
<tr>
<td>55-64</td>
<td>1.18 (0.96, 1.45)</td>
<td>1.75 (2.05, 2.41)</td>
<td>1.82 (1.74, 4.17)</td>
<td>1.75 (1.49, 2.04)</td>
</tr>
<tr>
<td>Region (North)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South-East</td>
<td>0.55 (0.46, 0.66)</td>
<td>1.35 (2.12, 2.50)</td>
<td>1.35 (0.79, 1.16)</td>
<td></td>
</tr>
<tr>
<td>South-West</td>
<td>0.76 (0.64, 0.91)</td>
<td>1.35 (1.83, 2.14)</td>
<td>1.17 (0.86, 1.25)</td>
<td></td>
</tr>
<tr>
<td>Mid-West</td>
<td>1.14 (0.91, 1.42)</td>
<td>1.56 (2.14, 2.59)</td>
<td>1.18 (1.77, 2.59)</td>
<td></td>
</tr>
<tr>
<td>Mid-East</td>
<td>0.96 (0.78, 1.18)</td>
<td>1.19 (1.20, 1.44)</td>
<td>1.18 (1.01, 1.44)</td>
<td></td>
</tr>
<tr>
<td>Number of observations</td>
<td>12,554 (10,904)</td>
<td>10,666 (10,296)</td>
<td>8,481 (8,569)</td>
<td>10,666 (14,618)</td>
</tr>
<tr>
<td>-2 Log L</td>
<td>10,904 (10,296)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There are many interesting outcomes in these consecutive missingness behaviour models. The results of the first survey are fairly ordinary, for example, men respond more poorly than women in both phases. The response propensities are also lower in the South than in the rest of the country. The differences between age groups are somewhat surprising since the middle-age groups respond most poorly.

The voluntariness estimates are different. People in the Mid-East and North are the least willing to participate in the second survey, but the response premia given that a person is voluntary do not differ much. By age, it seems that younger people are more willing to participate but do not, nevertheless, respond very well. Older people, thus, seem in this sense to be more prepared to make a commitment than young people. However, we see clearly that the oldest ones will be under-represented without adjustments.

When considering the two first internal auxiliary variables (Table 3), it is observed that the people who are not relatively healthy (variable Health) and who do not actively pursue recreation in nature (Outdoor), are not willing to receive any new questionnaire, either. This is seen from the very high odds ratios. Interestingly, the respective odds ratio for the variable Health is close to the one for the volunteers. This, thus, means that a non-healthy person is not very likely to volunteer, but if he/she does, she/he responds as well as a healthy one. The tendency is similar with the variable Outdoor. It should be noted that the non-healthy and non-outdoor domains are not very large and although their roles in the response propensity modelling are important, their impacts on the final estimates are not very dramatic (Section 4.3).

When adding the other two internal auxiliary variables, that is, Skiing and Fishing, the same selectiveness continues although not as substantial. As a conclusion, we see clearly that the response mechanism of the second survey does not seem to be very non-informative. Consequently, it is expected that this leads to some effects on reweights and on survey estimates. These are considered in the next two subsections.
Table 3
Logistic regressions using some auxiliary variables from the first phase respondents in addition to those used in Table 2. The model numbers in this table and in Table 2 correspond to each other so that the response variable and the datasets are the same

<table>
<thead>
<tr>
<th>Explanatory variables and other statistics</th>
<th>Model 2b Voluntariness</th>
<th>Model 3b Response for volunteers</th>
<th>Model 4b Second phase response</th>
<th>Model 4c Second phase response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (ref. Female)</td>
<td>0.94</td>
<td>0.77</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>(0.85, 1.04)</td>
<td>(0.69, 0.85)</td>
<td>(0.75, 0.88)</td>
<td>(0.68, 0.83)</td>
</tr>
<tr>
<td>Age group (65+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 and under</td>
<td>4.92</td>
<td>0.52</td>
<td>1.30</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>(4.07, 5.97)</td>
<td>(0.41, 0.65)</td>
<td>(1.12, 1.52)</td>
<td>(0.41, 0.64)</td>
</tr>
<tr>
<td>25-34</td>
<td>3.83</td>
<td>0.65</td>
<td>1.46</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>(3.18, 4.60)</td>
<td>(0.52, 0.81)</td>
<td>(1.25, 1.70)</td>
<td>(0.52, 0.81)</td>
</tr>
<tr>
<td>35-44</td>
<td>3.26</td>
<td>0.64</td>
<td>1.37</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>(2.74, 3.88)</td>
<td>(0.51, 0.80)</td>
<td>(1.18, 1.58)</td>
<td>(0.52, 0.80)</td>
</tr>
<tr>
<td>45-54</td>
<td>2.59</td>
<td>0.85</td>
<td>1.56</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>(2.20, 3.05)</td>
<td>(0.68, 1.06)</td>
<td>(1.34, 1.81)</td>
<td>(0.69, 1.06)</td>
</tr>
<tr>
<td>55-64</td>
<td>1.73</td>
<td>1.18</td>
<td>1.55</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>(1.45, 2.05)</td>
<td>(0.89, 1.46)</td>
<td>(1.32, 1.81)</td>
<td>(0.90, 1.47)</td>
</tr>
<tr>
<td>Region (North)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South-East</td>
<td>2.15</td>
<td>0.96</td>
<td>1.34</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>(1.81, 2.55)</td>
<td>(0.79, 1.16)</td>
<td>(1.16, 1.54)</td>
<td>(0.79, 1.16)</td>
</tr>
<tr>
<td>South-West</td>
<td>1.92</td>
<td>1.04</td>
<td>1.36</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>(1.64, 2.26)</td>
<td>(0.86, 1.25)</td>
<td>(1.19, 1.56)</td>
<td>(0.86, 1.25)</td>
</tr>
<tr>
<td>Mid-West</td>
<td>2.09</td>
<td>1.15</td>
<td>1.52</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>(1.71, 2.54)</td>
<td>(0.93, 1.43)</td>
<td>(1.29, 1.78)</td>
<td>(0.93, 1.43)</td>
</tr>
<tr>
<td>Mid-East</td>
<td>1.17</td>
<td>1.15</td>
<td>1.18</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>(0.98, 1.41)</td>
<td>(0.91, 1.42)</td>
<td>(1.00, 1.38)</td>
<td>(0.92, 1.43)</td>
</tr>
<tr>
<td>Outdoor</td>
<td>3.04</td>
<td>1.24</td>
<td>1.93</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>(3.43, 2.71)</td>
<td>(1.43, 1.07)</td>
<td>(2.15, 1.74)</td>
<td>(1.97, 1.59)</td>
</tr>
<tr>
<td>Health</td>
<td>3.61</td>
<td>1.02</td>
<td>2.71</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td>(4.61, 2.82)</td>
<td>(1.58, 0.66)</td>
<td>(3.51, 2.09)</td>
<td>(3.24, 1.92)</td>
</tr>
<tr>
<td>Fishing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of observations</td>
<td>10,666</td>
<td>8,481</td>
<td>10,666</td>
<td>10,666</td>
</tr>
<tr>
<td>-2 Log L</td>
<td>9,721</td>
<td>8,560</td>
<td>14,342</td>
<td>14,244</td>
</tr>
</tbody>
</table>

4.2 Comparison of different weights

As already explained, we provided several weights. Table 4 gives a summary of these with descriptive statistics in order to explain the changes that occur after each adjustment operation. The design weights cannot be used in our comparisons since no data on variables \( Y \) are available for the initial sample. It is, however, illustrative to see that it has the lowest relative variation measured here with \( 1 + cv^2 \) in which \( cv \) is the coefficient of variation. This formula is also used as an approximation of the design effect (DEFF). Rizzo et al. (1996) also use this indicator when comparing their weights.

The changes are not dramatic in the first step, that is, from design weights to first-phase basic weights (except for the average that is related to decreasing counts), but in the following two steps the DEFFs are higher. We also see that the variation for both calibrated weights is lower than that for the respective response propensity-based weights. The distribution for each weight is skewed to the right, least for the design weights, naturally. It is somewhat surprising that the skewness is the highest for the volunteer weights. More details about the weight distributions and the differences between the weights are presented in Figures 1 to 3.

Figure 1 illustrates well how some weights have increased substantially due to the response propensity modelling (Model 2b). It is possible to look in detail to see which types of units are under the plots with high weight increase. For example, behind the separate left-side plots with \( RP \) weights higher than 700 are persons who are not healthy and do not engage much in outdoor activities but are, nevertheless, still in the volunteer data file. Similarly, we can find other interesting groups by using the results from the model estimations. However, the majority of the plots are in the same area and, consequently, less changes can be expected in the estimates than in the area with more substantial weight changes.
Table 4 Descriptive statistics for different sampling weights. \( RP \) = Response Propensity

<table>
<thead>
<tr>
<th>Weight</th>
<th>Phase</th>
<th>Unit size</th>
<th>Average</th>
<th>Skewness</th>
<th>( 1 + cv^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design Weight</td>
<td>Zero</td>
<td>12,658</td>
<td>308</td>
<td>0.94</td>
<td>1.30</td>
</tr>
<tr>
<td>(Calibrated) Basic Weight</td>
<td>First</td>
<td>10,666</td>
<td>365</td>
<td>1.30</td>
<td>1.39</td>
</tr>
<tr>
<td>Calibrated Weight</td>
<td>Volunteers</td>
<td>8,481</td>
<td>460</td>
<td>2.52</td>
<td>1.63</td>
</tr>
<tr>
<td>RP Weight, Model 2b</td>
<td>Volunteers</td>
<td>8,481</td>
<td>460</td>
<td>4.60</td>
<td>1.82</td>
</tr>
<tr>
<td>Calibrated Weight</td>
<td>Second</td>
<td>5,480</td>
<td>712</td>
<td>1.64</td>
<td>1.62</td>
</tr>
<tr>
<td>TwoStep RP Weight, Models 2b and 3b</td>
<td>Second</td>
<td>5,480</td>
<td>712</td>
<td>3.60</td>
<td>1.84</td>
</tr>
<tr>
<td>OneStep RP Weight, Model 4b</td>
<td>Second</td>
<td>5,480</td>
<td>712</td>
<td>2.56</td>
<td>1.80</td>
</tr>
</tbody>
</table>

The dispersion in the Figure 2 scatter is somewhat stronger than that in Figure 1 but the profile is similar. Consequently, interesting sub-groups can be found behind distinct plots.

Finally, Figure 3 compares the two alternative second phase weights with each other. This scatter differs considerably from the previous two, since the relationship is rather linear. The maximum values of the two-step weights are higher than those of the one-step weights, but the weights of many one-step weights are, however, clearly higher. For example, non-healthy people who, however, engage in outdoor activities receive relatively high one-step weights but there is no clear age effect. On the other hand, people with little outdoor activities in older age groups receive relatively high two-step weights but health does not relate to them. Nevertheless, it is not expected that there will be big differences in respective estimates although one of these two alternatives should have been introduced into use. If this choice were a simpler one, that is, one-step weighting, it would still be useful to analyse both steps and their response propensities separately in order to understand better the reasons for both types of missingness.

4.3 Comparison of parameter estimates

We have not been able to make complete simulation studies with different assumptions in order to analyse which type of method would be best in each particular case. Fortunately, we can get quite close to this by comparing the effects on the estimates from three different perspectives. First, we have prepared the response/voluntariness models by using both \( X \) and some \( Y_i \) variables. Consequently, we know the ‘best’ parameter estimates based on these \( Y_i \) values from the first survey. Second, we add auxiliary variables \( Y_i \) in the model but exclude some \( Y_i \) values from them. However, we know the ‘best’ values in these cases and thus can make exact comparisons. Third, we can compare some estimates that are not known in any way. In
this last case, we can only deduce which values might be the best.

We present our explicit results based on the variables described in Section 2. Note that we do not consider it important to present standard errors for each estimate because we are concentrating on the biases in these estimates. However, it is good to notice that the standard errors are around 0.2-0.4 percentage points for the first-phase data set and around 0.3-0.5 percentage points for the second-phase data set (lowest always in Health, second in Jogging, and highest in Outdoor and Skiing).

Figure 4 gives the results based on the weights without using any adjustment (that is often the case in practice, unfortunately). We see that the bias is substantial in most estimates, lowest in Jogging, which was not very actively practised when compared to Outdoor, for example. In general, most users are unhappy with such big biases that are statistically significant and highly significant except for Jogging in the second phase (e.g., the 95% confidence interval of the bias for Health is from 1.7 points to 2.3 points). Here, as in later results, the bias means overestimation so that while missingness increases the estimate becomes too high. The results without good adjustments will be too optimistic, that is, people seem to do ‘too much’ of all exercises. Note that the same tendency is obviously also in the first-phase estimates but we cannot justify this. There are surprising differences between those two estimates, sometimes the ‘volunteer’ data give a more biased result, sometimes it takes the second-phase respondents data. We do not interpret these in detail but naturally they reflect differences in missingness, and can be considered to be warnings for a user.

For comparison, we show again in Figure 5 the same unadjusted results for volunteers as in Figure 4 but we have added the corresponding estimates based on post-stratified calibration and response propensity modelling. This graph clearly shows that post-stratification gives some benefit compared to the unadjusted solution. However, the response propensity method is the best in each case, and extremely good for Health and Outdoor that have been used as auxiliary variables (Model 2b in Table 3).

Figures 6 and 7 concern the final-step estimates and are thus the most important. Figure 6 shows the same conclusion as Figure 5 in the sense that the response propensity technique is superior to post-stratified calibration although all differences are not statistically highly significant (especially Jogging). The difference between the one-step method and the two-step method is fairly small and the bias varies from one variable to the next. Hence, basing on this study, we cannot say which of these two specifications is better.
Figure 7 presents some comparisons when the two new variables have been added to the response propensity model. The results are quite predictable since this reduces the bias in these estimates and in all other estimates to some extent as well. The bias is still too large in Boating and Biking in the opinion of many users, I suppose. We can reduce this bias, naturally, by adding new auxiliary variables to the model. How far could we go in this? This has not been examined further in this study. On the other hand, we have worse tools for reducing bias in such variables that have been based on the second survey only. We tested several such estimates and observed some changes in corresponding estimates, being of the same level as in the cases of Boating and Biking in Figure 7. In this case, however, we cannot check the bias. We can only believe basing on our previous exercises that these results are less biased than those based on more poorly adjusted ones.

![Figure 7: Bias in estimates in percentage points for respondents after both steps based on post-stratification, and the two ‘One-step’ response propensity methods so that variables Skiing and Fishing have also been used as auxiliary variables (Model 4c in Table 3). These are compared to those based on Model 4b.](image)

5. Discussion

The problem discussed in this paper is common in surveys. There are many surveys which are conducted in more than one step, and some inconsistencies have occurred between these surveys due to missingness and other discrepancies. An internationally well-known example is the European Social Survey (ESS) that includes two questionnaires, a core one and a supplementary one. The number of respondents is naturally smaller for the latter than for the former. This leads to some selectiveness, for example, responding to the second questionnaire being positively associated to political activity. This is awkward from the user’s point of view because an estimate based on a larger dataset differs from that based on a smaller one, although both concern the same variable and time period.

Similarly to the ESS, this study concerns two-phase surveyed data. The response rate in the second survey was substantially lower than in the ESS. The effect from selectiveness is also higher. Using the response propensity models we predicted this selectiveness and exploited the results in weighting adjustments, and as the final step we calibrated the sums of the weights to correspond to certain known population aggregates. This strategy aims at making the most of all available auxiliary information, derived both from registers and other external sources, and also of the previous phase of the survey at the micro level.

In our example, the second phase of the survey comprised two different steps but only one data collection. The first step concerned willingness to participate voluntarily in the second phase of the survey, and the second step the actual survey participation of these volunteers, respectively. We examined both steps separately and found interesting information on their response mechanisms. Moreover, we used the results from this analysis for reweighting adjustments. For the sake of comparison, we looked at these both steps in one occasion and built a respective model, and continued the reweighting analogously. Finally, we compared the estimates. It was somewhat surprising that the two results differed quite little in our examples. This is, on the other hand, a good point, since it is easier to work with one step, and hence this could be introduced into use.

We thus propose a certain methodology for two-phase sampling weighting, but cannot say definitely which specification would be the best in each particular case. Our methodology is quite easy to exploit, but the advantages from it depend naturally largely on the availability of good external and internal auxiliary data. If no direct auxiliary variables are available, it will not be clear how good the adjusted estimates will be. Our examples show that these will be easily less biased than the initial ones. However, our recommended technique seems to be somewhat conservative so that all the best adjusted estimates in our analysis are slightly overestimated although not statistically significantly. This is an interesting question for future research that is still needed especially because this problem is becoming more common in the survey world. Another interesting topic for future research is how to make an optimal choice of auxiliary variables in the two-phase survey setting.

Acknowledgements

The author would like to thank the editor and the anonymous referees for their precise and helpful comments.
References


Weighting in rotating samples: The SILC survey in France

Pascal Ardilly and Pierre Lavallée

Abstract

The European Union’s Statistics on Income and Living Conditions (SILC) survey was introduced in 2004 as a replacement for the European Panel. It produces annual statistics on income distribution, poverty and social exclusion. First conducted in France in May 2004, it is a longitudinal survey of all individuals over the age of 15 in 16,000 dwellings selected from the master sample and the new-housing sample frame. All respondents are tracked over time, even when they move to a different dwelling. The survey also has to produce cross-sectional estimates of good quality.

To limit the response burden, the sample design recommended by Eurostat is a rotation scheme consisting of four panels that remain in the sample for four years, with one panel replaced each year. France, however, decided to increase the panel duration to nine years. The rotating sample design meets the survey’s longitudinal and cross-sectional requirements, but it presents some weighting challenges.

Following a review of the inference context of a longitudinal survey, the paper discusses the longitudinal and cross-sectional weighting, which are designed to produce approximately unbiased estimators.

Key Words: Longitudinal survey; Panel; Weight share method; Longitudinal weighting; Cross-sectional weighting.

1. Introduction

Statistics on Income and Living Conditions (SILC) is a European survey that produces data on the income and living conditions of persons living in regular households (persons living in communal households are excluded). It was introduced in 2004 as a replacement for the European Panel. While it is a European Union (EU) survey and therefore under Eurostat responsibility, it is conducted independently in each EU member state. Hence, the member states - France in this case - are free to adjust the sample design suggested by Eurostat to meet their national requirements. The data are also processed by the individual member states, as is usually the case for Eurostat surveys in the EU. This article deals only with the SILC survey conducted in France, but it may also be of interest to other EU member states.

SILC is a longitudinal survey conducted once a year in May. It focuses on individuals rather than households, and data are collected through personal interviews with every person in the sampled dwellings. SILC can be thought of as the European version of the Statistics Canada’s Survey of Labour and Income Dynamics (SLID) (see Lavallée 1995, and Lévesque and Franklin 2000).

The SILC sample is rotating: each year, it is formed by combining nine panel subsamples selected under identical steady-state conditions, partly from the master sample and partly from the new-housing sample frame. The master sample and the new-housing sample frame are two dwelling frames constructed from the French census of population and the information and automated data processing system for dwelling and office space (SITADEL) respectively (see Ardilly 2006).

Each incoming panel includes all individuals living in the selected dwellings. Surveying all members of the households living in the selected dwellings makes it possible to produce both individual-level and household-level estimates and helps keep collection costs down by maximizing the number of individuals reached in each contact. On the other hand, some of the estimates are narrower in scope, applying only to the population aged 16 and over on December 31 of the survey year.

Each year, one subsample is rotated out and replaced with another subsample. In the survey’s starting year, 2004, each subsample consisted of 1,780 dwellings (give or take a few units because of rounding). In the second and subsequent years (i.e., from 2005 on), the size of the year’s incoming subsample was 3,000 dwellings. Note that at the outset in 2004, the sample was 16,000 dwellings, divided into nine equal parts. One of those parts was surveyed only once (in 2004), another twice (2004 and 2005), a third three times (2004, 2005 and 2006), and so on. After the start-up phase, a given panel will be surveyed for nine consecutive years. During the start-up phase, which will end in 2012 with the departure of the ninth and last subsample from the 2004 selection, the subsamples will have been surveyed fewer than nine times.

The sampling procedure itself is the standard method of selecting units from the master sample and the new-dwelling sample frame (see Ardilly 2006). In this case, no
category of individuals is overrepresented. The survey has a uniform sampling fraction - ignoring rounding - except for vacant rural dwellings and dwellings that were secondary residences in the 1999 census and became principal residences by survey date, which are traditionally under-represented.

Under the collection process, each subsample is considered a true panel of individuals. Panel members who move are tracked, and their files are sent to the appropriate regional branch of INSEE. More details on SILC’s sample design are available in the November 17, 2003, issue of the Official Journal of the European Union and internal INSEE documents describing sampling practices in France.

Since SILC is a longitudinal survey with panels that overlap in time, weighting the sample presents a special problem. This paper provides a detailed picture of the two types of weighting used for SILC. We will begin by discussing some general principles related to SILC’s sample design. Then we will examine longitudinal weighting, followed by cross-sectional weighting.

Note that we will not consider the topics of non-response correction and estimate adjustment. Those issues are dealt with in the same way as they are generally for any other longitudinal survey, such as the SLID (see Lavallée 1995, and Lévesque and Franklin 2000).

2. General principles

2.1 Two approaches: The longitudinal view and the cross-sectional view

Each year, we have a sample of fully panelized individuals, eight ninths of whom were interviewed at least once in previous years (barring non-response).

Two types of parameters may be of interest: annual totals \( Y_t \) (or their satellites), and changes in totals \( \Delta_{t+1} \) between two years, consecutive or otherwise. For simplicity, we will confine ourselves to differences in totals between two consecutive years. When discussing changes, we have to be clear about the inference populations involved. We can look at the data in two different ways: either as populations that change over time - the cross-sectional approach - or as a fixed population - the longitudinal approach. If we let \( \Omega_t \) be the entire in-scope population in year \( t \), the annual total for year \( t \) is given by \( Y_t = \sum_{i \in \Omega_t} Y_i' \), where \( Y_i' \) is a variable of interest measured for individual \( i \). When we look at change, we may want to estimate the difference \( \Delta_{t+1} \) between the total \( Y_{t+1} \) at \( t+1 \) over \( \Omega_{t+1} \) and the total \( Y_t \) at \( t \) over \( \Omega_t \), that is, \( \Delta_{t+1} = Y_{t+1} - Y_t \). This is a cross-sectional view. Alternatively, we may want to estimate the difference \( \Delta_{t+1} \) between the totals for the units that are common to populations \( \Omega_{t+1} \) and \( \Omega_t \), where the size difference between the two populations is due to their incoming units (births) and outgoing units (deaths). This is a longitudinal view. Let \( \Omega_{t+1} = \Omega_t \cap \Omega_{t+1}^* \), the population that is common to \( t \) and \( t+1 \). Then \( \Delta_{t+1} \) is defined as \( \Delta_{t+1} = \sum_{i \in \Omega_{t+1}} (Y_i'' - Y_i') \).

The two approaches are illustrated in the diagrams below. The upper rectangle represents the entire population at time \( t \), and the lower one represents the entire population at time \( t+1 \). The “minus” side represents deaths in the broad sense (persons who have died, emigrated, moved to a communal household, and so on), and the “plus” side represents births in the broad sense (newborns, new immigrants, persons who have become part of the survey population by passing an age threshold, and so on). The grey portion represents the inference population on each date.

2.2 Surveys repeated over time and potential strategies

The goal, of course, is to produce both longitudinal estimates and cross-sectional estimates. There are essentially three possible strategies:

1. An “independent” sampling each year. In fact, because we have a master sample and a new-housing sample frame, the panels are selected from the same localities each year, and as a result, the subsamples are not truly independent. This solution can be improved for estimating changes.

2. A fully panelized sampling, i.e., initial selection of a sample that is surveyed each year. This scenario presents a response burden problem, since the SILC survey is to continue indefinitely. It is therefore unrealistic.

3. A rotational sample. This is the scenario that was chosen, because of its advantages in satisfying both longitudinal and cross-sectional goals.
The rotation strategy has four major advantages:

i. It reduces the sampling error associated with measuring change (in principle, as do panels, though it is theoretically less efficient than a “pure” panel).

ii. It has a smaller response burden than a “pure” panel. Under the circumstances, since France has a nine-year panel, this argument must be used with restraint. It is more persuasive in the scenario recommended by Eurostat, which consists of an annual survey for four consecutive years.

iii. It takes into account very “naturally” how the population changes over time. This point will become clearer when we look at the coverage of new populations.

iv. It reduces observation errors (as do panels).

On the other hand, the strategy also has at least three weaknesses:

i. Participants have to be tracked over time, which results in tracing costs and non-response due to moves.

ii. The length of the individual series is limited to nine years, which is substantial, though not as informative as a pure panel.

iii. The longitudinal/cross-sectional weighting method is not straightforward.

### Table: Sample Design Alternatives

<table>
<thead>
<tr>
<th>Sample Type</th>
<th>Cross-Sectional Approach</th>
<th>Longitudinal Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Independent” each year</td>
<td>CUSTOMARY</td>
<td>POSSIBLE but less efficient</td>
</tr>
<tr>
<td>Panel</td>
<td>IMPOSSIBLE without a top-up sample</td>
<td>CUSTOMARY</td>
</tr>
<tr>
<td>Rotational</td>
<td>POSSIBLE</td>
<td>POSSIBLE</td>
</tr>
</tbody>
</table>

The essential difficulty is to represent population $\Omega_t$ on date $t$ using eight panel subsamples selected on different dates and therefore from different populations. Intuitively it makes sense that a given individual would have a probability of selection on date $t$ that would depend on the number of panel subsamples for which he or she could be chosen. For this discussion, it is assumed that there is no non-response. This situation can be expressed formally by letting $a_{t,k} = a_{t,k}$ a panel subsample to be surveyed in year $t$ for the $k^{th}$ time, and $s_{t+1} = \bigcup_{k=1}^{9} a_{t,k}$.

Note that we can write $a_{t,k+1} = a_{t,k}$ ($\forall t, \forall k \neq 9$) since we are obliged to use each (non-outgoing) panel subsample in its entirety year after year. This is pictured below.

![Diagram](image)

The grey part represents $s_{t,k+1}$, which is the sample used in this longitudinal approach. It is from the individuals in $s_{t,k+1}$ that we obtain both $Y_t$ and $Y_{t+1}$, i.e., information about individual $i$ on dates $t$ and $t+1$ respectively.

Suppose we have an individual $i$ in $\Omega_t$ who is in-scope on date $t$. We denote as $L_i$ the number of years in $\{t-7, t-6, ..., t-1, t\}$ during which individual $i$ was in-scope and therefore had a chance of being selected as a member of an incoming panel. It is assumed here that each year, the sample frame covers the survey population exactly. We have $L_i \in \{1, 2, 3, ..., 8\}$. In addition, we denote as $K_i$ the set of $k$ indexes out of $1, 2, 3, ..., 8$ for which $i \in a_{t,k}$. These are the numbers of the panels of which individual $i$ is a member on date $t$. For all $i$ in $s_{t,k+1}$, $K_i$ will be construed as a set containing at least one element. Most of the time, $K_i$ will in fact have only one index, but in some cases, it will have two or even more indexes. That will be the case if $i$ is selected for a panel, he/she moves and his/her new dwelling is chosen for another panel in a subsequent year. Note that our scenario excludes the possibility of selecting a
given dwelling twice, since dwellings from the master sample and the new-housing sample frame are not supposed to be surveyed more than once. This is just a practical convention, however, as the theory can easily accommodate a system in which dwellings can be selected multiple times.

If \( i \in a_{t,k} \), let \( W_i(t,k) \) be his/her “raw” sampling weight. In fact, it is the sampling weight of the dwelling in which \( i \) was living at the time he/she was chosen as a panel member, i.e., at the time of the annual selection from \( \Omega_{t-1} \). This weighting system allows direct inference from subsample \( a_{t,k} \) to the entire population \( \Omega_{t-1} \). In particular, \( \sum_{a_{t,k}} W_i(t,k) \) provides an unbiased estimate of the total number of in-scope individuals who are members of population \( \Omega_{t-1} \). For SILC in France, that total is roughly 60 million. The longitudinal weight assigned to each individual \( i \) in \( s_{t,i+1} \) will therefore be as follows:

\[
W_{i,t+1} = \frac{1}{L_i} \sum_{k \in a_i} W_i(t,k).
\]

(1)

This equation is derived from the application of the weight share method (see Lavallée 1995, and Lavallée 2002) in which the initial population (the population of sampling units) is defined as the union of the populations \( \Omega_{-7}, \ldots, \Omega_{-1}, \Omega \) and the final population (the population of observation units) as \( \Omega \). This is illustrated in the diagram below; for greater clarity, only three of the initial subpopulations are shown. Clearly, the number of links is equal to \( L_i \) (in this case, \( i \) has exactly eight links, while \( j \) must have fewer than eight because it does not appear in the oldest sample frames). In practice, it is realistic to proceed as if \( \Omega_{t-7} \subset \Omega_{t-6} \subset \ldots \subset \Omega_{t-1} \subset \Omega \). We can work with nested populations, since all individuals who leave the survey population in the time before \( t \) will not be part of \( s_{t,i+1} \).

Equation (1) is the most general formula for the “raw” longitudinal weight. It can then be simplified for specific situations. For example, if we ignore the cases in which a panel member can be selected more than once, we have

\[
W_{i,t+1} = \frac{W_i}{L_i}.
\]

(2)

where \( W_i \) is the weight of \( i \) relative to the one panel subsample of which he/she was a member on date \( t \). In France’s case, because of the sample sizes involved, it seems quite appropriate to use that equation. If we assume that we are in the ideal position - though that seems simplistic in our circumstances - of having a population that does not change over time, we will have \( L_i = 8 \) for all \( i \). The population changes a great deal in nine years, but with shorter panel lives, the ideal case might be an acceptable approximation. Moreover, if the panels are selected with equal probabilities, \( W_i \) will be equal to a constant \( W \) and we will have

\[
W_{i,t+1} = \frac{W}{8}.
\]

(3)

Such a scenario is highly improbable in France’s case. First, up to 2012, the sample will contain subsamples with very different raw weights. Second, the sampling process is likely to focus on generating a predetermined number of dwellings (as the total number of dwellings increases), and not a constant sampling fraction.

Note that equation (3) is intuitive. Ultimately, everything proceeds “as if” any individual in the longitudinal sample \( s_{t,i+1} \) had a selection probability eight times the selection probability of each panel subsample \( s_{i,t+1} \) that is part of.

The foregoing applies to the survey in its steady state and must be adapted slightly during the start-up phase, i.e., until 2012. The first longitudinal operation is performed on the combined 2004-2005 data, to estimate the changes between 2004 and 2005 with the 2004 reference population (from which the “deaths” are removed in 2005). In this case, we need only to divide all the weights \( W_i \) of the eight subsamples \( a_{2004,1} \) to \( a_{2004,8} \) by 8; in other words, \( L_i = 8 \) for all \( i \). In 2006, when we look at the 2005-2006 changes, the denominator \( L_i \) may take only two values. In the first scenario, panel member \( i \) was in the sample frame used in 2004 (and hence could have been selected in 2004) and so \( L_i = 8 \). This is due to the fact that everything proceeds as if, in 2004, the seven selection processes for panels \( a_{2005,2} \) to \( a_{2005,8} \) had been carried out under exactly the same conditions. In the second scenario, individual \( i \) was not in the 2004 sample frame - but is in the 2005 frame and is necessarily in \( a_{2005,1} \) - and \( L_i = 1 \). For the 2006-2007 changes, \( L_i \) can be equal to 1, 2 or 8, and so on. We will not
have the set of all possible values of \( L \) in \{1, 2, 3, ..., 8\} until we measure the 2011-2012 changes.

Once we reach this point in the longitudinal weighting process, we can calculate the longitudinal weights \( W_{t+1} \) and then derive the estimator of the difference \( \Delta_{t+1} \) using

\[
\hat{\Delta}_{t+1} = \sum_{i \in \Omega} W_{t+1}(Y_{t+1} - Y_t).
\]

Logically, the weights \( W_{t+1} \) are used only to estimate change. They are of no value for point estimates because the inference population has little meaning on a particular date. Note that up to this point, the \( W_{t+1} \) have not been corrected for non-response or adjusted in any other way. In practice, equation (4) will be subject to adjustments in the case of the SILC survey.

Estimation of the difference \( \Delta_{t+1} = Y_{t+1} - Y_t \) is a cross-sectional matter and therefore involves the weighting process described in the next section.

4. Cross-sectional weighting

The aim is to make an inference about the total in-scope population \( \Omega_t \) on the current date \( t \). The essential difficulty lies in the fact that in theory, a given (sampled) subsample provides adequate coverage of the population only in the year in which it was selected. After that year, the panel subsample no longer represents the new population of “births”, the units that become in-scope. That is the case for newborns, immigrants, individuals who reach specific age thresholds, homeless people who start living in a regular dwelling, people who leave communal dwellings, and so on. While in practice we might consider this coverage defect acceptable for a period of time, it very quickly becomes a serious problem (that is true each year for most panel subsamples), and a top-up sample must be obtained in some fashion. It is worth noting that the problem of population change over time is highly dissymmetrical, since the subpopulation that disappears from year to year (the “deaths”) presents no particular difficulties for weighting.

In the SILC survey, the top-up sample is obtained as follows. We survey all individuals in the household of each panel member interviewed in the longitudinal tracking process. Thus, every household surveyed in the cross-sectional process is made up of two types of people: panel members and cohabitants (people who are surveyed but are not panel members). This method covers a large portion of the “births” (in the broad sense) in the population. However, it does not cover households consisting entirely of “births”, such as households of new immigrants. “Birth” status is usually determined by asking the birth date of newborns and the landing date of immigrants. Moreover, in practice, the weakness in births coverage is generally regarded as very minor because it is partially corrected with adjustments.

The main technique used to produce cross-sectional weights is the weight share method (Lavallée 2002). As noted previously, in year \( t \) we have nine panel subsamples \( a_{t,k} \) (1 \( \leq k \leq 9 \)). We will describe two different ways of using the weight share method. The information that must be collected in the questionnaire is the same for both methods.

4.1 Method 1

The more rigorous approach involves linking all nine subsamples \( a_{t,k} \) to the cross-sectional sample for year \( t \), which we will denote \( \bar{a}_t \) (Merkouris 2001). In other words, the sample \( \bar{a}_t \) is the same as \( s_{t} = \bigcup_{k=1}^{9} a_{t,k} \). First, we must determine the links associated with this approach. When a panel member in one of the nine subsamples \( a_{t,k} \) is selected, he/she points to himself/herself as a member of the cross-sectional sample at \( t \) (similar to what is shown in the diagram in 3.1). Under these conditions, when the survey is in steady-state mode, the cross-sectional weight \( W^i(t) \) of an individual \( i \) in \( \bar{a}_t \) is calculated as shown below. The household of which \( i \) is a member is denoted \( m \). We have

\[
W^i(t) = \frac{\sum_{k=1}^{9} \sum_{j \in m}^1 W_j(t, k)}{\sum_{k=1}^{9} \sum_{j \in \Omega_{1-t+1}}^1 1}
\]

where \( W_j(t, k) \) is the sampling weight from sample \( a_{t,k} \).

This expression shows that all members of the same household ultimately have the same weight. In the numerator, we have the sum of all the “raw” weights (the sampling weights) of the household’s panel members. It is understood that a panel member generally appears in only one subsample, but that there may be cases in which a panel member is selected two or more times over a period of nine consecutive years (usually because he/she has moved). Note that dwellings selected from the master sample and the new-housing sample frame are not supposed to be selected again and therefore, in the case of SILC, the probability that an individual who has not moved will appear in two different panels is zero.

As in the longitudinal case (see 3.1), weighting can be carried out only if the data management system is capable of linking each panel member in \( \bar{a}_t \) to all panel samples \( a_{t,k} \) in which he/she is included. In the denominator, for each of the nine years \( t-8 \) to \( t \) considered, we count the household members (both panel members and cohabitants) who are in the sample frame from which the incoming panel subsample
for the year in question is selected. This calculation clearly requires the information provided by the questionnaire.

There are two advantages to this approach: it is completely general, and it produces unbiased cross-sectional weights directly because every cross-sectional household is necessarily linked to one of the nine subsamples involved. The fact that there is an incoming subsample each year ensures the completeness of the cross-sectional population \( \Omega_i \); that is, in more technical terms, it ensures that there is at least one link for each household considered at \( t \). This is a useful property of rotational sampling, as discussed in section 2.2. On the other hand, the weighting formula has a disadvantage, which is its (relative) complexity both in theoretical terms and for computer programming purposes.

In the start-up phase (up to and including 2011), the formula must be adjusted. The numerator remains the same, but the denominator covers all individuals who could be sampled in 2004 (the survey’s first year) and subsequent years. In 2004, weighting is trivial since there is no weight share, but in 2005, we have

\[
W^{t(1)}_i = \sum_{j=1}^{9} \sum_{m_{j=1}^{m_{a=1}}} W_j(t, k) \left( \sum_{j \in \Omega_{2005}} 1 \right) + 8 \left( \sum_{j \in \Omega_{2004}} 1 \right).
\]

In 2006, the formula will be

\[
W^{t(1)}_i = \sum_{j=1}^{9} \sum_{m_{j=1}^{m_{a=1}}} W_j(t, k) \left( \sum_{j \in \Omega_{2005}} 1 \right) + 7 \left( \sum_{j \in \Omega_{2004}} 1 \right).
\]

### 4.2 Method 2

It is possible to take an alternative approach to cross-sectional weighting one that leads to a “slightly” simpler equation and is easier to program, but one that presents a difficulty that was not present in the previous method and may make the final weights somewhat less precise. The idea is to use one subsample at a time rather than all of them at once. We take one of the nine subsamples \( a_{t, k} \) and the sample of households to which it leads. We then apply the weight share, which when the survey is in steady-state mode yields an individual weight equal to

\[
\tilde{W}_j(t, k) = \frac{\sum_{j \in \Omega_{a_{t, k}}} W_j(t, k)}{\sum_{j \in \Omega_{a_{t, k}}} 1}.
\]

for any individual \( i \) in household \( m \). It is very easy to verify that if \( k = 1 \) (which is the case for the incoming subsample), \( \tilde{W}_j(t, 1) \) is the sampling weight of household \( m \).

The problem with this approach lies in the existence \( (a \ priori) \) on date \( t \) of individuals who cannot be surveyed because they belong to households that cannot be “reached” through the sampling \( a_{t, k} \) (as long as \( k \geq 2 \)), i.e., individuals whose probability of being surveyed at \( t \) is zero. This impediment did not exist in the previous method because taking all the subsamples into account at once ensured that on date \( t \), every household had a non-zero probability of being selected, at least through \( a_{t, 1} \). This illustrates once again one of the key advantages of rotational sampling, which is that it covers the entire population each year. In our approach, it is clear that if we consider \( a_{t, k} \) for \( k \geq 2 \), the population of households consisting exclusively of “immigrants” (in the broad sense) between \( t - k + 1 \) and \( t \) is not covered. To formalize the situation and produce the final cross-sectional weight, we will use \( \Omega^{\text{immig}}_{a_{t}} \) to denote the population of “immigrants” (in the broad sense) on date \( t \) in households that consist only of immigrants who can be sampled after year \( a_{t} \), with \( t - 8 \leq a_{t} \leq t - 1 \). To be more precise, we should say “who can be sampled on or after a date that is strictly subsequent to the collection date in year \( a_{t} \).”

On date \( t \), the entire population \( \Omega_{a_{t}} \) is partitioned into nine components: the eight subpopulations \( \Omega^{\text{immig}}_{a_{t}} \), with \( t \) ranging from \( t - 8 \) to \( t - 1 \), and the subpopulation consisting of individuals who either were already surveyable at \( t - 8 \) or became surveyable on a date subsequent to \( t - 8 \) (i.e., who immigrated after \( t - 8 \)) but at \( t \) are members of a household containing at least one person who is surveyable at \( t - 8 \). We consider that if the household at \( t \) contains at least one person who is surveyable at \( t - 8 \), that will be the case on any date between \( t - 8 \) and \( t - 1 \). This ignores situations in which an individual who is in-scope on a given date becomes out-of-scope for a time (as a result of emigration, for example) and then becomes in-scope again.

Next, we use \( \bar{u}_{t, k} \) to denote the cross-sectional sample at \( t \) from panel \( a_{t, k} \), which leads to \( \bigcup_{k=1}^{9} \bar{u}_{t, k} = \bar{u}_{t} \). Let \( Y^{\text{immig}}_{a_{t}} \) be the total of the \( Y_{t} \) defined on \( \Omega^{\text{immig}}_{a_{t}} \). Following the weight share performed for all \( k = 2, \ldots, 9 \), we have

\[
E \left( \sum_{j \in \bar{u}_{t}} \tilde{W}_j(t, k) \cdot Y_j \right) = \sum_{j \in \Omega_{a_{t}}} Y_j - \sum_{a=t-k+1}^{t-1} Y^{\text{immig}}_{a_{t}}
\]

\[
= Y_t - \sum_{a=t-k+1}^{t-1} Y^{\text{immig}}_{a_{t}}
\]

Statistics Canada, Catalogue No. 12-001-X
and
\[ E\left( \sum_{j=1}^{\Omega} \tilde{W}_j(t,1) \cdot Y_j \right) = \sum_{\Omega} Y'_{\Omega} = Y_t \]  
(10)

since \( \tilde{a}_{i,t} = a_{i,t} \).

If we were using shorter-duration panels, we might be able to ignore the \( Y_{\text{im}g} \) and take the actual total over \( \Omega_t \). In that case, the “raw” final cross-sectional weight of any individual \( i \) would be \( \tilde{W}_t(t,k)/9 \) if \( i \) is from \( a_{t,k} \), which would yield the final estimator
\[ \hat{Y}_t = \frac{1}{9} \sum_{k=1}^{9} \sum_{\Omega \in \Omega_{a,t}^\text{im}g} \tilde{W}_t(t,k) \cdot Y'_i \]
(11)

However, since the panels used in France have long lives, we will probably not be able to ignore the \( Y_{\text{im}g} \) (an analysis of the collection files will provide the answer), which will mean having to compute specific weights for the individuals in \( \Omega_{a,t}^\text{im}g \). In those circumstances, we check that any individual \( i \) in \( \Omega_{a,t}^\text{im}g \) who ends up in the cross-sectional sample \( a_{i,t} \) will have a raw cross-sectional weight \( W_{i}^{(2)} \) equal to the weight share value \( \tilde{W}_t(t,k) \) divided by \( t - \alpha \) (and therefore \( 1 \leq t - \alpha \leq 8 \)). Any individual in \( \Omega_t \) who does not belong to any of the \( \Omega_{a,t}^\text{im}g \) (i.e., the vast majority of individuals) will have a final weight of \( \tilde{W}_t(t,k)/9 \). Note that if \( i \) is in \( \Omega_{a,t}^\text{im}g \), he/she can be surveyed only through \( a_{i,1}, a_{i,2}, \ldots, a_{i,t-\alpha} \). Thus we have
\[ W_{i}^{(2)} = \begin{cases} \tilde{W}_t(t,k)/(t-\alpha) & \text{if } i \in \Omega_{a,t}^\text{im}g \\ \tilde{W}_t(t,k)/9 & \text{otherwise} \end{cases} \]  
(12)

In the start-up phase, the weighting process has to be adjusted. In 2005, the final cross-sectional weight of individuals in \( \Omega_{2005,2005}^\text{im}g \) will come directly from the selection of the dwelling from \( a_{2005,1} \) (they can only be reached through this incoming panel). In contrast, all other individuals can be surveyed “normally” in the nine panels \( a_{2005,k} \) \( (1 \leq k \leq 9) \), so that their weights as calculated by the weight share method will all be divided by 9. In 2006, the weights of the individuals in \( \Omega_{2006,2006}^\text{im}g \) will be equal to the weight of the dwelling in which they live, a weight that directly reflects the sampling from \( a_{2006,1} \); the weights of the individuals in \( \Omega_{2004,2006}^\text{im}g \) will be the weights from the weight share divided by 2; and the weights of all other individuals will be the weights from the weight share divided by 9.

This procedure can be carried out for one subsample after another and does not have to take account of what happens in other subsamples. If an individual is surveyed at \( t \) through two (or more) different subsamples \( a_{t,k} \), we carry out the full procedure for each of the two (or more) subsamples. This could occur, for example, in the case of a household composed of two panel members from two different subsamples \( a_{t,k} \) who married each other and before their marriage were each tracked separately as one-person households. In that scenario, each individual would be “formally” surveyed twice, once as a panel member and once as a cohabitant.

Finally, to estimate the difference \( \hat{\Delta}_{t+1} \), we can use the weights \( W_{i}^{(1)} \) from method 1 and calculate
\[ \hat{\Delta}_{t+1} = \sum_{i \in \Omega_{a,t}} W_{i}^{(1)} Y_{i+1}^{(1)} - \sum_{i \in \Omega_{a,t}} W_{i}^{(1)} Y_{i}^{(1)}. \]  
(13)

Alternatively, we can use the weights \( W_{i}^{(2)} \) from method 2. In that case, the estimator of the difference \( \hat{\Delta}_{t+1} \) will be given by
\[ \hat{\Delta}_{t+1} = \sum_{i \in \Omega_{a,t}} W_{i}^{(2)} Y_{i+1}^{(2)} - \sum_{i \in \Omega_{a,t}} W_{i}^{(2)} Y_{i}^{(2)}. \]  
(14)

References


Cell collapsing in poststratification

Jay J. Kim, Jianzhu Li and Richard Valliant ¹

Abstract

Poststratification is a common method of estimation in household surveys. Cells are formed based on characteristics that are known for all sample respondents and for which external control counts are available from a census or another source. The inverses of the poststratification adjustments are usually referred to as coverage ratios. Coverage of some demographic groups may be substantially below 100 percent, and poststratifying serves to correct for biases due to poor coverage. A standard procedure in poststratification is to collapse or combine cells when the sample sizes fall below some minimum or the weight adjustments are above some maximum. Collapsing can either increase or decrease the variance of an estimate but may simultaneously increase its bias. We study the effects on bias and variance of this type of dynamic cell collapsing theoretically and through simulation using a population based on the 2003 National Health Interview Survey. Two alternative estimators are also proposed that restrict the size of weight adjustments when cells are collapsed.

Key Words: Bias; Combining cells; Coverage error; Poststratification; Under-coverage; Weight trimming.

1. Introduction

Poststratification is a common technique used in survey weighting that can serve to (1) reduce variances or (2) adjust for deficient coverage by the sample of some groups in the target population. In household surveys in the U.S. the second purpose is especially important because some demographic groups, like young Black males, are covered less well than others (e.g., see Kostanich and Dippo 2000, chapter 16). Adjusting for undercoverage can lead to differential weights, which may correct for bias but will also increase standard errors. Practitioners often avoid making extreme weight adjustments, in effect trading-off some bias reduction in order to keep variances under control.

One method of controlling the size of weight adjustments is to collapse the initial poststratification cells together if the adjustment in a cell exceeds some limit. Little (1993) and Lazzeroni and Little (1998) cover methods of collapsing categories of ordinal poststratifiers. Other strategies for how to collapse strata or construct estimators have been suggested by Fuller (1966), Kalton and Maligalig (1991), and Tremblay (1986). Kim, Thompson, Wolman, and Vajs (1982) give some practical applications. In this paper, we study the effects on bias and variance of combining cells, assuming that more finely defined cells would be preferable if the sample sizes and sizes of weight adjustments were within some tolerances set by the survey designers.

Two criteria are often used to decide whether a cell should be collapsed with another. The first is the inverse coverage ratio or initial adjustment factor (IAF), and is defined as the ratio of the control count to the initially weighted sample count for the cell. A ratio which is significantly different from 1 indicates that coverage is either low or high for the group represented by the cell. When the IAF for a cell falls outside some bounds set in advance, the cell is combined with another. For example, the collapsing threshold for “high” ratio might be 2 and the threshold for “low” ratio 0.6, which are the bounds used in the Current Population Survey (CPS) conducted by U.S. Bureau of the Census (see Kostanich and Dippo 2000, page 10-7). The second criterion is the sample size. A cell whose raw sample count is too small may be collapsed on the grounds that the IAF is unstable. We will refer to a cell as sparse if it violates one or the other of the criteria and is collapsed with another cell.

The categories of the variables that define poststrata are usually sorted based on a natural ordering (e.g., age or income categories) or a convenient ordering (e.g., race-ethnicity). Common practice is to collapse a cell with an adjacent one which is similar in characteristics, disregarding different coverage ratios of the individual cells.

Kalton and Flores-Cervantes (2003, page 95) observed that “methods that automatically restrict the range of the adjustments are redistributing the excess adjustments that would otherwise be given to some respondents to other respondents. The appropriateness of this redistribution should be examined.” This paper indeed examines its appropriateness and identifies circumstances where the weight redistribution due to collapsing may be quite harmful.

An obvious weakness of popular collapsing strategies is that coverage bias for some groups will be incompletely corrected. For example, suppose that the survey estimate for the number of units in a group is only 1/3 of the census count, so that initial weights would have to be multiplied by 3 to correct for undercoverage. If cell collapsing restricts the

¹ Jay J. Kim, National Center for Health Statistics, Centers for Disease Control and Prevention; Jianzhu Li, Joint Program in Survey Methodology, University of Maryland; Richard Valliant, Survey Research Center, University of Michigan.
weight adjustment for units in that group to a factor of 2, then the survey estimate for the number of units in the group will be only 2/3 of the census count. In addition, if cells with much different means are combined, bias can be introduced rather than corrected. The incomplete correction for undercoverage and collapsing of cells with disparate characteristics may lead to bias in totals, means, and other types of estimates.

Table 1 gives some illustrative coverage ratios, i.e., survey estimates prior to poststratification divided by census counts, for the March 2002 U.S. Current Population Survey and the 2003 Behavioral Risk Factors Surveillance Survey (BRFSS) in a set of 44 counties in the southwestern U.S. The survey estimates include a nonresponse adjustment for both CPS and BRFSS. Coverage ratios shown for the subset of demographic groups in the CPS range from 0.70 to 0.93 with Black-only typically being less than for other groups. BRFSS is a telephone survey with low response rates in this set of counties, and the ratios for BRFSS are much smaller than for CPS. There are also substantial differences in coverage ratios for different groups in BRFSS. For example, the ratio for 35 - 44 year old Hispanic males is 0.18 but is 0.37 for Black/Multiracial/Other males in the same age range. If these two groups were collapsed, incomplete coverage would be under-corrected for the Black/Multiracial/Other group. Another example is the 2003 National Health Interview Survey (NHIS) where American Indians and Asians were collapsed with Whites within age groups. In the cell for ages 25-29, for example, the coverage rates for Whites, American Indians, and Asians were 0.60, 0.44, and 0.31, respectively (Tompkins and Kim 2006).

This paper demonstrates the weaknesses of current cell collapsing procedures and proposes some alternatives. Section 2 discusses the bias of some standard estimators when there is undercoverage. Section 3 introduces two new estimators that retain more of the undercoverage adjustment than the standard method when cells are collapsed. Empirical properties of the standard and alternative methods are investigated through simulation in section 4. We conclude in section 5 with a summary and some possibilities for future research.

2. Some standard estimators

Three standard estimators of the population mean are the Hájek estimator, the poststratified estimator, and the poststratified estimator with initial poststrata collapsed together where necessary. Each of these is defined in detail below. When the sampling frame covers units in the target population at different rates, each estimator can be biased. Kim et al. (2005) give some numerical illustrations of the effects of collapsing pairs of cells with different coverage rates.

To derive theory for alternative estimators of means, we model a unit’s being covered by the sampling frame and a cell’s being sparse or not as random events. Define three indicator variables: \( \delta_k = 1 \) if unit \( k \) is selected for the sample and 0 if not; \( c_k = 1 \) if unit \( k \) is covered by the frame and 0 if not; \( d_i = 1 \) if poststratum \( i \) is classified as sparse in a particular sample and 0 if not (\( i = 1, \ldots, I \)). These indicators are assumed to be mutually independent and to have expectations \( \pi_z, \phi_k, \) and \( p_i \), respectively. Consider a stratified, two-stage probability sample design. A design stratum is denoted by \( h; s_k \) is the set of primary sampling units (PSU’s) selected from design stratum \( h; s_{bij(i)} \) is the set of sample units from sample PSU \( j \) in stratum \( h \) that are also in poststratum \( i; U_h \) is the population of PSU’s in stratum \( h; U_{hj} \) is the population of units in PSU \( j \) within stratum \( h; \) and \( U_{hbi(i)} \) is the population of units in PSU \( j \) within stratum \( h \) that are in poststratum \( i \). For the analysis in this section, the sample design does not need to be specified in more detail. Note that some summations in sections 2 and 3 of the form \( \sum_k \sum_{j=1}^{I} \sum_{s_{bij(i)}} y_k / \pi_k \) for units within poststratum \( i \) could be simplified to \( \sum_k \sum_{s_{bij(i)}} 1 / \pi_k \) without loss of generality. We have used the more elaborate notation to make clear how the stages of sampling should be treated.

2.1 Hájek estimator

First, consider the Hájek estimator of a mean, which is

\[
\hat{\pi} = \frac{\sum_{i=1}^{I} \sum_{h} \sum_{j=1}^{I} \sum_{s_{bij(i)}} y_k / \pi_k}{\sum_{i=1}^{I} \sum_{h} \sum_{j=1}^{I} \sum_{s_{bij(i)}} 1 / \pi_k} \equiv \hat{T} / \hat{N}. \quad (1)
\]

The expectation of \( \hat{T} / \hat{N} \) with respect to sampling and the coverage mechanism is 

\[
E[E_i(\hat{T})] = \sum_{i=1}^{I} \sum_{h} \sum_{j=1}^{I} \sum_{s_{bij(i)}} \phi_k y_k \equiv T^c ,
\]

where the c superscript denotes “covered”. Similarly, the expectation of \( \hat{N} \) is 

\[
E[E_i(\hat{N})] = \sum_{i=1}^{I} \sum_{h} \sum_{j=1}^{I} \sum_{s_{bij(i)}} \phi_k \equiv N^c .
\]

Expanding \( \hat{T} / \hat{N} \) around \( (T^c, N^c) \), its linear approximation is 

\[
\hat{T} / \hat{N} \approx \frac{T^c}{N^c} + 1/N^c \times (\hat{T} - (T^c/N^c) \hat{N}).
\]

Next, consider the bias of \( \hat{T} / \hat{N} \) as an estimator of \( T = \sum_{i=1}^{I} T_i / N \) with \( T_i \) being the total for the full population of units in poststratum \( i \) (not just the covered portion). After some calculation, the bias is

\[
\text{bias}(\hat{T} / \hat{N}) = \frac{T^c}{N^c} \left( 1 - \frac{1}{N} \sum_{i=1}^{I} T_i \right) = \frac{C_w}{\bar{\phi}} \quad (2)
\]

where \( \bar{\phi} = \sum \phi_k / N, \quad C_w = \sum (\phi_k - \bar{\phi}) (y_k - \bar{y}) / N, \quad \bar{y} = T / N, \) and \( \sum \) denotes the sum over \( i, h, j \in U_h, \) and \( k \in U_{hj(i)} \). Consequently, \( \hat{T} / \hat{N} \) is biased if there is any correlation between the variable measured, \( y \), and the coverage probability \( \phi_k \). The bias in (2) is \( O(1) \), meaning that it remains important even in large samples.
If the coverage probability is the same for every unit in poststratum \( i \), i.e., \( \phi_k = \phi(i) \) for any \( k \in U_{y(i)} \), then the approximate bias reduces to \( \text{bias}(\bar{y}_i) = \phi^{-1} \sum W_i \times (\phi(i) - \bar{\phi})(\bar{y}_i - \bar{Y}) \) where \( W_i = N_i / N \) and \( \bar{Y} = \sum_{h \in U_{y(i)}} y_h / N_i \).  If there is a correlation between the poststratum coverage probabilities and the poststratum means, the Hájek estimator will again be biased, and the bias could be either positive or negative.  If the coverage rates or the means are constant across poststrata, i.e., \( \phi(i) = \bar{\phi}_o \) or \( \bar{Y} = \bar{y} \), then the Hájek estimator will be unbiased, but poststrata are usually not formed this way.  Also, the bias exists even when the appropriate set of poststrata, that subdivide the population into groups with different means, is unknown to the sampler.

### 2.2 Poststratified mean with no cell collapsing

The poststratified mean is defined as 
\[
\bar{y}_{PS1} = \frac{1}{N} \sum_{i=1}^{I} \left( N_i / \hat{N}_i \right) \hat{T}_i
\]
where \( \hat{T}_i \) and \( \hat{N}_i \) are defined as in (1) but excluding the summation over \( i \).  Define 
\[
T_i^c = \sum_{h \in U_{y(i)}} \phi_h y_h \quad \text{and} \quad N_i^c = \sum_{h \in U_{y(i)}} \phi_h \]
These are the expected (with respect to the coverage mechanism) total and count of covered units in poststratum \( i \).  Expanding \( \hat{y}_{PS1} \) around \( (T_i^c, N_i^c) \), \( i = 1, \ldots, I \), its linear approximation is
\[
\hat{y}_{PS1} \approx \frac{1}{N} \sum_{i=1}^{I} \frac{N_i}{N_i^c} T_i^c + \sum_{i=1}^{I} \frac{N_i^c}{N_i} \left( \hat{T}_i - T_i^c / N_i^c \hat{N}_i \right)
\]
The bias of the poststratified estimator is then the first term of this expression minus \( \sum_{i=1}^{I} T_i / N \), and after some manipulation, can be written as
\[
\text{bias}(\hat{y}_{PS1}) \approx \sum_{i=1}^{I} W_i \frac{C_{\phi(i)}}{\phi(i)} \tag{3}
\]
where \( \bar{\phi} = \sum \phi_k / N \), \( C_{\phi(i)} = \sum (\phi_k - \bar{\phi})(y_k - \bar{y}) / N_i \), and \( \bar{\phi} \) denotes the sum over \( h, j \in U_{\phi(i)} \), and \( k \in U_{y(i)} \).  Thus, \( \hat{y}_{PS1} \) is biased if there is any correlation between the \( y \) variable measured and the coverage probability \( \phi_k \) in any of the poststrata.  If the coverage rate is constant at \( \phi_k = \phi(i) \) within poststratum \( i \), then the poststratified estimator is approximately unbiased.  From (3) it is apparent that poststrata should be formed so that either coverage rates or the \( y \)'s are homogeneous within each poststratum.  This is similar to the recommendations of Eltinge and Yansaneh (1997), Kalton and Maligalig (1991), and Little and Vartivarian (2005) for the formation of nonresponse adjustment cells.  In large surveys, the initial set of candidate poststrata is often more extensive than the sample can support.  With few exceptions, some of the initial poststrata are collapsed to control weight adjustments.  If no collapsing occurs, this is usually because small categories are pre-collapsed based on prior experience in the same or a similar survey.  In that sense, PS1 does not really exist in practice.  More common is the collapsing approach, PS2, described below.

### 2.3 Poststratified mean with collapsing

Turning to the poststratified estimator with collapsing, the sparse cells are identified and combined with other cells considered to be their nearest neighbors.  This could result in more than one sparse cell being collapsed with a given nonsparse cell.  Neighbors can be defined in various ways, e.g., cells with similar estimated coverage rates, \( \hat{N}_k / N_i \), cells that are adjacent in some substantive sense like nearby income classes, or cells that have similar means on some important survey variables.  The general algorithm for collapsing, given an initial set of cells, is:
(1) Compute the collapsing criteria for each cell, e.g., the IAF’s, \(N_1/\hat{N}_{g}\), and the cell sample sizes;

(2) Identify the sparse cells, i.e., those whose criteria fall outside the bounds for collapsing;

(3) Determine the nearest, non-sparse neighbor of each sparse cell and combine the sparse cell with its neighbor.

The poststratified mean with collapsing is then

\[
\hat{y}_{PS2} = \frac{1}{N} \sum_{g} \frac{N_g}{N} \left( \frac{y_{g}}{N} \hat{T}_{g} - \frac{\hat{T}_{g} - T_{g}}{N_g} \hat{N}_{g} \right).
\]

It follows that

\[
\text{bias}(\hat{y}_{PS2}) \cong \sum_{g} W_g \frac{C_{g,g}}{\hat{\phi}_g} \tag{4}
\]

with

\[
W_g = N_g/N, \quad \hat{\phi}_g = \sum \phi_k/N_g, \quad C_{g,g} = \sum (\phi_k - \hat{\phi}_g)(y_k - \bar{y}_g)/N_g,
\]

and the summations in \(\bar{y}_g\) and \(C_{g,g}\) are over \(i \in A_g, h, j \in U_h, \) and \(k \in U_{h(j)}\). If \(\hat{\phi}_g\) is constant within collapsed group \(g\), this estimator is unbiased, but if \(\phi_k = \phi(i)\), i.e., the coverage rate is constant within poststratum \(i\) but can differ across the poststrata, then the bias becomes

\[
\text{bias}(\hat{y}_{PS2}) \cong \sum_{g} W_g \frac{C_{g,g}^*}{\hat{\phi}_g} \tag{5}
\]

with \(\hat{\phi}_g = \sum W_g \phi(i), \quad C_{g,g}^* = \sum W_g (\hat{\phi}_g - \phi(i))(\bar{y}_g - \bar{y}_g)/N_g\), and the summations are over \(i \in A_g\).

Thus, in the case where \(\bar{y}_g\) will be unbiased, \(\hat{y}_{PS2}\) will be biased if poststrata are collapsed together that have different coverage rates and different population means. Since \(\hat{\phi}_g\) and \(C_{g,g}^*\) are both \(O(1)\), the bias does not decrease as the sample increases; thus, the bias-squared will eventually be the dominant part of the mean square error. If cells are collapsed, the cells in each group should have the same coverage rates, the same means, or both to avoid bias.

### 3. Weight restricted estimators

We examine two alternative methods of weight computation when collapsing of poststrata is used, extending work of Kim (2004). The alternatives are designed to be compromises between (a) use of all poststrata and the potential for large weight adjustments and (b) collapsing of poststrata yielding less variable weights but potentially biased estimates. We refer to these as weight restriction (WR) methods. The two alternatives presented in this section use cell collapsing but retain a larger share of the weight adjustment for individual cells than does the standard collapsing method.

The first alternative is denoted PS.WR1 and consists of the following algorithm. Denote the maximum allowable weight adjustment by \(f_{\text{max}}\) with \(f_{\text{max}} > 1\).

(1) Execute steps (1) - (3) of the algorithm in section 2.3 for PS2.

(2) Censor any IAF greater than \(f_{\text{max}}\) to \(f_{\text{max}}\) and adjust each weight in the corresponding initial cell to \(\tilde{w}_k = w_k f_{\text{max}}\) with \(w_k = 1/\pi_k\). For units in cells with \(\text{IAF} \leq f_{\text{max}}\), set \(\tilde{w}_k = w_k\).

(3) Compute a collapsing adjustment factor (CAF) for a collapsed group \(g\) as

\[
\tilde{f}_g = N_g \sum_{i \in A_g} \sum_{h, j \in \Phi_{h(i)}} \tilde{w}_k.
\]

(4) The final adjusted weight is then \(\tilde{w}_k \tilde{f}_g\) for unit \(k\) in group \(g\).

This method will reduce the largest values of the final weight adjustment below the without-collapsing adjustments, \(N_i/\hat{N}_{g}\), though there may be one or more groups that have CAF’s greater than the \(f_{\text{max}}\) cutoff. The control total for group \(g\), \(\hat{N}_{g}\), is met in the sense that

\[
\sum_{i \in A_g} \sum_{h, j \in \Phi_{h(i)}} \tilde{w}_k \tilde{f}_g = N_g
\]

but the control totals for the individual cells in \(A_g\) are not.

To analyze the properties of PS.WR1, define \(A_{g,up}\) and \(A_{g,sp}\) to be the sets of sparse and non-sparse poststrata in collapsed group \(g\). PS.WR1 can be expressed as

\[
\hat{y}_{PS.WR1} = \frac{1}{N} \sum_{g} \frac{N_g}{\hat{N}_{g,WR1}} \hat{y}_{g,WR1}\]

where

\[
\hat{y}_{g,WR1} = \sum_{i \in A_g} \sum_{h, j \in \Phi_{h(i)}} \sum_{k \in \Phi_{h(k)}} \tilde{w}_k y_k + \sum_{i \in A_g} \sum_{h, j \in \Phi_{h(i)}} \sum_{k \in \Phi_{h(k)}} \tilde{w}_k y_k
\]

and \(\hat{N}_{g,WR1}\) has a similar definition with \(y_k\) set to 1. The expectation of \(\hat{y}_{g,WR1}\) over the coverage, sparseness, and sampling mechanisms is \(E_{g} E_{w} E_{x}(\hat{y}_{g,WR1}) = T_g + (f_{\text{max}} - 1) \times \hat{T}_{g}\), where \(\hat{T}_g = \sum_{i \in A_g} \pi_i T_i\). Likewise, \(E_{g} E_{w} E_{x}(\hat{N}_{g,WR1}) = \hat{N}_g + (f_{\text{max}} - 1) \hat{N}_g\) with \(\hat{N}_g = \sum_{i \in A_g} \pi_i N_i\). \(\hat{y}_{PS.WR1}\) can be expanded around the expectations of \(\hat{y}_{g,WR1}\) and \(\hat{N}_{g,WR1}\). After some manipulation, the approximate bias of \(\hat{y}_{PS.WR1}\) becomes

\[
\text{bias}(\hat{y}_{PS.WR1}) \cong \sum_{g} W_g \frac{C_{ab, y, g}}{\hat{\alpha}_g} \tag{6}
\]

where

Statistics Canada, Catalogue No. 12-001-X
\[
(\alpha \phi)_i = \sum \alpha_i \phi_k / N_g, \quad \alpha_i = 1 + (f_{\text{max}} - 1) p_i,
\]
\[
C_{\alpha y, y, g} = \sum (\alpha \phi_k - \bar{\phi}_g)(y_k - \bar{y}_g) / N_g,
\]
and the summations are over \( i \in A_g, h, j \in U_h, \) and \( k \in U_{h(i)} \). In the case of a common coverage probability in poststratum \( i, \) i.e., \( \phi_k = \phi(i) \), we have \( (\alpha \phi)_i = \bar{\phi}_g + (f_{\text{max}} - 1)(p \phi)_i \) and
\[
\alpha, \phi(i) - (\alpha \phi)_i = (\phi(i) - \bar{\phi}_g) + (f_{\text{max}} - 1)(p \phi(i) - (p \phi)_i)
\]
where \( (p \phi)_i = \sum A_g W_g p \phi(i) \). From this it follows that
\[
C_{\alpha y, y, g} = C_{\phi y, g} + (f_{\text{max}} - 1)C_{p \phi, y, g}
\]
with
\[
C_{p \phi, y, g} = \sum A_g W_g (p \phi(i) - (p \phi)_i)(\bar{y}_g - \bar{y}_g).
\]
If the cell means \( \bar{y}_g \) are all equal within a collapsed group, then \( C_{\phi y, g} = C_{p \phi, y, g} = 0 \) and \( \hat{y}_g \) will be approximately unbiased. In the special case in which coverage is constant in a group, i.e., \( \phi(i) = \bar{\phi}_g \), then \( C_{\alpha y, y, g} = (f_{\text{max}} - 1) \bar{\phi}_g \sum A_g W_g (p \phi(i) - \bar{\phi}_g)(\bar{y}_g - \bar{y}_g) \) with \( \bar{y}_g = \sum A_g W_g p \phi(i) \). Thus, if \( p \) and \( \phi(i) \) are constant within \( A_g \), then \( \hat{y}_g \) will be nearly unbiased even if the cell means \( \bar{y}_g, i \in A_g \), differ. This condition is almost sure to be false as long as one poststratum in a group has a probability of being sparse that is substantially different from the others.

In the case of a common coverage probability in poststratum \( i, \phi(i) \), we can also compare the biases of the collapsed cell estimator, \( \hat{y}_g \), with that of \( \hat{y}_g \). Using results in the previous paragraph, the bias in (6) can be expressed as
\[
\text{bias}(\hat{y}_g) = \sum W_g \left[ \frac{C_{\phi y, g}}{\bar{\phi}_g} + (f_{\text{max}} - 1)C_{p \phi, y, g} / \bar{\phi}_g \right]
\]
Since \( 1 + (f_{\text{max}} - 1)(p \phi)_i / \bar{\phi}_g \geq 1 \), we can use (5) to obtain
\[
\text{bias}(\hat{y}_g) \leq \sum W_g \left[ \frac{C_{\phi y, g}}{\bar{\phi}_g} + (f_{\text{max}} - 1)C_{p \phi, y, g} / \bar{\phi}_g \right]
\]
If \( p \phi(i) \) and \( \bar{y}_g \) are uncorrelated, the absolute bias of \( \hat{y}_g \) is less than or equal to that of \( \hat{y}_g \) because \( 1 + (f_{\text{max}} - 1)(p \phi)_i / \bar{\phi}_g \geq 1 \). When \( p \phi(i) \) and \( \bar{y}_g \) are correlated, there are two cases to consider: (i) bias \( \text{bias}(\hat{y}_g) \geq 0 \) and (ii) bias \( \text{bias}(\hat{y}_g) < 0 \). In the former, the last line of (7) will be less than or equal to the absolute bias of \( \hat{y}_g \) if
\[
-2 \frac{\text{bias}(\hat{y}_g)}{f_{\text{max}} - 1} \leq \sum W_g C_{p \phi, y, g} / \bar{\phi}_g \leq 0.
\]
In case (ii), the requirement is
\[
0 \leq \sum W_g C_{p \phi, y, g} / \bar{\phi}_g \leq 2 \frac{\text{bias}(\hat{y}_g)}{f_{\text{max}} - 1}.
\]
If the covariance between the probability of being sparse and covered, \( p \phi(i) \), and the cell means, \( \bar{y}_g \), is small in all groups and the opposite sign of bias \( \hat{y}_g \), then \( \hat{y}_g \) will be less biased than \( \hat{y}_g \).

The second alternative is denoted PS.WR2 and is intended to exercise more control over the size of the final weight adjustment than does PS.WR1. In PS.WR1 the final adjustment can be larger than \( f_{\text{max}} \). PS.WR2 seeks to limit the final adjustment to \( f_{\text{max}} = 2 \) or some other maximum set in advance. The general idea is to first determine which cells should be collapsed together, as was done for PS.WR1. Then weights in the sparse cells are multiplied by \( f_{\text{max}} \). The weights in the non-sparse cell in a collapsed group are then adjusted by a constant factor to bring the estimated population count in the group to the control count. The detailed algorithm for computing weights for PS.WR2 is the following:

1. Execute steps (1) - (3) of the algorithm in section 2.3 for PS2.
2. In a group containing at least one non-sparse cell, compute the control total in group \( g \) as \( N_g = \sum A_g N_i \) and the adjusted weight for all units \( k \in A_g \) as \( \tilde{w}_k = w_k f_{\text{max}} \).
3. Compute the adjusted weight for all units \( k \in A_g \) as \( \tilde{w}_k = w_k (N_g - \hat{N}_g) / \tilde{N}_g \) where \( \hat{N}_g = \sum A_g \sum_{h \in A} \sum_{j \in A} \sum k \in A \sum_{s \leq k} \tilde{w}_k \) and \( \tilde{N}_g = \sum A_g \sum_{h \in A} \sum_{j \in A} \sum_{s \leq k} \tilde{w}_k \).
4. The final adjusted weight is then \( \tilde{w}_k \) for unit \( k \) in group \( g \).

This second weight restricted estimator can be written as \( \tilde{y}_g = (1 / N_g) \tilde{W}_g \) where
\[
\tilde{W}_g = \sum A_g \sum_{h \in A} \sum_{j \in A} \sum_{s \leq k} \tilde{w}_k \tilde{y}_k,
\]
\[
\tilde{W}_g = \sum A_g \sum_{h \in A} \sum_{j \in A} \sum_{s \leq k} \tilde{w}_k \tilde{y}_k,
\]
\[
\tilde{W}_g = \sum A_g \sum_{h \in A} \sum_{j \in A} \sum_{s \leq k} \tilde{w}_k \tilde{y}_k.
\]

The expectation of \( \hat{y}_g \) with respect to the coverage, sparseness, and sampling mechanisms is
\[
E_{c, E_{ps}, E_s}(\hat{y}_g) = \sum A_g \sum_{h \in A} \sum_{j \in A} \tilde{w}_k \tilde{y}_k
\]
where \( \hat{N}_g = \sum_{h \in A_g} N_i \tilde{w}_k \). After some calculation, the approximate bias of \( \hat{y}_g \) can be written as
Next, note that in the case of a common coverage proportion covered in sparse cells in group \( g \), the common mean of 0.20 regardless of the unit’s poststratum membership. In that case all estimators, including the Hájek estimator, will be unbiased regardless of coverage rates. Also, the conventional thinking that collapsing of cells may reduce variances by smoothing out extreme weight adjustments may hold for this variable.

4.2 Sample design

Two sample PSU’s were selected in each stratum with probability proportional to size (PPS) with the size being the count of persons in each PSU. Sampling of PSU’s was done without-replacement to simplify variance estimation. If without-replacement sampling had been used, then a more elaborate method of selection and variance estimation would have been needed (see, e.g., Särndal, Swenson, and Wretman 1992, chapter 3). In each sample PSU, 20 persons were selected by simple random sampling without replacement for a total of 1,000 persons in each sample. For each combination of parameters discussed below, 2,000 samples were selected. Sixteen initial poststrata were used which were the cross of the eight age groups, shown in Table 2, with gender. In public-use file. A subset of the NHIS was created with 21,664 persons. These were divided into 25 strata with each having six PSUs. The strata and PSU’s are based on those in the NHIS public use file, but sets of three strata were collapsed together to create new design strata for the study population. We used four binary variables (0-1 characteristics) for the simulation, each of which is based on a person’s self-report:

- Health insurance coverage - whether a person was covered by any type of health insurance;
- Physical, mental, or emotional limitation - whether a person was limited in any of these ways;
- Medical care delayed - whether a person delayed medical care or not because of cost in last 12 months;
- Overnight hospital stay - whether a person stayed overnight in a hospital in last 12 months.

Table 2 shows the percentages of persons with these four characteristics in cells formed by age and sex. These 16 (age x sex) cells are the initial set of poststrata used in estimation. The percentages can vary substantially among the cells, depending on the characteristic. For example, 18-24 year olds are much more likely to have no health insurance; children under age 5 and the elderly age 65 and over are much more likely to have had a hospital stay. Collapsing cells together that have different means, or proportions in this case, has the potential to introduce bias, as noted earlier.

We also created one artificial binary variable that had a common mean of 0.20 regardless of the unit’s poststratum membership. In that case all estimators, including the Hájek estimator, will be unbiased regardless of coverage rates. Also, the conventional thinking that collapsing of cells may reduce variances by smoothing out extreme weight adjustments may hold for this variable.

4. An empirical investigation

To test some of the ideas presented earlier, we conducted a simulation study of the bias properties of alternative methods of poststratification. We also examined the performance of one variance estimator that is often used in practice.

4.1 Study population

The population used in the simulation was extracted from the 2003 National Health Interview Survey (NHIS) person
each sample, we computed the estimators of population proportions, described earlier in sections 2-3 - the Hájek estimator, \( \hat{\pi}_n \), the poststratified estimator \( \hat{\pi}_{PS} \), that uses all 16 poststrata, the poststratified estimator with collapsing of cells, \( \hat{\pi}_{PS2} \), and the two weight-restricted estimators, \( \hat{\pi}_{PS.WR1} \) and \( \hat{\pi}_{PS.WR2} \). The simulation code was written in the R language (R Development Core Team 2005) with extensive use of the R survey package (Lumley 2004, 2005).

4.3 Coverage mechanisms

Five sets of coverage mechanisms, shown in Table 3, were employed to filter the population before the PSU’s were sampled. The coverage ratios varied by poststratum and were different for each of the five characteristics for which proportions were estimated. The coverage ratios specific to each of the five characteristics are named C1 through C5 in Table 3. These coverage ratios were artificially created based on the population means for each age and sex group. Poorer coverage was assigned to groups with larger percentages with a characteristic for health insurance coverage and limitations; the opposite was true for delayed medical care and hospital stays. In C5 the coverage ratios are quite variable and are intended to lead to coverage adjustments that vary substantially among the initial set of 16 poststrata. Although the rates in Table 3 are low, they are comparable to or higher than those for BRFSS in Table 1. In applying these rates, we randomly selected a subset of the population to be in the sample frame for each sample that was selected. For example, if the coverage ratio in the poststratum of males younger than 5 years old is 0.9, then 90% of the population in that poststratum was randomly selected to stay in the sampling frame while the rest had a zero probability of being sampled.

4.4 Collapsing rules

We set up situations where the conditions for unbiasedness in sections 2 and 3 can be violated when cells were collapsed in the simulations. Each of the estimators, \( \hat{\pi}_{PS2} \), \( \hat{\pi}_{PS.WR1} \), and \( \hat{\pi}_{PS.WR2} \) involve cell collapses. If the IAF (poststratification factor) in an initial poststratum, \( \frac{N_i}{N} \), exceeds the maximum allowable adjustment, \( f_{i,\text{max}} \), or if the cell sample size is less than a minimum, \( n_{i,\text{min}} \), we call this poststratum a “sparse” cell and collapse it with a neighboring cell. We used two methods of determining neighbors, designated here as “adjacency” and “close-mean”.

In adjacency collapsing, the neighbors of a specific cell are defined as the cells either horizontally or vertically adjacent to it in the age \( \times \) sex table. For example, in the following, abbreviated table, the neighbors of cell 3 are the shaded cells 2, 4, and 7.

![Table 3 Coverage ratios used in the simulations](image)

![Table 2 Percentages of persons with four health-related characteristics in groups formed by age and sex](image)
4.5 Variance estimation

For each of the estimators of a proportion, a linearization variance estimate was calculated. Each of the variance estimators is based on the linear substitute method (e.g., see Särndal et al. 1992, chapter 5). The variance estimates for all estimators of proportions were computed using the svydesign, poststratify, and svymean functions in the R survey package. The general, theoretical approach is to make a linear approximation for a particular estimator. The linear approximation is rearranged so that the estimator is written as a sum of weighted PSU totals, and the variance estimator for with-replacement PSU sampling is used. The estimators \( \hat{Y}_{PSI}, \hat{Y}_{PS} , \hat{Y}_{PS.WR1}, \) and \( \hat{Y}_{PS.WR2} \) are treated as standard poststratified estimators for the purposes of variance estimation. For \( \hat{Y}_{PSI} \) define the following:

\[
\begin{align*}
  u_k &= \frac{N_i}{\hat{N}_{ig}} (y_k - \hat{y}), k \in s_i, \quad \text{with } \hat{y} = \frac{\hat{N}_{ig}}{N_g}, \\
  \bar{u}_{hg} &= \sum_{i, k \in s_{ig}} \frac{w_i}{d_i} u_k, \quad \text{and} \\
  \bar{u}_{hg} &= \frac{1}{n_d} \sum_{j \in s_{ig}} \bar{u}_{hg},
\end{align*}
\]

The variance estimator for \( \hat{Y}_{PSI} \) is then

\[
y(\hat{Y}_{PSI}) = \frac{1}{N^2} \sum_k n_k \left( \frac{n_k}{n_k - 1} \sum_{j \in s_k} (\bar{u}_{hg} - \bar{u}_{hg})^2 \right). \tag{9}
\]

For the collapsed stratum estimator, \( \hat{Y}_{PS2} \), \( \hat{Y}_{PS2} \) applies with the linear substitute defined as

\[
u_k = \frac{N_g}{\hat{N}_{ig}} (y_k - \hat{y}), k \in s_g,
\]

with \( \hat{y} = \frac{\hat{N}_{ig}}{N_g} \) and \( s_g \) is the set of all sample units in group \( g \).

In the cases of PS.WR1 and PS.WR2, we calculate the final weights as described in section 3 and call the R poststratify function. This results in the linear substitute being computed as

\[
u_k = \frac{N_g}{\hat{N}_{ig}} (y_k - \hat{y}) = y_k - \hat{y}/\hat{N}_{ig}
\]

because \( \hat{N}_{ig} = \sum_{i, k \in s_{ig}} \bar{u}_k = N_g \). The mean \( \hat{y} \) is computed as

\[
\hat{y} = \frac{\sum_{i, k \in s_{ig}} \bar{w}_k y_k / \sum_{i, k \in s_{ig}} \bar{w}_k}{N_g}.
\]

The weighted linear substitute is then \( u_k = \bar{w}_k (y_k - \hat{y}/\hat{N}_{ig}) \) and

\[
\bar{u}_{hig} = \sum_{j \in s_{ig}} \bar{w}_j u_k.
\]

In the cases of PS2, PS.WR1, and PS.WR2, these variance estimators do not account for the dynamic nature of cell collapsing which can vary from sample to sample. Consequently, there is a source of variation that is not accounted for, and we can anticipate that the variance estimates will be somewhat too small compared to empirical, simulation variances.

4.6 Simulation results

Tables 4-7 summarize results for coverage correction errors, relative biases of estimated proportions, variances of alternative estimators, and confidence interval coverage using linearization variance estimators. Table 4 shows average absolute coverage correction error, defined as

\[
\bar{e} = (DF)^{-1} \sum_{i=1}^{D} \sum_{d=1}^{d} \left| \hat{N}_{di} / N_i - 1 \right| \tag{10}
\]

where \( d \) is one of the \( D=2,000 \) samples and \( \hat{N}_{di} \) is the estimated number of units in poststratum \( i \) based on the final weights for a particular estimator (Hájek, PS2, PS.WR1, or PS.WR2). The value of \( \bar{e} \) is 0 for the poststratified estimator with no cell collapsing, PSI, since it corrects coverage error completely in each of the 16 poststrata. To illustrate how the average coverage correction errors can vary, we estimated the proportions for the health insurance and common mean \( Y \) variable using the C1 and C5 frame coverage ratios. For most combinations of coverage ratios, collapsing method, and adjustment bound, PS.WR1 more effectively corrects for coverage error than the standard.

Kim, Li and Valliant: Cell collapsing in poststratification.
collapsing estimator, PS2. For example, $\bar{\tau} = 0.086$ with PS.WR1 for (health insurance, adjacency collapsing, $f_{\text{max}} = 2$) while PS2 has $\bar{\tau} = 0.120$. In contrast, PS.WR2 is somewhat worse than PS2 in coverage correction.

Table 5 presents the relative biases (rebias), defined as $100\sum_{d=1}^{D}(\hat{y}_d - \bar{Y})/\bar{Y}$ where $\hat{y}_d$ is one of the estimates of proportion for sample $d$. The Hájek estimates are badly biased for the first four characteristics since they include no correction for the differential undercoverage among the cells. The rebias ranges from -12.1% to 13.4% for hospital stay. As noted in section 2, the bias can be either positive or negative, depending on the correlation of coverage rates and cell means.

Poststratification with no collapsing of cells (PS1) gives nearly unbiased estimates while the alternatives - PS2, PS.WR1, and PS.WR2 - all introduce a bias when using adjacency collapsing for the first four characteristics. The number of poststrata after collapsing, shown in Table 5, ranges from 6 to 16 when $f_{\text{max}} = 2$ and from 5 to 13 when $f_{\text{max}} = 1.8$. The relative biases of PS2, using adjacency collapsing, range from -4.4% to 6.2% when $f_{\text{max}} = 2$ and from -6.5 to 9.4% when $f_{\text{max}} = 1.8$. With adjacency collapsing, The alternatives, PS.WR1 and PS.WR2, have biases that are intermediate between PS1 (no collapsing) and PS2. PS.WR1, in particular, is reasonably competitive with PS1 in terms of bias with adjacency collapsing. In contrast, close-mean collapsing yields PS2, PS.WR1, and PS.WR2 estimates that are essentially unbiased when $f_{\text{max}} = 2$. With mean collapsing and $f_{\text{max}} = 1.8$, PS2 and PS.WR2 are still somewhat biased, but PS.WR1 compares well with PS1. For the fifth characteristic (Common mean $Y$), all estimators are nearly unbiased, regardless of collapsing method, as expected.

One justification that is conventionally given for collapsing cells is that extreme weights will be reduced and variations of estimates will, in turn, be reduced. Table 6 shows the ratios of the empirical variances of estimated proportions as a proportion of the variance of PS1. The Hájek estimates have variances that are about 12% and 18% smaller than those of PS1 for health insurance and limitations, but are more variable than PS1 for delayed care and hospital stay. These results also make it clear that the variance of a poststratified estimator can be either increased or decreased by collapsing. There are some minor variance gains from using PS2 for some combinations for the first four variables, but with (adjacency, $f_{\text{max}} = 2$) the PS2 variance of hospital stay is 17% larger than that of PS1. With (adjacency, $f_{\text{max}} = 1.8$), PS2 is 23% more variable for hospital stay. PS.WR1 does not have the extreme variances of PS2 in adjacency collapsing; like PS2, PS.WR2 has larger variance for hospital stay in adjacency collapsing.

When close-mean, rather than adjacency, collapsing is used, variances of PS2, PS.WR1, and PS.WR2 are much closer to those of PS1. However, for the Common Mean $Y$ variable, collapsing always reduces variance. The reductions are almost 20% for adjacency collapsing.

The right-hand section of Table 6 lists the ratios of the empirical mean square errors (MSEs) of estimated proportions as a proportion of the MSE of PS1. With a few exceptions, PS2 is the worst choice of the poststratified estimators for the first four characteristics regardless of the combination of variable, $f_{\text{max}}$, and collapsing method. When $f_{\text{max}} = 1.8$, the choice that leads to more collapsing, the MSEs of PS2 range from 1.8% to 44.2% larger than those of PS1. The MSEs of both PS.WR1 and PS.WR2 are near those of PS1 with the exception of (hospital stay, $f_{\text{max}} = 1.8$, adjacency) where the 6.3% bias of PS.WR2 leads to an MSE 25.6% larger than that of PS1. Close-mean collapsing is preferable to adjacency collapsing, although for the first four characteristics none of the estimators have smaller MSEs than PS1, which does not use collapsing.

The estimators again perform differently for the Common mean $Y$ variable. The MSEs of Hájek, PS2, PS.WR1, and PS.WR2 are all less than that of PS1. The Hájek estimator has the smallest MSE, owing to the fact that poststratification is unnecessary to correct bias in estimating the mean.

<table>
<thead>
<tr>
<th>Collapsing Method</th>
<th>Adjustment Bound</th>
<th>Hájek (standard collapsing)</th>
<th>PS2 (truncated weights then collapse)</th>
<th>PS.WR1 (fixed maximum weight adjustment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjacency</td>
<td>2</td>
<td>0.257</td>
<td>0.120</td>
<td>0.086</td>
</tr>
<tr>
<td>Close mean</td>
<td>2</td>
<td>0.257</td>
<td>0.080</td>
<td>0.127</td>
</tr>
<tr>
<td>Adjacency</td>
<td>1.8</td>
<td>0.256</td>
<td>0.150</td>
<td>0.085</td>
</tr>
<tr>
<td>Close mean</td>
<td>1.8</td>
<td>0.256</td>
<td>0.101</td>
<td>0.109</td>
</tr>
</tbody>
</table>

|         | Adjacency | 2                | 0.442                                | 0.326                                    |
|         | Close mean | 2                | 0.441                                | 0.321                                    |
|         | Adjacency | 1.8              | 0.442                                | 0.330                                    |
|         | Close mean | 1.8              | 0.442                                | 0.337                                    |

Table 4 Average absolute coverage correction error as defined in expression (10)

Statistics Canada, Catalogue No. 12-001-X
Table 5  Relative biases (in percent) of estimated proportions. (Figures for Hájek and PS1 are not affected by collapsing and are repeated in the four sections of the table to facilitate comparisons)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Range of no. of poststrata after collapsing</th>
<th>Hájek (no collapsing)</th>
<th>PS1 (standard collapsing)</th>
<th>PS2 (truncature weights then collapse)</th>
<th>PS.WR1 (fixed maximum weight adjustment)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Hájek</td>
<td>PS2</td>
<td>PS.WR1</td>
<td>PS.WR2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>(10, 16)</td>
<td>-11.5</td>
<td>0.1</td>
<td>-4.4</td>
<td>1.0</td>
</tr>
<tr>
<td>Limitations</td>
<td>(8, 15)</td>
<td>-12.1</td>
<td>-0.3</td>
<td>-2.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Delayed care</td>
<td>(6, 14)</td>
<td>8.2</td>
<td>-0.2</td>
<td>2.2</td>
<td>-0.6</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>(9, 16)</td>
<td>13.4</td>
<td>0.2</td>
<td>6.2</td>
<td>-0.7</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>(5, 11)</td>
<td>0.3</td>
<td>0</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>(10, 16)</td>
<td>-11.5</td>
<td>0.1</td>
<td>-0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Limitations</td>
<td>(8, 15)</td>
<td>-12.1</td>
<td>-0.3</td>
<td>-1.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Delayed care</td>
<td>(6, 14)</td>
<td>8.2</td>
<td>-0.2</td>
<td>-0.3</td>
<td>-0.3</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>(9, 16)</td>
<td>13.4</td>
<td>0.2</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>(5, 11)</td>
<td>0.3</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>(7, 13)</td>
<td>-11.5</td>
<td>0.1</td>
<td>-6.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Limitations</td>
<td>(7, 12)</td>
<td>-12.1</td>
<td>-0.3</td>
<td>-3.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Delayed care</td>
<td>(5, 11)</td>
<td>8.2</td>
<td>-0.2</td>
<td>3.5</td>
<td>-0.4</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>(5, 12)</td>
<td>13.4</td>
<td>0.2</td>
<td>9.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>(5, 9)</td>
<td>0.3</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>(6, 13)</td>
<td>-11.5</td>
<td>0.1</td>
<td>-1.6</td>
<td>0.3</td>
</tr>
<tr>
<td>Limitations</td>
<td>(7, 12)</td>
<td>-12.1</td>
<td>-0.3</td>
<td>-2.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Delayed care</td>
<td>(5, 10)</td>
<td>8.2</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>(5, 12)</td>
<td>13.4</td>
<td>0.2</td>
<td>1.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>(5, 10)</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 6  Ratio of variances (or MSES) to the variance (or MSE) of the poststratified estimator (PS1) with no collapsing. (Figures for Hájek are repeated in the four sections of the table to facilitate comparisons)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Ratio of variances to the variance of the poststratified estimator (PS1)</th>
<th>Ratio of MSES to the MSE of the poststratified estimator (PS1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hájek</td>
<td>PS2 (standard collapsing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>0.877</td>
<td>1.014</td>
</tr>
<tr>
<td>Limitations</td>
<td>0.821</td>
<td>0.966</td>
</tr>
<tr>
<td>Delayed care</td>
<td>1.099</td>
<td>1.023</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>1.290</td>
<td>1.169</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>0.755</td>
<td>0.805</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>0.877</td>
<td>1.013</td>
</tr>
<tr>
<td>Limitations</td>
<td>0.821</td>
<td>0.999</td>
</tr>
<tr>
<td>Delayed care</td>
<td>1.099</td>
<td>0.997</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>1.290</td>
<td>1.011</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>0.776</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>0.877</td>
<td>0.960</td>
</tr>
<tr>
<td>Limitations</td>
<td>0.821</td>
<td>0.939</td>
</tr>
<tr>
<td>Delayed care</td>
<td>1.099</td>
<td>1.051</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>1.290</td>
<td>1.225</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>0.780</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance</td>
<td>0.877</td>
<td>1.010</td>
</tr>
<tr>
<td>Limitations</td>
<td>0.821</td>
<td>0.983</td>
</tr>
<tr>
<td>Delayed care</td>
<td>1.099</td>
<td>1.003</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>1.290</td>
<td>1.052</td>
</tr>
<tr>
<td>Common mean Y</td>
<td>0.771</td>
<td>0.924</td>
</tr>
</tbody>
</table>
Table 7 reports the empirical coverages of 95% CI's computed using the estimated proportions and the linearization variance estimator that naturally accompanies each. A $t$-distribution with 25 degrees of freedom is used in all cases. The Hájek coverage rates are extremely poor, as expected, ranging from 70.9% to 92% for the first four characteristics. The poststratified estimators, PS1 and PS.WR1 provide 93.8% to 94.7% coverage, i.e., near the nominal 95%. In contrast, PS2 coverage is somewhat poor for Health insurance and hospitalization, especially for (adjacency, $f_{\text{max}} = 1.8$) where the coverages are 87.5% and 88.4%. Coverage rates for PS.WR2 are slightly less than for PS.WR1 but are reasonably close to nominal. Use of close-mean collapsing generally improves the cases of poor coverage found with adjacency. For Common Mean $Y$ coverages are good, ranging from 92.2% to 94.9%.

In summary, the weight-restricted estimators, PS.WR1 and PS.WR2, have some advantage over the standard collapsing estimator, PS2. They are generally less biased and retain more of the undercoverage adjustment than does PS2. However, the most critical element in bias-control is how the cells are collapsed in the first place. Collapsing using nearness of cell means or coverage rates is far more preferable than collapsing using some adjacency criterion based on neither of these. Only when cell means were equal did we observe any gain in MSE from collapsing cells. However, equality of cell means is the exception in practice.

5. Concluding remarks

Designers of surveys of households or establishments often have a lengthy list of poststrata or cells in mind when they develop weighting systems. If the sample size in a poststratum is small or the sample estimate of the population count in a poststratum is much different from an external control count, the poststratum may be collapsed with an adjacent one. The conventional justification for collapsing is that the possibility of creating extreme weights is reduced as are variances of estimates.

However, a poor choice of the method for collapsing has at least two undesirable consequences: (i) deficient frame or sample coverage in some cells is not completely corrected and (ii) estimates from the standard approach to collapsing may be quite biased. The latter problem can result in confidence intervals that cover at much less than the nominal rate. Collapsing leads to bias when coverage rates, cell means, or both are correlated within a collapsed poststratum. The bias can be either positive or negative, depending on the correlation.

Cells should be collapsed based on similarity of coverage rates, population cell means, or both in order to avoid bias. This method of collapsing can be much different from standard procedures that only collapse “adjacent” cells, e.g., by combining contiguous age groups. If the adjacency coincides with cells that have similar coverage rates or
means, no bias results. But, this should be checked rather than assumed.

There are at least two practical issues with collapsing based on cell means. One is that, while the theory directs us to collapse based on population means, in a particular sample we will only have estimates for the population covered by the frame. Coverage may be so deficient that the means of the covered and non-covered parts of the population are substantially different, even within the initial poststrata. This would be a case of “nonignorable non-coverage.” If so, poststratification based only on the initial set of cells or combinations of them cannot correct coverage bias. A second practical issue is that data on many items are collected in most surveys. Collapsing based on the cell means for one variable may not work well for other variables. In that case, the compromise, suggested by Little and Vartivarian (2005) for nonresponse adjustment, of collapsing based on some weighted average of the means of an important set of variables should be a good solution.

Extensions of this research would be to examine the performance of the class of calibration estimators in correcting coverage errors. Poststratification is a special case. When categories of qualitative auxiliaries are combined due to small sample sizes or other reasons, the same bias problems we have illustrated here may be introduced in more general calibration estimators. One method of allowing some flexibility to depart from controls while retaining important auxiliaries is already available in Rao and Singh (1997). The effect of their proposals on coverage bias needs to be investigated.

Acknowledgements

The authors are indebted to the Associate Editor and the referees for their careful reviews and constructive comments. This paper reports the general results of research undertaken in part by National Center for Health Statistics (NCHS) staff. The views expressed are attributable to the authors and do not necessarily reflect those of the NCHS. The work of R. Valliant was partially supported by Professional Services Contracts 200-2004-M-09302 and 200-2006-M-17916 between the University of Michigan and the National Center for Health Statistics.

References


Kim, Li and Valliant: Cell collapsing in poststratification


A single frame multiplicity estimator for multiple frame surveys

Fulvia Mecatti

Abstract

Multiple Frame Surveys were originally proposed to foster cost savings on the basis of an optimality approach. As surveys on special, rare and difficult-to-sample populations are becoming more prominent, a single list of population units to be used as a sampling frame is often unavailable in sampling practice. In recent literature multiple frame designs have been put forward in order to increase population coverage, to improve response rates and to capture differences and subgroups. Alternative approaches to multiple frame estimation have appeared, all of them relying upon the virtual partition of the set of the available overlapping frames into disjointed domains. Hence the correct classification of sampled units into the domains is required for practical applications. In this paper a multiple frame estimator is proposed using a multiplicity approach. Multiplicity estimators require less information about unit domain membership hence they are insensitive to misclassification. Moreover the proposed estimator is analytically simple so that it is easy to implement and its exact variance is given. Empirical results from an extensive simulation study comparing the multiplicity estimator with major competitors are also provided.

Key Words: Difficult-to-Sample populations; Dual frame survey; Misclassification; Raking ratio; Variance estimation.

1. Introduction

In classic finite population sampling a basic hypothesis is the availability of a unique and complete list of units forming the target population to be used as a sampling frame. In some cases a set of two or more lists is available for survey purposes. The general case of \( Q \geq 2 \) lists, singularly partial and possibly overlapping, is known as Multiple Frame Survey. Multiple frame surveys were originally introduced (Hartley 1974) as a device for reducing survey costs by achieving the same precision as a customary unique-frame survey. In modern sampling practice, as surveys of special, rare and difficult-to-sample populations are becoming more common (Kalton and Anderson 1986; Sudman and Kalton 1986; Sudman, Sirken and Cowan 1988) it is often the case that a unique list of units does not exist and the population size \( N \) is an unknown parameter to be estimated. Recent literature considers multiple frame surveys with the main aim of increasing population coverage, of improving response rates and of capturing differences and subgroups more accurately (Iachan and Dennis 1993; Carlson and Hall 1994; Haines and Pollock 1998; Eurostat 2000). In a recent paper Lohr and Rao (2006) stated: “As the U.S., Canada, and other nations grow in diversity, different sampling frames may better capture subgroups of the population. […] We anticipate that modular sampling designs using multiple frames will be widely used in the future”. A contemporary application could be found in web surveys: the population coverage can be improved and the bias due to the features of the site used for data collection can be reduced by using two or more independent web sites simultaneously. Since the same unit can visit more than one site involved in the survey, the sites overlap configuring a multiple frame framework.

Estimation in multiple frame surveys, as first developed by Hartley (1962, 1974), is based on the virtual partition of the population \( i.e. \), the unknown union of the \( Q \) overlapping frame) into \( 2^Q - 1 \) disjointed domains \( i.e. \), the mutually exclusive intersections of frames). Hence the total \( Y \) of a study variable \( y \), taken as the parameter to be estimated, is expressed as a sum of domain totals. Sample data from the \( Q \) frames are used to produce estimates for the domain totals. Estimated domain totals are finally combined to provide estimation for the population total \( \bar{Y} \). A number of estimators have been developed according to alternative approaches to multiple frame estimation (see Section 2). Since all estimators appearing in literature rely on the partition into the domains as mentioned above, the correct identification of the domain membership of each sampled unit is required for their practical application. This is a strong assumption that may not always be true in practice, as argued for instance in Lohr and Rao (2006). Indeed this implies that every sampled unit should be questioned on both the survey value and on its membership to each frame involved in the survey, in order to be able to correctly classify them into the domains. In addition to the natural risk of misclassification there might also be a risk connected with confidentiality and with the sensitivity of units to the frame membership which could both increase the rate of non-response and affect the estimator precision. This situation could apply, for instance, when surveying sensitive characteristics (private behaviours, addictions …) or when sampling elusive populations (illegal immigrants,  

1. Fulvia Mecatti, Department of Statistics, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8. Ed. U7, 20126 Milan, Italy. E-mail: fulvia.mecatti@unimib.it.
ex-prisoners, patients...). In the present paper a different approach to estimation in multiple frame surveys is adopted. The concept of unit *multiplicity*, corresponding to the *number* of frames to which units belong, is proposed in alternative to the existing approaches based on the domain membership, *i.e.*, to which frames units belong. An unbiased estimator, naturally insensitive to domain misclassification and applying to any number of frames, is presented. The proposed multiplicity estimator has a simple analytical structure so that it can be easily implemented, while its exact variance is given in a closed form and hence readily estimated for any sample size.

In Section 2 an overall discussion of the main contributions to multiple frame estimation is presented in a unified view and the necessary notation is introduced. In Section 3 a multiplicity estimator is proposed and variance estimation is analysed. An extensive simulation study comparing the proposed estimator with major competitors is presented in Section 4.

**2. Optimum, pseudo-optimum and single frame estimation**

Although literature has mostly dealt with the dual frame case (*Q* = 2), a general theoretical framework for multiple frame surveys (*Q* ≥ 2) has been recently provided in Lohr and Rao (2006). By using their multiple frame notation, different estimation approaches are briefly reviewed and the available estimators are presented in a unified way which highlights their dependency on the domain membership of the sampled units.

Let \( A_1 \cdot A_y \cdot A_Q \) be a collection of \( Q \geq 2 \) overlapping frames, the union of which offers a population coverage adequate for survey objectives. Let the index sets \( K \) be the subsets of the range of the frame index \( q = 1 \cdot Q \). For every index set \( K \subseteq \{1 \cdot q \cdot Q\} \) a domain is defined as the set \( D_K = (\bigcap_{q \in K} A_q) \cap (\bigcap_{q \not\in K} A_q^C) \), where \( C \) denotes complementation.

Let the domain membership indicator be the indicator \( \delta_i(K) \) taking value 1 if unit \( i \) is included in domain \( D_K \) and 0 otherwise. The estimating total \( Y \) over the (unknown) union of the \( Q \) overlapping frames is then expressed as a sum over the set of \( 2^Q - 1 \) disjointed domains

\[
Y = \sum_{i \in \bigcup_A A_q} y_i = \sum_{K} \sum_{i \in D_K} \delta_i(K) y_i. \tag{1}
\]

Let \( s_q \) be a sample selected from frame \( A_q \) under a given design, independently for \( q = 1 \cdot Q \). A general expression of a multiple frame estimator based on the domain classification is then

\[
\hat{Y} = \sum_k \sum_{q \in K} \sum_{i \in D_q} w_i^{(q)} \delta_i(K) y_i. \tag{2}
\]

Note that when a unbiased estimator for the total \( Y \) is given, an estimator for the population size \( N \) is also given by simply substituting sample values \( y_i \) by 1's.

Estimators available in literature result from setting weights \( w_i^{(q)} \) in (2) according to three main approaches. Since multiple frame surveys were originally put forward with the aim of fostering cost savings by achieving equal or greater precision than a customary unique-frame survey, an *optimum* approach was first suggested by using optimum weights \( w_i^{(q)} \mid opt \) in (2), *i.e.*, by minimizing the estimator variance (Hartley 1962, 1974; Lund 1968; Fuller and Burmeister 1972). Optimum estimators have optimal theoretical properties (Skinner 1991; Lohr and Rao 2000) but present practical problems due mainly to their complexity (explicit though complex formulae for optimum weights \( w_i^{(q)} \mid opt \) with any number of frames are given in Lohr and Rao 2006, Section 3). Moreover, optimum weights depend on unknown population covariances so that they must be estimated from sample data. This is both computationally complex and affects optimality since the extra variability in estimating the covariances leads to larger mean square errors (Lohr and Rao 2006, Section 7).

In order to improve the applicability, a *single frame* (SF) approach has been proposed by using *fixed weights* which ensure design-unbiasedness. For simple random sampling in every frame, the SF estimator is given by substituting weights \( w_i^{(q)} \) in (2) with \( w_i^{(q)} \mid SF = w_i^{(k)} = (\sum_{q \in K} f_q) ^{-1} \) where \( f_q = n_q / N_q \) denotes the frame sampling fraction (Bankier 1986; Kalton and Anderson 1986; Skinner 1991; Skinner, Holmes and Holt 1994). Since fixed weights usually differ from optimum weights, the SF estimator is generally less efficient than an optimum estimator (Lohr and Rao 2000). Finally a *pseudo-optimum approach* was proposed (Skinner and Rao 1996; Lohr and Rao 2000) in order to achieve both a wider applicability than optimum estimators and to improve efficiency compared with the SF approach. A pseudo-maximum likelihood (PML) estimator for multiple frame surveys is given by substituting in (2): \( w_i^{(q)} \mid PML = w_i^{(k)} = \hat{N}_k / \sum_{q \in K} \sum_{i \in D_q} \delta_i(K) = \hat{N}_k / n_k \) where the estimated domain sizes \( \hat{N}_k \) are the solution of a system of non linear equations. Although complex to implement for practical applications (an iterative linear approximation of \( \hat{N}_k \) under simple random sampling is given in Lohr and Rao 2006, Section 4.1) the PML estimator retains good theoretical properties from the optimum approach.

Note that formula (2) involves the domain membership indicator \( \delta_i(K) \); hence optimum, pseudo-optimum and SF estimators apply only if the correct classification of sample data into the \( 2^Q - 1 \) domains is accomplished.
In the next Section a multiple frame estimator is presented on the basis of a single frame multiplicity approach which does not require domain classification.

3. The single frame multiplicity estimator

The notion of multiplicity was first introduced in connection with Network Sampling (Casady and Sirken 1980; Sirken 2004). It is also a tool of the Generalized Weight Share Method (Lavallée 2002; 2007) as well as of the Center Sampling estimation theory (Mecatti 2004) since center sampling and multiple frame surveys are equivalent under certain conditions. In Lohr and Rao (2006), the multiplicity of domain \( D_k \) is defined as the cardinality of the index set \( K \). Since domains are mutually exclusive, multiplicity is also a characteristic of every population unit, being the number of frames in which each unit is included among the \( Q \) involved in the survey.

Let \( m_i \) be the multiplicity of unit \( i \). Note that unit multiplicity may be collected simply by asking sampled units how many frames they belong to.

Since clearly \( \sum_{q\in A_i} y_i = \sum_{i\in s_q} m_i y_i \), it follows that

\[
Y = \sum_{q=1}^{Q} \sum_{i\in s_q} y_i m_i^{-1}. \tag{3}
\]

Notice that expression (3), which involves exclusively sums over the frames, represents a practical advantage with respect to equation (1). In fact the domains provide a virtual (unknown) partition of the population while the sample selection is actually performed in the \( Q \) overlapping frames. This leads to a SF multiplicity estimator as given by

\[
\hat{Y}_M = \sum_{q=1}^{Q} \sum_{i\in s_q} w_i^{(q)} y_i m_i^{-1} \tag{4}
\]

with fixed weights \( w_i^{(q)} \) ensuring, for instance, design-unbiasedness. For simple random sampling of every frame we have \( w_i^{(q)} = f_q^{-1}, \forall i \in s_q \).

Unlike the optimum, PML and SF estimators discussed in Section 2, estimator (4) does not involve the sample membership indicator and it is very simple to implement in practical applications. Furthermore, it is to be noted that for simple random sampling of every frame, the sampled values in multiplicity estimator (4) are weighted by \( (f_q m_q)^{-1} \), i.e., by a specific frame coefficient; vice versa, in the SF estimator sampled values are weighted by \( w_i^{(k)} = (\sum_{q\in k} f_q)^{-1} \), i.e., by an average coefficient over the frames involved in each domain. As a consequence \( \hat{Y}_M \) is expected to be more accurate than the SF estimator, as confirmed by simulation results. Moreover, owing to its Horvitz-Thompson structure, the exact variance of \( \hat{Y}_M \) can be derived in closed form. For simple random sampling of every frame the estimator variance is given by

\[
V(\hat{Y}_M) = \sum_{q=1}^{Q} \sum_{i\in s_q} \left[ N_q \sum_{i\in s_q} y_i^2 m_i^{-2} - \left( \sum_{i\in s_q} y_i m_i^{-1} \right)^2 \right]. \tag{5}
\]

An unbiased variance estimator for simple random sampling of every frame is then

\[
\hat{V}(\hat{Y}_M) = \sum_{q=1}^{Q} \frac{N_q (N_q - n_q)}{n_q^2 (N_q - 1)} \left[ N_q \sum_{i\in s_q} y_i^2 m_i^{-2} - f_q^{-1} \left( \sum_{i\in s_q} y_i m_i^{-1} \right)^2 \right]. \tag{6}
\]

The performance of the multiplicity estimator for finite sample sizes has been empirically studied under simple random sampling and compared to major competitors in a simulation study.

4. Simulation study

Several simulation results concerning dual frame estimators have appeared in literature (Bankier 1986; Skinner and Rao 1996; Lohr and Rao 2000). In the general case of \( Q \geq 2 \) frames, Lohr and Rao (2006) extensively investigated the empirical mean squared errors of a set of eight estimators under optimum, pseudo-optimum and single frame approaches, in a three-frame framework under a two-stage design. Their results suggest that optimum estimators are theoretically optimal but in practice the extra variability in estimating optimum weights leads to larger mean squared errors. Hence the PML estimator appears as the best performer in terms of empirical relative efficiency. Furthermore, their study regarded a case of about 10% of sampled units misclassified into the domains and more research on the effects of misclassification on the estimator performances is recommended.

In the present study pseudo-optimum and single frame estimators are compared with the multiplicity estimator (4), with three main objectives:

i) to investigate empirical conditions in which the multiplicity estimator results more efficient than the SF estimator (Section 4.2);

ii) to consider the raking ratio correction to known frame sizes \( N_q \) as already proposed in order to improve efficiency of the SF estimator (Section 4.3);

iii) to explore the effects of increasing rates of misclassification upon the empirical properties of the PML and SF estimators (simple and raked).
versus the natural insensitivity of the multiplicity estimator (Section 4.4).

4.1 Implementation

The simulation study was performed in an artificial three-frame setup and implemented as follows. $N$ population pseudo-values $y_i$ are generated from a Gamma distribution. Some preliminary simulations indicated that both increasing values of the population size $N$ and different values for the Gamma parameters (leading to an asymmetrical and almost symmetrical shape) do not produce significant differences in the pattern of the relative performance of the estimators considered. The study was then conducted by setting $N = 1,200$ and by generating from a Gamma distribution with parameters of 1.5 and 2. Every pseudo-value $y_i$ is randomly assigned to the $Q = 3$ frames according to 3 independent Bernoulli trials with probability $\alpha_q = N_q / N$, $q = 1, 2, 3$. Different scenarios regarding both frame coverage and frame overlapping result from different choices for the $\alpha_q$, namely the total rate of population units classified into the four overlapping domains.

Chosen a set of sampling fractions $f_q = n_q / N_q$, $q = 1, 2, 3$, a simple random sampling is selected independently from every frame, iteratively for 10,000 simulation runs. For a given estimator, say $\hat{Y}$, the collection of values $\{\hat{Y}_p, p = 1...10,000\}$ is assumed as its monte carlo distribution and the empirical mean $E_{mc}(\hat{Y}) = \sum_p \hat{Y}_p / 10,000$ and the empirical mean squared error $\text{MSE}_{mc}(\hat{Y}) = \sum_p (\hat{Y}_p - \hat{Y})^2 / 10,000$ are calculated. The monaco carlo error is controlled by only accepting simulations giving empirical relative bias $\text{RB}_{mc}(\hat{Y}) = 100 \cdot |E_{mc}(\hat{Y}) - \hat{Y} / \hat{Y}|$ less than 1.5% for those estimators known to be unbiased. Furthermore, by using the exact variance of the multiplicity estimator as given by (6), simulations ensure $|\text{MSE}_{mc}(\hat{Y}_M) - V(\hat{Y}_M)| \leq 0.03$. Several different scenarios have been investigated by combining different levels of frame coverage, of frame overlapping and of sampling disproportion, leading to 29 simulated populations. In Figure 1 the simulated populations are represented as points in the plane formed by the two main simulation parameters, namely the total frame coverage on the horizontal axis (as given by $\sum_q \alpha_q$) and the sampling disproportion on the vertical axis, i.e., the dispersion among the sampling fractions $f_q$ as measured by $\sum_q |f_q - f_q| / 3^2$. The different shape of populations points in Figure 1 indicates different levels of overlapping, namely the total rate of population units classified into the four overlapping domains.

![Figure 1 Simulated populations](image)

4.2 Multiplicity versus simple single frame estimation

As noted in Section 3, the multiplicity estimator involves specific frame weights whereas the SF estimator is based on average coefficients. As a consequence the two estimators coincide for constant sample fraction $f_q = f$ in every frame, i.e., for proportionate sampling, and they offer different estimates for disproportionate sampling. Simulation results provide empirical evidence that the multiplicity estimator is more accurate than the simple SF estimator. Estimator $\hat{Y}_M$ is shown to be more efficient in all the cases explored except in one extreme case in which the three frames are almost complete and hence the total overlapping is close to 100%. Neglecting this single case, efficiency gains of $\hat{Y}_M$ over the SF estimator, as measured by a customary empirical efficiency ratio (see Table 1), range from 5% to 48%, and are never less than 26% in half of the simulations. Efficiency of the multiplicity estimator over the SF estimator increases as the sampling disproportion increases (see Table 2) whereas it has resulted as being essentially independent with respect to increasing levels of frame coverage and overlapping.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Empirical efficiency ratio of $\hat{Y}_M$ versus SF estimator: Elementary statistics over 28 simulated populations</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>Max</td>
</tr>
<tr>
<td>0.7425</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Empirical efficiency ratio of $\hat{Y}_M$ versus SF estimator for increasing levels of sampling disproportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Disproportion</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Empirical efficiency ratio averaged for different levels of frames coverage/overlapping 0.92 0.81 0.68 0.57
4.3 Raking ratio adjustment

It has been suggested that the raking ratio adjustment using the known frame sizes $N_q$ (Bankier 1986) be used in order to improve efficiency of the simple SF estimator. Theoretical and empirical results have already appeared in literature confirm that the raking ratio SF estimator (SFrac) can be considerably more efficient than the simple SF estimator (Skinner 1991; Lohr and Rao 2000, 2006; Mecatti 2005).

In order to adjust the multiplicity estimator via raking ratio, knowledge of the domain membership of sampled units has to be assumed. By using this additional, though redundant, information $\hat{Y}_M$ may be rewritten as

$$\hat{Y}_M = \sum_{K} \sum_{q \in K} \left( |K| f_q \right)^{-1} \sum_{i \in q} \delta_i(K) y_i$$  \hspace{1cm} (7)

where $|K|$ indicates the number of frames involved in domain $D_K$ and it equals unit multiplicity $m_i$ for all $i \in D_K$. Setting the initial weights at $h_{Kq}^{(0)} = (|K| f_q)^{-1}$, the $t$th iteration of the raking ratio multiplicity estimator (Mrak) is obtained by substituting the following raked weights in (7)

$$h_{Kq}^{(t)} = \begin{cases} \frac{N_q h_{Kq}^{(t-1)}}{\sum_{q \in K} h_{Kq}^{(t-1)} n_q} & \text{if } q \in K \\ h_{Kq}^{(t-1)} & \text{if } q \notin K \end{cases}$$

where $q = Q$ if $t$ is a multiple of $Q$ otherwise $q = t \mod(Q)$, for $t = 1, 2, \ldots$ until convergence.

Simulations regarded different levels of frames coverage combined with different sets of sampling fractions, leading to increasing sampling disproportion.

Empirical results show that Mrak is more efficient than SFrac in 38% of cases explored and it is equally or less efficient in the remaining cases. Efficiency gains range from 3% to 74% and occur for low levels of frame coverage. For increasing frame coverage (and hence increasing overlapping) Mrak estimator is superior to SFrac estimator for high sampling disproportion only. In the other cases, namely for increasing frame coverage/overlapping combined with low to medium sampling disproportion, Mrak can be considerably less efficient than SFrac (see Table 3 for the ten indicative cases) and also severely biased. Thus empirical results suggest that the raking ratio adjustment has better effects under a single frame approach than under a multiplicity approach, although there are conditions in which the latter is still superior. With this respect more research is needed. Particularly, since the raking ratio procedure is in fact a special case of calibration (Deville and Särndal 1992; Deville, Särndal and Sautory 1993), potential improvements might follow by applying the more general calibration to estimator $\hat{Y}_M$. Calibration of the multiplicity estimator, as viewed as a particular case of the Generalized Weight Share Method, is outlined in Lavallée (2002, 2007).

### Table 3 Efficiency of Mrak versus SFrak: Ten indicative simulation runs

<table>
<thead>
<tr>
<th>Frame coverage $\alpha_q = N_q/N$</th>
<th>Sampling fractions $f_q = n_q/N_q$</th>
<th>Empirical efficiency ratio Mrak versus SFrak</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.60 0.60 0.60 0.01 0.95 0.15</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>0.35 0.35 0.70 0.80 0.50</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>0.85 0.85 0.01 0.95 0.15</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>0.35 0.40 0.50 0.70 0.80 0.60</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>0.60 0.60 0.60 0.70 0.80 0.60</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>0.80 0.50 0.35 0.01 0.95 0.15</td>
<td>1.22</td>
<td></td>
</tr>
<tr>
<td>0.35 0.40 0.50 0.01 0.95 0.15</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>0.70 0.05 0.95 0.70 0.80 0.60</td>
<td>2.09</td>
<td></td>
</tr>
<tr>
<td>0.70 0.05 0.95 0.80 0.20 0.50</td>
<td>5.79</td>
<td></td>
</tr>
</tbody>
</table>

4.4 Misclassification

The aim of the final part of the simulation study is to investigate the sensitivity of the pseudo-optimum (PML) and single frame estimators (simple and raked) to increasing levels of misclassification of sampled units into the domains, with respect to the structural insensitivity of the proposed multiplicity estimator. For a chosen rate of misclassification, the desired number of sampled units to be inexactely classified is taken from the domain with the largest size and randomly assigned to the remaining domains, independently for each frame.

Tables 4 and 5 show elementary statistics summarizing simulation results in the case of exact classification and in the case of slight misclassification equal to 1% of sampled units. Note that for exact classification all the estimators appear unbiased (or nearly unbiased). As regards efficiency, according to other simulation results (Lohr and Rao 2006) SFrak and PML estimators show similar performances. As expected, for exact classification they are more efficient than $\hat{Y}_M$ in all the cases explored (except for two isolated cases) as a consequence of the different amount of information used in the estimation process. However the SF (simple and raked) and PML estimators tend to become biased and less efficient than $\hat{Y}_M$ in presence of just a small amount of misclassification.

### Table 4 Relative bias in case of 1% of misclassification: Elementary statistics over the 29 simulated populations

<table>
<thead>
<tr>
<th>RB_{mc} 1% of sampled units misclassified</th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>75th quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Y}_M$</td>
<td>SF</td>
<td>SFrak</td>
<td>PML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>2.5880</td>
<td>1.7632</td>
<td>2.7352</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.83</td>
<td>0.80</td>
<td>0.73</td>
<td>0.23</td>
<td>4.67</td>
<td>3.46</td>
</tr>
<tr>
<td>0.70</td>
<td>2.65</td>
<td>1.97</td>
<td>3.46</td>
<td>2.13</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.97</td>
<td>2.87</td>
</tr>
</tbody>
</table>

Statistics Canada, Catalogue No. 12-001-X
Finally we focused on the case of maximum efficiency of SF, SFrak and PML over \( \hat{Y}_M \) for exact classification, namely the case of high frame overlapping/coverage and low sampling disproportion. In this set up, increasing rates of misclassification of sampled units into domains (from 0 to 50%) were investigated. Table 6 and 7 show respectively the relative bias and the efficiency ratio of \( \hat{Y}_M \) versus SF, SFrak and PML estimators, for increasing levels of misclassification. It is to be noticed that although the negative effects of misclassification are rapid and severe for all the competitors, the PML estimator emerges as the least affected.

As a conclusion the proposed multiplicity estimator, besides being simple, is recommended when the risk of (even slight) misclassification of sampled units into the domains is a concrete possibility.

### Table 5
Empirical efficiency ratio of \( \hat{Y}_M \) versus SF and PML estimators: Elementary statistics over the 29 simulated populations for exact classification and for slight misclassification

<table>
<thead>
<tr>
<th>Empirical efficiency ratio</th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>75th quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFrak</td>
<td>1.43</td>
<td>0.69</td>
<td>3.21</td>
<td>1.51</td>
<td>1.28</td>
</tr>
<tr>
<td>PML</td>
<td>1.41</td>
<td>0.72</td>
<td>3.30</td>
<td>1.47</td>
<td>1.25</td>
</tr>
<tr>
<td>1% misclassification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF</td>
<td>0.39</td>
<td>0.13</td>
<td>0.71</td>
<td>0.54</td>
<td>0.34</td>
</tr>
<tr>
<td>SFrak</td>
<td>0.78</td>
<td>0.13</td>
<td>1.98</td>
<td>0.95</td>
<td>0.74</td>
</tr>
<tr>
<td>PML</td>
<td>0.77</td>
<td>0.14</td>
<td>1.94</td>
<td>0.98</td>
<td>0.70</td>
</tr>
</tbody>
</table>

### Table 6
(absolute) Relative bias for increasing rate of misclassification

<table>
<thead>
<tr>
<th>% misclassification</th>
<th>( \hat{Y}_M )</th>
<th>SF</th>
<th>SFrak</th>
<th>PML</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>( \approx 0 )</td>
</tr>
<tr>
<td>1%</td>
<td>0</td>
<td>2.57</td>
<td>1.38</td>
<td>4.3</td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
<td>13.57</td>
<td>7.15</td>
<td>2.75</td>
</tr>
<tr>
<td>10%</td>
<td>0</td>
<td>17.80</td>
<td>14.14</td>
<td>4.56</td>
</tr>
<tr>
<td>20%</td>
<td>0</td>
<td>25</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>50%</td>
<td>0</td>
<td>144</td>
<td>68</td>
<td>39</td>
</tr>
</tbody>
</table>

### Table 7
Empirical efficiency ratio of \( \hat{Y}_M \) versus SF, SFrak and PML estimators for increasing rate of misclassification

<table>
<thead>
<tr>
<th>% misclassification</th>
<th>( \hat{Y}_M ) versus SF</th>
<th>( \hat{Y}_M ) versus SFrak</th>
<th>( \hat{Y}_M ) versus PML</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.640</td>
<td>3.210</td>
<td>3.300</td>
</tr>
<tr>
<td>1%</td>
<td>0.260</td>
<td>1.040</td>
<td>1.100</td>
</tr>
<tr>
<td>5%</td>
<td>0.020</td>
<td>0.060</td>
<td>0.370</td>
</tr>
<tr>
<td>10%</td>
<td>0.010</td>
<td>0.020</td>
<td>0.150</td>
</tr>
<tr>
<td>20%</td>
<td>0.004</td>
<td>0.004</td>
<td>0.080</td>
</tr>
<tr>
<td>50%</td>
<td>( \approx 0 )</td>
<td>0.001</td>
<td>0.006</td>
</tr>
</tbody>
</table>

**Acknowledgements**

This work was partially supported by a grant from the Italian Ministry of University and Research. The author wishes to thank Jon N.K. Rao for the useful discussion and resourceful advice. Thanks are also due to the editor, two associate editors and two anonymous referees for their constructive comments and suggestions.

**References**


Variance estimation for a ratio in the presence of imputed data

David Haziza

Abstract
In this paper, we study the problem of variance estimation for a ratio of two totals when marginal random hot deck imputation has been used to fill in missing data. We consider two approaches to inference. In the first approach, the validity of an imputation model is required. In the second approach, the validity of an imputation model is not required but response probabilities need to be estimated, in which case the validity of a nonresponse model is required. We derive variance estimators under two distinct frameworks: the customary two-phase framework and the reverse framework.

Key Words: Imputation model; Nonresponse model; Marginal random hot deck imputation; Reverse framework; Two-phase framework; Variance estimation.

1. Introduction

Variance estimation in the presence of imputed data for simple univariate parameter such as population totals and population means has been widely treated in recent years; see for example, Särndal (1992), Deville and Särndal (1994), Rao and Shao (1992), Rao (1996) and Shao and Steel (1999). In practice, it is often of interest to estimate the ratio of two population totals, \( R = Y / X \), where \( (Y, X) = \sum_{i \in U} (y_i, x_i) \), \( y \) and \( x \) denote two variables of interest potentially missing and \( U \) denotes the finite population (of size \( N \)) under study. Although variance estimation for a ratio in the presence of imputed data is a problem frequently encountered in practice (especially in business surveys), it has not been, to our knowledge, fully studied in the literature. In this paper, we consider the case Marginal Random Hot Deck (MRHD) imputation performed within the same set of imputation classes for both variables \( y \) and \( x \). In other words, to compensate for nonresponse, random hot deck imputation is performed separately for both variables within the same set of imputation classes. This situation occurs frequently in practice. For simplicity, we consider the case of a single imputation class. Extensions to multiple imputation classes are relatively straightforward for most derivations presented in this paper.

In this paper, we derive variance estimators that take sampling, nonresponse and imputation into account. Two distinct frameworks for variance estimation have been studied in the literature: (i) the customary two-phase framework (e.g., Särndal (1992)) and (ii) the reverse framework (e.g., Shao and Steel (1999)). In the two-phase framework, nonresponse is viewed as a second phase of selection. That is, a random sample is selected from the population according to a given sampling design. Then, given the selected sample, the set of respondents is generated according to the nonresponse mechanism. In the reverse framework, the order of sampling and response is reversed. That is, the population is first randomly divided into a population of respondents and a population of nonrespondents according to the nonresponse mechanism. Then, a random sample is selected from the population (containing respondents and nonrespondents) according to the sampling design. As we will see in section 4, the reverse framework facilitates the derivation of variance estimators but unlike the two-phase framework, it requires the additional assumption that the nonresponse mechanism does not depend on which sample is selected. This assumption is satisfied in many situations encountered in practice. For each framework, inference can be based either on an Imputation Model (IM) or a Nonresponse Model (NM). The IM approach requires the validity of an imputation model, whereas the NM approach requires the validity of a nonresponse model.

In section 2, we introduce notation, assumptions and the imputed estimator of a ratio under weighted MRHD imputation. The IM and NM approaches are then presented in sections 2.1 and 2.2. In section 2.3, the bias of the imputed estimator is discussed. In section 3, variance estimators are derived under the two-phase framework and the IM approach using the method proposed by Särndal (1992). We show that, under MRHD imputation, the naïve variance estimator (that treats the imputed values as observed values) generally overestimates the sampling variance when \( y \) and \( x \) are positively correlated. In section 4, we derive variance estimators under the reverse framework and both the IM and the NM approaches using the method proposed by Shao and Steel (1999). Finally, we conclude in section 5.
2. Notation and assumptions

Our goal is to estimate \( R \). We select a random sample, \( s \), of size \( n \), according to a given sampling design \( p(s) \). A complete-data estimator is given by

\[
\hat{R} = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}},
\]

(2.1)

where \( (\hat{Y}_{HT}, \hat{X}_{HT}) = \sum_{i \in s} w_i(Y_i, X_i) \) denote the Horvitz-Thompson estimators for \( Y \) and \( X \), respectively and \( w_i = 1/\pi_i \) denotes the sampling weight of unit \( i \), where \( \pi_i \) is its probability of inclusion in the sample. The estimator \( \hat{R} \) in (2.1) is asymptotically \( p \)-unbiased for \( R \), i.e., \( E_p(\hat{R}) = R \), where the subscript \( p \) denotes the expectation and variance with respect to the sampling design \( p(s) \).

Since \( \hat{R} \) is a nonlinear function of estimated totals, its exact design variance, \( V_p(\hat{R}) \), cannot be easily obtained. To overcome this problem, Taylor linearization is often applied in order to approximate the exact variance. An asymptotically \( p \)-unbiased estimator of the approximate variance of \( \hat{R} \) is given by

\[
\hat{V}_{SAM} = \sum_{i \in s} \sum_{j \in s} \Delta_y e_i e_j,
\]

(2.2)

where \( e_i = 1/\hat{X}_{HT}(Y_i - \hat{R}X_i) \), \( \Delta_y = (\pi_y - \pi_i \pi_j)/\pi_y \pi_i \pi_j \) and \( \pi_y \) is the joint selection probability of units \( i \) and \( j \). Note that \( \pi_y = \pi_i \). In the case of simple random sampling without replacement, the variance estimator (2.2) reduces to

\[
\hat{V}_{SAM} = \left(1 - \frac{n}{N}\right) \frac{1}{n^2} \left[ s_y^2 + \hat{R}^2 s_x^2 - 2 \hat{R} s_{xy} \right],
\]

(2.3)

where

\[
s_y^2 = \frac{1}{n-1} \sum_{i \in s} (Y_i - \bar{y})^2, \quad s_x^2 = \frac{1}{n-1} \sum_{i \in s} (X_i - \bar{x})^2
\]

and

\[
s_{xy} = \frac{1}{n-1} \sum_{i \in s} (X_i - \bar{x})(Y_i - \bar{y})
\]

with

\[
(\bar{y}, \bar{x}) = \frac{1}{n} \sum_{i \in s} (Y_i, X_i).
\]

We now turn to the case for which both variables \( x \) and \( y \) may be missing. Let \( a_i \) be the response indicator of unit \( i \) such that \( a_i = 1 \) if unit \( i \) responds to variable \( y \) and \( a_i = 0 \), otherwise. Similarly, let \( b_i \) be the response indicator of unit \( i \) such that \( b_i = 1 \) if unit \( i \) responds to variable \( x \) and \( b_i = 0 \), otherwise. Let \( s^{(3)}_x \) be the set of respondents to variable \( x \) of size \( r_x \) and \( s^{(3)}_y \) be the set of respondents to variable \( x \) of size \( r_x \). Also, let \( r_y \) be the number of respondents to both variables \( y \) and \( x \). Finally, let \( y_i^* \) and \( x_i^* \) denote the imputed values to replace the missing values \( y_i \) and \( x_i \), respectively. An imputed estimator of \( R \) is given by

\[
\hat{R}_i = \frac{\sum_{i \in s} w_i \hat{y}_i}{\sum_{i \in s} w_i \hat{x}_i},
\]

(2.4)

where \( \hat{y}_i = a_i y_i + (1 - a_i) y_i^* \) and \( \hat{x}_i = b_i x_i + (1 - b_i) x_i^* \). Under weighted MRHD imputation, to compensate for the missing value \( y_i^* \), a donor \( j \) is selected at random with replacement from \( s^{(3)}_y \) so that

\[
P(y_i^* = y_j) = \frac{w_j}{\sum_{l \in s} w_l a_l}.
\]

Similarly, to compensate for the missing value \( x_i^* \), a donor \( j \) is selected at random with replacement from \( s^{(3)}_x \) so that

\[
P(x_i^* = x_j) = \frac{w_j}{\sum_{l \in s} w_l b_l}.
\]

Note that, when both \( y_i \) and \( x_i \) are missing, \( j \) is generally not equal to \( k \) under weighted MRHD imputation.

Random hot-deck imputation within classes is widely used in practice because (i) it preserves the variability of the original data; and (ii) it leads to plausible values. The latter is particularly important in the case of categorical variables of interest. However, random hot-deck imputation within classes suffers from an additional component of variance due to the use of a random imputation mechanism. The main reason weighted MRHD imputation is used is that it leads to asymptotically unbiased estimator under the nonresponse model approach (see section 2.1) unlike unweighted MRHD imputation.

Let \( E_i(\cdot | s, s^{(3)}_x, s^{(3)}_y) \), \( V_i(\cdot | s, s^{(3)}_x, s^{(3)}_y) \) and \( \text{Cov}(\cdot, \cdot | s, s^{(3)}_x, s^{(3)}_y) \) denote the conditional expectation, the conditional variance and the conditional covariance operators with respect to the random imputation mechanism (here, weighted MRHD imputation). Using a first-order Taylor expansion, it can be shown that

\[
E_i(\hat{R}_i | s, s^{(3)}_x, s^{(3)}_y) \approx \frac{\bar{y}_x}{\bar{y}_y} \equiv \hat{R}_x,
\]

(2.5)

where

\[
\bar{y}_x = \frac{\sum_{i \in s} w_i a_i y_i}{\sum_{i \in s} w_i a_i},
\]

and

\[
\bar{x}_y = \frac{\sum_{i \in s} w_i b_i x_i}{\sum_{i \in s} w_i b_i},
\]
denote the weighted means of the respondents to variables \( y \) and \( x \), respectively. The approximation in (2.5) will be valid if the sample size within classes is sufficiently large, which we assume to be the case. Now, let
\[
S_{yr}^2 = \frac{1}{n_r} \sum_{i \in s_r} w_i (a_i - \bar{y}_i)^2
\]
and
\[
S_{xr}^2 = \frac{1}{n_r} \sum_{i \in s_r} w_i (b_i - \bar{x}_i)^2
\]
denote the variability of the \( y \)-values and the \( x \)-values in the set of respondents \( s_r^{(o)} \) and \( s_r^{(s)} \), respectively. Noting that, under weighted MRHD imputation,
\[
V_i(y_i^*) = S_{yr}^2, \ V_j(x_j^*) = S_{xr}^2
\]
we can approximate \( V_i(\tilde{R}_i | s, s_r^{(o)}, s_r^{(s)}) \) by
\[
V_i(\tilde{R}_i | s, s_r^{(o)}, s_r^{(s)}) \approx \frac{1}{x_r} \left[ \sum_{i \in s} w_i (1 - a_i) S_{yr}^2 + \tilde{R}_i \sum_{i \in s} w_i (1 - b_i) S_{xr}^2 \right]. \tag{2.6}
\]
Expressions (2.5) and (2.6) will be useful in subsequent sections when discussing the bias and the variance of the imputed estimator \( \tilde{R}_i \). As we will see in sections 3 and 4, the conditional variance (2.6) is a measure of the variability due to the imputation mechanism.

Next, we describe two approaches to inference that will be used to obtain variance estimators in sections 3 and 4: the Nonresponse Model (NM) approach and the Imputation Model (IM) approach.

### 2.1 The nonresponse model approach

In the NM approach, inference is made with respect to the joint distribution induced by the sampling design and the nonresponse model. The nonresponse model is a set of assumptions about the unknown distribution of the response indicators \( R_i = \{a_i, b_i\}; i \in s \}. \) This unknown distribution is often called the nonresponse mechanism. Let
\[
p_{yi} = P(a_i = 1 | s, Z_i) \text{ be the response probability of unit } i \text{ to variable } y, \text{ where } Z_i = \{z_i; i \in s\} \text{ and } z_i \text{ is a vector of auxiliary variables available for all sample units used to form the imputation classes. Similarly, let } p_{xi} = P(b_i = 1 | s, Z_i) \text{ be the response probability of unit } i \text{ to variable } x. \] We assume that units respond independently; i.e., \( p_{yi} = P(a_i = 1, a_j = 1 | s, Z_i) = p_{yi} p_{yj} \) for \( i \neq j \). However, we do not assume that, for a given unit \( i \), response to variable \( y \) is independent of response to variable \( x \). In other words, if we let \( p_{yi} = P(a_i = 1, b_j = 1 | s, Z_i) \), then we have \( p_{yi} \neq p_{yi} p_{yj} \) in general. Within an imputation class, we assume a uniform response mechanism such that \( p_{yi} = p_{yi}, p_{xi} = p_{xi} \) and \( p_{yi} = p_{yi} \).

We also assume that, after conditioning on \( s \) and \( Z_i \), the nonresponse mechanism is independent of all other variables involved in the imputed estimator (2.4) as well as the joint selection probabilities. In other words, the distribution of \( R_i \) does not depend on \( Y_i = \{y_i; i \in s\} \), \( W_i = \{w_i; i \in s\} \) and \( \Pi_i = \\{\pi_{ij}; i \in s, j \in s\} \), after conditioning on \( s \) and \( Z_i \). As a result, except for the response indicators \( a_i \) and \( b_i \), we assume that all the variables involved in the imputed estimator (2.4) as well as the joint selection probabilities are treated as fixed when taking expectations and variances with respect to the nonresponse model. From this point on, we use the subscript \( q \) to denote the expectation and variance with respect to the nonresponse mechanism.

### 2.2 The imputation model approach

In the IM approach, inference is made with respect to the joint distribution induced by the imputation model, the sampling design and the nonresponse model. The imputation model is a set of assumptions about the unknown distribution of \( (Y_i, X_i) = \{y_i, x_i; i \in U\} \). Within an imputation class, the imputation model, \( m \), in the case of MRHD imputation, is given by
\[
m: \begin{cases} y_i = \mu_y + \varepsilon_i \\ x_i = \mu_x + \eta_i \end{cases} \tag{2.7}
\]
where \( \varepsilon_i \) is a random error term such that \( E_m(\varepsilon_i) = 0, \) \( E_m(\varepsilon_i, \varepsilon_j) = 0, \) for \( i \neq j \), \( V_m(\varepsilon_i) = \sigma^2_\varepsilon \) and \( \eta_i \) is a random error term such that \( E_m(\eta_i) = 0, \ V_m(\eta_i) = 0, \) for \( i \neq j \). \( V_m(\varepsilon_i, \eta_i) = \sigma^2_\varepsilon \). Furthermore, we assume that \( E_m(\varepsilon_i, \eta_i) = \sigma_{\varepsilon \eta} \). Here, \( E_m(\cdot), V_m(\cdot) \) and \( \text{Cov}_m(\cdot) \) denote respectively the expectation, the variance and the covariance operators with respect to model \( m \). It is implicit in the notation that expectations or variances with respect to model \( m \) are conditional on \( Z_U = \{z_i; i \in U\} \). In this approach, we assume that the distribution of the model errors \( (\varepsilon_i, \eta_i, \text{ and } \eta_i) = \{(\varepsilon_i, \eta_i); i \in U\} \) does not depend on \( s, s_r^{(o)}, s_r^{(s)}, W_i = \{w_i; i \in s\} \) and \( \Pi_i = \{\pi_{ij}; i \in U, j \in U\} \), after conditioning on \( Z_U \). As a result, except for the variables of interest \( y \) and \( x \), all variables involved in the imputed estimator (2.4) are treated as fixed when taking expectations and variances with respect to the imputation model.
2.3 Bias of the imputed estimator

To study the bias of the imputed estimator (2.4), we use the standard decomposition of the total error of \( \hat{R}_i \):

\[
\hat{R}_i - R = \left[ \hat{R} - R \right] + \left[ E_I(\hat{R}_i | s, s_r^{(1)}, s_r^{(2)}) - \hat{R} \right] + \left[ \hat{R}_i - E_I(\hat{R}_i | s, s_r^{(1)}, s_r^{(2)}) \right].
\]

(2.8)

The first term \( \hat{R} - R \) on the right-hand side of (2.8) is called the sampling error, the second term \( E_I(\hat{R}_i | s, s_r^{(1)}, s_r^{(2)}) - \hat{R} \) is called the nonresponse error, whereas the third term \( \hat{R}_i - E_I(\hat{R}_i | s, s_r^{(1)}, s_r^{(2)}) \) is called the imputation error.

Using a first-order Taylor expansion, it can easily be shown that, under the NM approach, the imputed estimator (2.4) is asymptotically \( pql \)-unbiased; that is, \( E_{pqI}(\hat{R}_i - R) \approx 0 \). Also, under the IM approach and model (2.7), it can be shown that the imputed estimator (2.4) is asymptotically \( mqpI \)-unbiased under the IM approach; that is, \( E_{mqI}(\hat{R}_i - R) \approx 0 \). Thus, the imputed estimator is robust in the sense that it is valid under either the NM approach or the IM approach. Note that for the asymptotic bias to be equal to 0 under both approaches, we require that the sample size within each imputation class is sufficiently large. From this point on, we thus assume that the bias of \( \hat{R}_i \) is negligible.

3. Variance estimation: The two-phase framework

In this section, we derive variance estimators under the two-phase framework and the IM approach according to the method proposed by Särndal (1992) and Deville and Särndal (1994). Using the decomposition (2.8), the total variance of \( \hat{R}_i \) can be approximated by

\[
V_{mpqI}(\hat{R}_i - R) \approx V_{pqI}(\hat{R}_i - R)^2
= V_{SAM} + V_{NR} + V_{I} + 2V_{mix},
\]

(3.1)

where \( V_{SAM} = E_n(V_a(\hat{R})) = E_n(V_{SAM}) \) is the sampling variance of the complete-data estimator \( \hat{R} \), \( V_{NR} = E_{pqI} V_n(E_I(\hat{R}_i | s, s_r^{(1)}, s_r^{(2)}) - \hat{R} | s, s_r^{(1)}, s_r^{(2)}) \) is the nonresponse variance of the imputed estimator \( \hat{R}_i \), \( V_I = E_{pqI} V_I(\hat{R}_i - E_I(\hat{R}_i | s, s_r^{(1)}, s_r^{(2)}) | s, s_r^{(1)}, s_r^{(2)}) \) is the imputation variance of the imputed estimator \( \hat{R}_i \), and \( V_{mix} = E_{pqm} [E_I(\hat{R}_i | s, s_r^{(1)}, s_r^{(2)}) - \hat{R}](\hat{R} - R) | s, s_r^{(1)}, s_r^{(2)}) \) is a mixed component. Note that the expression (3.1) contains only one cross product term, \( 2V_{mix} \), because the other cross product terms are asymptotically equal to 0.

3.1 Estimation of the sampling variance \( V_{SAM} \)

Let \( \hat{V}_{ORD} \) be the naive variance estimator of \( \hat{R}_i \); i.e., the variance estimator obtained by treating the imputed values as observed values. The variance estimator \( \hat{V}_{ORD} \) is thus obtained by replacing \( e_i \) by \( \tilde{e}_i = 1/\tilde{X}_i(\tilde{y}_i - \tilde{R}_i \tilde{x}_i) \) in (2.2) which leads to

\[
\hat{V}_{ORD} = \sum_{i=1}^{n} \Delta_{ij} \tilde{e}_i \tilde{e}_j.
\]

(3.2)

As we show now in the case of simple random sampling without replacement, \( \hat{V}_{ORD} \) overestimates \( V_{SAM} \) under MRHD imputation whenever \( \sigma_{en} > 0 \) (as it is usually the case in practice). After some algebra, we obtain

\[
E_{ml}(\hat{V}_{ORD} - \hat{V}_{SAM} | s, s_r^{(1)}, s_r^{(2)}) \\
\approx 2 \mu_\tau \left( \mu_\tau - \frac{1}{n} \right) \frac{1 - r_{xy}}{n} \sigma_{en}.
\]

Expression (3.3) shows that \( \hat{V}_{ORD} \) is \( mpql \)-biased for \( \hat{V}_{SAM} \) unless \( \sigma_{en} = 0 \), \( r_{xy} = n \) (which is the case of complete data) or \( n = N \) (which is the census case). The fact that \( \hat{V}_{ORD} \) is not a valid estimator of \( V_{SAM} \) can be easily explained by noting that although MRHD imputation preserves the variability, \( s_x^2 \) and \( s_y^2 \), corresponding to variables \( x \) and \( y \), it does not preserve the covariance, \( s_{xy} \), in (2.3). Indeed, imputation tends to underestimate relationships between variables that are positively correlated. As a result, \( \hat{V}_{ORD} \) overestimates \( V_{SAM} \) because of the presence of the minus sign in front of \( s_{xy} \) in (2.3). To overcome this difficulty, Särndal (1992) proposed to estimate \( V_{DIF} = E_{ml}(\hat{V}_{SAM} - \hat{V}_{ORD} | s, s_r^{(1)}, s_r^{(2)}) \) by a \( ml \)-unbiased estimator \( \hat{V}_{DIF} \); i.e., \( E_{ml}(\hat{V}_{DIF} | s, s_r^{(1)}, s_r^{(2)}) = V_{DIF} \). However, the derivation of this component for an arbitrary design involves very tedious algebra in the case of a ratio. Therefore, we propose an alternative that does not require any derivation but involves the construction of a new set of imputed values. It can be described as follows: whenever \( a_i = 0 \) and/or \( b_i = 0 \), select a donor \( j \) at random with replacement from the set of respondents to both variables \( x \) and \( y \) (i.e., the set of sampled units for which \( a_i = 1 \) and \( b_i = 1 \)) with probability \( w_j/\sum_{i=1}^{n} w_i a_i b_i \) and impute the vector \( (x_j, y_j) \). In other words, whenever one variable is missing, the observed value is discarded and set to missing; the missing values are then replaced by the values of a donor selected at random among the set of respondents to both variables \( x \) and \( y \) (often called the set of common donors). Similarly, when both variables are missing, the vector \((x_j, y_j)\) of a donor \( j \) is imputed. Then, use the standard variance estimator (2.2) valid in the complete response case using these imputed values. Let
\( \hat{V}_{\text{ORD}} \) denote the resulting variance estimator. Note that this new set of imputed values is used only to obtain a valid estimator of the sampling variance and is not used to estimate the parameter of interest \( R \). It can be shown that \( \hat{V}_{\text{ORD}} \) is an asymptotically \( mpqI \)-biased estimator of \( V_{\text{SAM}} \). In practice, one could, for example, create a variance estimation file containing the new set of imputed values and use standard variance estimation systems (used in the complete data case) to obtain an estimate of the sampling variance.

### 3.2 Estimation of the nonresponse variance \( V_{\text{NR}} \)

An estimator \( \hat{V}_{\text{NR}} \) of \( V_{\text{NR}} = E_{pq} V_m(E_j(\hat{R}_j | s, s_r^{(i)})) \) can simply be obtained by estimating \( V_m(E_j(\hat{R}_j | s, s_r^{(i)})) \). Using a first-order Taylor expansion, we obtain

\[
\hat{V}_{\text{NR}} = \frac{1}{\mu_r} \left[ \sum_{i=1}^{N_a} \frac{w_i^2 \sigma^2}{N_a - 2} - \frac{\sigma_i^2}{N_a} \right] + \left[ \sum_{i=1}^{N_b} \frac{w_i^2 \sigma_n^2}{N_b} - \frac{\sigma_n^2}{N_b} \right] - \mu_r \left( \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} \right) \left( \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} \right) \left( \mu_r \right) (3.4)
\]

where \( (\hat{N}, \hat{N}_a, \hat{N}_b) = \sum_{i=1}^{N} w_i (1, a_i, b_i) \). Now, let \( s_i^2 = 1/N \sum_{i=1}^{N} w_i (x_i - \bar{x})^2 \) and \( s_r^2 = 1/N \sum_{i=1}^{N} w_i (y_i - \bar{y})^2 \) with \( (\bar{x}, \bar{y}) = 1/N \sum_{i=1}^{N} w_i (\hat{x}_i, \hat{y}_i) \). Note that \( s_i^2 \) and \( s_r^2 \) denote the sample variability of the \( x \)-values and the \( y \)-values after imputation. It can be shown that \( s_i^2 \) and \( s_r^2 \) are respectively asymptotically \( ml \)-biased for the model variances \( \sigma^2 \) and \( \sigma_n^2 \). Also, let \( s_{xy} = 1/N \sum_{i=1}^{N} w_i a_i b_i (x_i - \bar{x}) (y_i - \bar{y}) \), where \( \hat{N}_{ab} = \sum_{i=1}^{N} w_i a_i b_i \) and \( (\bar{x}, \bar{y}) = \hat{N}_{ab}^{-1} \sum_{i=1}^{N} w_i a_i b_i (x_i, y_i) \). Note that \( s_{xy} \) is \( m \)-biased for the model covariance \( \sigma_{xy} \). It follows that \( \hat{V}_{\text{NR}} \) is obtained by estimating the unknown quantities in (3.4), which leads to

\[
\hat{V}_{\text{NR}} = \frac{1}{\mu_r} \left[ \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} (1) - \frac{\sigma_i^2}{N_a} \right] + \left[ \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} (1) - \frac{\sigma_n^2}{N_b} \right] - \mu_r \left( \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} \right) \left( \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} \right) (3.5)
\]

The estimator (3.5) is asymptotically \( mpqI \)-biased for \( V_{\text{NR}} \). In the special case of simple random sampling without replacement, expression (3.5) reduces to

\[
\hat{V}_{\text{NR}} = \frac{1}{\mu_r} \left[ \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} (1) - \frac{\sigma_i^2}{N_a} \right] + \left[ \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} (1) - \frac{\sigma_n^2}{N_b} \right] - \mu_r \left( \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} \right) \left( \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} \right) (3.6)
\]

It follows that \( \hat{V}_{\text{NR}} \) in (3.6) is asymptotically \( mpqI \)-biased for \( V_{\text{NR}} \). In the special case of simple random sampling without replacement, expression (3.6) reduces to

\[
\hat{V}_{\text{NR}} = \frac{N^2}{\mu_r} \left[ \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} (1) - \frac{\sigma_i^2}{N_a} \right] + \left[ \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} (1) - \frac{\sigma_n^2}{N_b} \right] (3.7)
\]

### 3.3 Estimation of the imputation variance \( V_{\text{IMP}} \)

An estimator \( \hat{V}_I \) of \( V_I = E_{pq} V_I (\hat{R}_I - \hat{R} | s, s_r^{(i)}) \) can simply be obtained by estimating \( V_I (\hat{R}_I - \hat{R} | s, s_r^{(i)}) \) given by (2.6). An asymptotically \( I \)-unbiased of \( V_I \) is then given by

\[
\hat{V}_I = \frac{1}{\mu_r} \left[ \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} (1 - a_i) s_i^2 + \hat{R}_I \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} (1 - b_i) s_r^2 \right] (3.8)
\]

### 3.4 Estimation of the mixed component \( V_{\text{MIX}} \)

Finally, we obtain an estimator \( \hat{V}_{\text{MIX}} \) of \( V_{\text{MIX}} \) by estimating

\[
\hat{V}_{\text{MIX}} = \frac{1}{\mu_r} \left[ \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} (1) - \frac{\sigma_i^2}{N_a} \right] + \left[ \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} (1) - \frac{\sigma_n^2}{N_b} \right] - \mu_r \left( \sum_{i=1}^{N_a} \frac{w_i^2}{N_a} \right) \left( \sum_{i=1}^{N_b} \frac{w_i^2}{N_b} \right) (3.9)
\]
\[ E_n[(E_I(\hat{R}_i | s, s^{(y)}, s^{(x)}) - \hat{R})(\hat{R} - R) | s, s^{(y)}, s^{(x)})]. \]

Using a first-order Taylor expansion, we obtain
\[
E_n[(E_I(\hat{R}_i | s, s^{(y)}, s^{(x)}) - \hat{R})(\hat{R} - R) | s, s^{(y)}, s^{(x)}] \\
\approx 1 \left[ \frac{\sum_{i \in s} w_i^2 a_i}{NN_a} - \frac{\sum_{i \in s} w_i^2}{N^2} \right] \sigma_{\xi}^2 \\
+ \left[ \frac{\sum_{i \in s} w_i^2 b_i}{NN_b} - \frac{\sum_{i \in s} w_i^2}{N^2} \right] \frac{\mu_{\xi}}{\sigma_{\eta}} \sigma_{\eta}^2 \\
- 2 \left[ \frac{\sum_{i \in s} w_i^2 a_i}{NN_a} + \frac{\sum_{i \in s} w_i^2 b_i}{NN_b} - 2 \frac{\sum_{i \in s} w_i^2}{N^2} \right] \frac{\mu_{\xi}}{\sigma_{\eta}} \sigma_{\eta}. \tag{3.7} \]

An estimator of (3.7) is thus given by
\[ \hat{V}_{\text{MIX}} = \]
\[
\frac{1}{\hat{x}_I} \left[ \frac{\sum_{i \in s} w_i^2 a_i}{NN_a} - \frac{\sum_{i \in s} w_i^2}{N^2} \right] s^2_\delta + \frac{\sum_{i \in s} w_i^2 b_i}{NN_b} - \frac{\sum_{i \in s} w_i^2}{N^2} \hat{R}_i s^2_{\delta y} \\
- 2 \left[ \frac{\sum_{i \in s} w_i^2 a_i}{NN_a} + \frac{\sum_{i \in s} w_i^2 b_i}{NN_b} - 2 \frac{\sum_{i \in s} w_i^2}{N^2} \right] \hat{R}_i s^2_{\delta y}. \tag{3.8} \]

The estimator (3.8) is asymptotically mpql-unbiased for \( V_{\text{MIX}} \). In the case of simple random sampling without replacement, the component \( \hat{V}_{\text{MIX}} \) is equal to zero. More generally, the component \( \hat{V}_{\text{MIX}} \) is equal to zero for any unistage self-weighting design (i.e., a sampling design for which the sampling weight are all equal). For unequal probability designs, it is important to include the component \( \hat{V}_{\text{MIX}} \) because its contribution (positive or negative) to the overall variance could be substantial (Brick, Kalton and Kim (2004)).

Finally, an asymptotically mpql-unbiased estimator of the total variance \( V_{\text{TOT}} = V_{\text{mpql}}(\hat{R}_i - R) \) is thus given by
\[
\hat{V}_{\text{TOT}}^{(TP)} = \hat{V}_{\text{ORD}} + \hat{V}_{\text{NR}} + \hat{V}_{\text{I}} + 2\hat{V}_{\text{MIX}}. \]

4. Variance estimation: The reverse framework

In this section, we derive variance estimators under the reverse framework and both the NM and the IM approaches according to the method proposed by Shao and Steel (1999). Recall that, under this framework, we require the additional assumption that the response probabilities do not depend on the sample \( s \). Under the NM approach, the total variance of \( \hat{R}_i \) can be approximated by
\[
V(\hat{R}_i - R) \approx E_q V_q E_I(\hat{R}_i - R | a, b) \\
+ E_{pq} V_I(\hat{R}_i - R | a, b) + V_q E_{pl}(\hat{R}_i - R | a, b), \tag{4.1} \]

where \( a = (a_1, ..., a_y)' \) and \( b = (b_1, ..., b_y)' \) denote the vectors of response indicators to variables \( y \) and \( x \), respectively.

Under the IM approach, the total variance of \( \hat{R}_i \) can be approximated by
\[
V(\hat{R}_i - R) \approx E_m V_m E_I(\hat{R}_i - R | a, b) \\
+ E_{mq} V_I(\hat{R}_i - R | a, b) \\
+ E_q V_m E_{pl}(\hat{R}_i - R | a, b). \tag{4.2} \]

Under both the NM and the IM approaches, an estimator of the first term on the right hand side of (4.1) and (4.2) can be obtained by finding an asymptotically mpql-unbiased estimator of \( V_q E_q(\hat{R}_i | a, b) \). Also, the second term on the right hand side of (4.1) and (4.2) can be estimated by \( \hat{V}_{\text{I}} \) given by (3.6). Under the NM approach, an estimator of the last term on the right hand side of (4.1) can be obtained by estimating \( V_q E_{pl}(\hat{R}_i | a, b) \), whereas an estimator of the last term on the right hand side of (4.2) can be obtained by estimating \( V_m E_{pl}(\hat{R}_i - R | a, b) \) under the IM approach. As a result, the estimators of the first two terms in (4.1) and (4.2) are identical and thus are valid regardless of the approach (NM or IM) used for inference. Only the third term on the right hand side of (4.1) and (4.2) will depend on the approach used. In the case of the IM approach, specification and validation of the imputation model is crucial to achieve asymptotic unbiasedness of the third component, whereas in the case of the NM approach, the asymptotic unbiasedness of the third component relies on the correct specification of the nonresponse model.

4.1 Estimation of \( V_p E_I(\hat{R}_i - R | a, b) \)

Using a first-order Taylor expansion and expression (2.5), an estimator of \( V_p E_I(\hat{R}_i - R | a, b) \), denoted by \( \hat{V}_p \), is given by
\[
\hat{V}_p = \sum_{i \in s} \sum_{j \in s} A_{ij} \xi \bar{x}_j, \tag{4.3} \]

where
\[ \xi_i = \frac{1}{N_a} \left[ \frac{1}{N_a} a_i (y_i - \bar{y}) - \left( \frac{\bar{y}}{N_y} \right) \frac{1}{N_b} b_i (x_i - \bar{x}) \right]. \]
In other words, the estimator \( \hat{V}_1 \) is obtained from the complete data variance estimator (2.2) by replacing \( c_i \) by \( \xi_i \). In the case of simple random sampling without replacement, the estimator (4.3) reduces to

\[
\hat{V}_1 = \left( 1 - \frac{n}{N} \right) \frac{1}{X'_2} \left[ \frac{s_{xy}^2}{r_y} + \hat{R}_1^2 \frac{s_{xy}^2}{r_x} - 2 \hat{R}_1 \left( \frac{r_{xy}}{r_y} - \frac{r_{xy}}{r_x} \right) \frac{s_{xy}}{r_y} \right]. 
\] (4.4)

### 4.2 Estimation of \( V_q E_{pl} (\hat{R}_1 - R | a, b) \) under the NM approach

First, note that

\[
E_{pl} (\hat{R}_1) \approx \frac{Y_q N_b}{N_a X_b},
\]

where \((Y_a, N_a) = \sum_{i \in U} a_i (y_i, 1)\) and \((X_b, N_b) = \sum_{i \in U} b_i (x_i, 1)\).

Using a first-order Taylor expansion, it can be shown that \( V_q E_{pl} (\hat{R}_1 - R | a, b) \) can be approximated by

\[
V_q E_{pl} (\hat{R}_1 - R | a, b) \approx \frac{1}{N X'_2} \left[ \left( \frac{1 - p_y}{p_y} \right) S_y^2 + R^2 \left( \frac{1 - p_x}{p_x} \right) S_x^2 \right] - 2 R \left( \frac{p_{xy} - p_x p_y}{p_x p_y} \right) S_{xy},
\] (4.5)

where

\[
S_y^2 = \frac{1}{N - 1} \sum_{i \in U} (y_i - \bar{Y})^2,
\]

\[
S_x^2 = \frac{1}{N - 1} \sum_{i \in U} (x_i - \bar{X})^2,
\]

and

\[
S_{xy} = \frac{1}{N - 1} \sum_{i \in U} (x_i - \bar{X})(y_i - \bar{Y})
\]

with

\[
(\bar{Y}, \bar{X}) = \frac{1}{N} \sum_{i \in U} (y_i, x_i).
\]

An estimator of \( V_q E_{pl} (\hat{R}_1 - R | a, b) \) is obtained by estimating unknown quantities in (4.5), which leads to

\[
\hat{V}_2^{\text{NM}} = \frac{1}{N X'_2} \left[ \left( \frac{1 - \hat{p}_y}{\hat{p}_y} \right) \hat{s}_{y}^2 + \hat{R}_1^2 \left( \frac{1 - \hat{p}_x}{\hat{p}_x} \right) \hat{s}_{x}^2 \right] - 2 \hat{R}_1 \left( \frac{\hat{p}_{xy} - \hat{p}_x \hat{p}_y}{\hat{p}_x \hat{p}_y} \right) \hat{s}_{xy}
\]

\[
= \frac{1}{X'_2} \left[ \left( \frac{1}{N_a} - \frac{1}{N} \right) \hat{s}_{y}^2 + \hat{R}_1^2 \left( \frac{1}{N_b} - \frac{1}{N} \right) \hat{s}_{x}^2 \right] - 2 \hat{R}_1 \left( \frac{\hat{N}_{ab}}{N_{b} \cdot N_{ab}} - \frac{1}{N} \right) \hat{s}_{xy},
\] (4.6)

where \( \hat{p}_y = \frac{\hat{N}_a}{N_a}, \hat{p}_x = \frac{\hat{N}_b}{N_b} \) and \( \hat{p}_{xy} = \frac{\hat{N}_{ab}}{N_{b} \cdot N_{ab}} \).

The estimator (4.6) is asymptotically \( pqI \)-unbiased for the approximate variance (4.5), noting that \( s_{y}^2, s_{x}^2 \) and \( s_{xy} \) are asymptotically \( pqI \)-unbiased for \( S_y^2, S_x^2 \) and \( S_{xy} \), respectively. In the case of simple random sampling without replacement, the estimator (4.6) reduces to

\[
\hat{V}_2^{\text{NM}} = \frac{1}{X'_2} \left[ \left( \frac{N_a - r_y}{N_a} \right) \hat{s}_{y}^2 + \hat{R}_1^2 \left( \frac{N_b - r_x}{N_b} \right) \hat{s}_{x}^2 \right] - 2 \hat{R}_1 \left( \frac{r_{xy}}{N_a} - \frac{r_{xy}}{N_b} \right) \hat{s}_{xy}. 
\] (4.7)

### 4.3 Estimation of \( E_q V_m E_{pl} (\hat{R}_1 - R | a, b) \) under the IM approach

Using a first-order Taylor expansion, it can be shown that \( E_q V_m E_{pl} (\hat{R}_1 - R | a, b) \) can be approximated by

\[
E_q V_m E_{pl} (\hat{R}_1 - R | a, b) \approx \frac{1}{\mu_x} \left[ \left( \frac{1}{E_q (N_a)} - \frac{1}{N} \right) \sigma_c^2 \right] + \left( \frac{1}{\mu_x} \right) \left( \frac{1}{E_q (N_b)} - \frac{1}{N} \right) \sigma_{e1}^2 - 2 \left( \frac{1}{\mu_x} \right) \sum_{i \in U} \left[ \left( \frac{N_{ab}}{N_a N_b} - \frac{1}{N} \right) \sigma_{e1} \right],
\] (4.8)

where \( N_{ab} = \sum_{i \in U} a_i b_i \). An estimator of \( E_q V_m E_{pl} (\hat{R}_1 - R | a, b) \) is obtained by estimating unknown quantities in (4.8), which leads to

\[
\hat{V}_2^{\text{IM}} = \frac{1}{X'_2} \left[ \left( \frac{1}{N_a} - \frac{1}{N} \right) \hat{s}_{y}^2 + \hat{R}_1^2 \left( \frac{1}{N_b} - \frac{1}{N} \right) \hat{s}_{x}^2 \right] - 2 \hat{R}_1 \left( \frac{\hat{N}_{ab}}{N_a N_{ab}} - \frac{1}{N} \right) \hat{s}_{xy}. 
\] (4.9)

The estimator (4.9) is asymptotically \( mpql \)-unbiased for the approximate variance (4.8). It is interesting to note that, under weighted MRHD imputation, the estimator \( \hat{V}_2^{\text{NM}} \) in (4.6) obtained under the NM approach is identical to \( \hat{V}_2^{\text{IM}} \) in (4.9) obtained under the IM approach. However, this may not be the case with a different imputation method. Also, the component \( \hat{V}_2 \) is negligible with respect to \( \hat{V}_1 \) when the sampling fraction \( n/N \) is negligible, where \( \hat{V}_1 \) stands for \( \hat{V}_2^{\text{NM}} \) or \( \hat{V}_2^{\text{IM}} \). In this case, the component \( \hat{V}_2 \) may be omitted from the calculations.

Finally, an estimator of the total variance under the reverse framework is given by

\[
\hat{V}_{\text{TOT}}^{\text{(RE)}} = \hat{V}_1 + \hat{V}_2.
\]
Under the reverse framework, both the NM approach and the IM approach lead to the same estimator of the total variance. Thus, the variance estimator $\hat{V}_{TOT}^{(RE)}$ is robust in the sense that it is valid under either the NM approach or the IM approach.

5. Summary and conclusions

In this paper, we have derived variance estimators for the imputed estimator of a ratio under two different frameworks. The reverse framework facilitates the derivation of the variance expressions (in comparison with the customary two-phase framework), especially if the sampling fraction is small, in which case we can omit the component $\hat{V}_2$. However, unlike the two-phase framework, it requires an additional assumption that the response probabilities do not depend on the realized sample $s$. Also, the two-phase framework uses a natural decomposition of the total error that leads to a natural decomposition of the total variance. That is, the total variance can be expressed as the sum of the sampling variance, the nonresponse variance and the imputation variance, which allows the survey statistician to get an idea of the relative magnitude of each component. Under the reverse approach, there is no easy interpretation for the variance components (except the imputation variance).

We have considered the case of weighted MRHD imputation within classes. Another version of weighted random hot-deck imputation, which we call weighted joint random hot deck (JRHD) imputation, is identical to weighted MRHD imputation, except that when both variables are missing, a donor $j$ is selected at random from the set of common donors (i.e., the set of respondents to both variables $y$ and $x$) with probability $w_j / \sum_{i \in c} w_i a_i b_i$ and the vector $(x_j, y_j)$ is imputed. This version of the method helps preserving relationships between survey variables, contrary to imputing independently each variable. The results for JRHD imputation can be obtained using similar techniques presented in this paper. Finally, the results presented in this paper can be easily extended to the case of both deterministic and random regression imputation performed within imputation classes.

Acknowledgements

The author thanks an Associate Editor for constructive comments and suggestions that helped improving the quality of the paper. David Haziza’s work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References


Efficient bootstrap for business surveys

James Chipperfield and John Preston

Abstract

The Australian Bureau of Statistics has recently developed a generalized estimation system for processing its large scale annual and sub-annual business surveys. Designs for these surveys have a large number of strata, use Simple Random Sampling within Strata, have non-negligible sampling fractions, are overlapping in consecutive periods, and are subject to frame changes. A significant challenge was to choose a variance estimation method that would best meet the following requirements: valid for a wide range of estimators (e.g., ratio and generalized regression), requires limited computation time, can be easily adapted to different designs and estimators, and has good theoretical properties measured in terms of bias and variance. This paper describes the Without Replacement Scaled Bootstrap (WOSB) that was implemented at the ABS and shows that it is appreciably more efficient than the Rao and Wu (1988)'s With Replacement Scaled Bootstrap (WSB). The main advantages of the Bootstrap over alternative replicate variance estimators are its efficiency (i.e., accuracy per unit of storage space) and the relative simplicity with which it can be specified in a system. This paper describes the WOSB variance estimator for point-in-time and movement estimates that can be expressed as a function of finite population means. Simulation results obtained as part of the evaluation process show that the WOSB was more efficient than the WSB, especially when the stratum sample sizes are sometimes as small as 5.

Key Words: Variance; Bootstrap; Stratified sampling.

1. Introduction

In 2000, the Australian Bureau of Statistics (ABS) first obtained a register of businesses containing taxation data from the Australian Taxation Office (ATO). The data items included turnover, sales, and other expense items. In 2001, the ABS used this register as a sampling frame for some surveys in order to improve the efficiency of its sample designs. This data is updated for each business at least annually. To make maximum use of these administrative data items in estimation the ABS developed a generalized estimation system called ABSEST, with the capability of supporting generalized regression estimation (GREG) and variance estimation. ABSEST has been routinely used for the monthly ABS Retail Survey since July 2005.

A generalized estimation system is highly desirable for statistical agencies as it supports a variety of survey output requirements at high levels of statistical rigor for an acceptable cost. The ABS has invested considerable resources into its generalized estimation system for business surveys. Prior to 1998, the ABS’s generalized estimation system was capable of Horvitz-Thompson, ratio, and two-phase estimation with variance estimates based on Taylor Series (TS) approximations. In 1999, the Taylor Series method was replaced with the Jackknife method. Subsequent feedback about the computer design and usability were that changes to the generalized estimation system made it increasingly complex to maintain and develop and that processing time could be undesirably long. These key features were important when choosing the variance estimation method for ABSEST.

Core survey output statistics for ABS business surveys are estimates at a point in time, estimates of movement between two time points, and estimates of rates. Business surveys are equal probability designs within stratum, are highly stratified (100s of strata), can be either single or two phase sample designs, and for surveys that sample on more than one occasion the overlapping sample can range from 0 to 100%. The sample size for business surveys range from less than 1,000 to 15,000; stratum level sample sizes can be as low as 3 and as high as several hundred.

Section 2 introduces the GREG estimator. Section 3 discusses alternative variance estimators for GREG and justifies why the Bootstrap variance estimator was chosen for ABSEST. Section 4 describes the Without Replacement Scaled Bootstrap (WOSB) and Rao and Wu (1988)’s With Replacement Bootstrap (WSB) variance estimators for point-in-time estimates under single-phase designs. Section 5 describes the WOSB for movement estimates. Section 6 measures the bias and variance properties of WOSB and WSB in a simulation study. Section 7 gives some concluding remarks.

2. Generalised regression (GREG) estimator

In this section we briefly describe the GREG that is implemented in ABSEST. Consider a finite population \( U \) divided into \( H \) strata \( U = \{U_1, U_2, ..., U_h\} \), where \( U_h \) is comprised of \( N_h \) units. The finite population total of interest is \( Y = \sum_{h=1}^{H} Y_h \), where \( Y_h = \sum_{i \in u_h} y_{hi} \) and \( h = 1, ..., H \). Within stratum \( h \), the sample \( s_h \) of \( n_h \) units is selected.

1. James Chipperfield, Australian Bureau of Statistics. E-mail: james.chipperfield@abs.gov.au; John Preston, Australian Bureau of Statistics. E-mail: john.preston@abs.gov.au.
from $U_h$ by Simple Random Sampling without Replacement (SRSWOR). The complete sample set is denoted by $s = \{s_1, s_2, \ldots, s_H\}$.

Consider the case where a $K$ vector of auxiliary variables $x_i = (x_{i1}, x_{i2}, \ldots, x_{iK})^T$ is available for $i \in s$ and the corresponding vector of population totals $X = \sum_{s_i} x_i$ are known. The GREG estimator (Särndal, Swensson and Wretman 1992, page 227) is given by

$$\hat{Y}_{\text{reg}} = \sum_{s_i} \tilde{w}_i y_i = \sum_{s_i} \tilde{w}_i y_i + (X - \hat{X})^T \tilde{B},$$

where $\tilde{w}_i = w_i g_i$, $w_i = N_h / n_{ih}$ and $\tilde{B} = \hat{T}^{-1}$. It is well known that $\hat{Y}_{\text{reg}}$ is unbiased to $O(n^{-1})$. The weights $\tilde{w}_i$ are stored for ready calculation of estimates. In practice bounds will be placed on the weights, $\tilde{w}_i$. If the weights, $\tilde{w}_i$, given by the above equation, are outside these bounds, they are calculated through iteration (see Method 5 of Singh and Mohl 1996).

The expression for $\hat{Y}_{\text{reg}}$ can be adapted to a range of estimates, including domains and multi-phase (see Estevao, Hidiroglou and Särndal 1995). For example, when $x_i = 1$, $\hat{Y}_{\text{reg}}$ becomes the Horvitz-Thompson estimator given by

$$\hat{Y} = \sum_{i} w_i \sum_{s_{ih}} y_{ih},$$

with estimated variance $\text{Var}(\hat{Y}) = \sum_{i} N_i^2 \text{Var}(\tilde{y}_i)$, where $\tilde{y}_i = (n_i - 1)^{-1} \sum_{s_{ih}} (y_{ih} - \tilde{y}_h)^2 / n_i$ and $\tilde{y}_h = \sum_{s_i} y_{ih} / n_h$. $\tilde{y}_i$ is the response vector from unit $i$ with elements $y_{ih}$, and $\tilde{y}_h$ is a smooth function. Estimating the variance using a replication method involves the following steps:

(i) independently sub-sampling from the set $s$ a total of $R$ times;

(ii) for each of the $R$ sub-samples computing $w_i^* = b_i^* w_i$, where $b_i^*$ depends upon the number of times unit $i$ is selected in the sub-sample;

(iii) calculate $\hat{\theta}^* = \hat{\theta}(\hat{Y}^*)$, where $\hat{Y}^* = \sum_{s_{ih}} w_i^* y_{ih};$

(iv) estimate variance of $\theta$ by $\hat{\text{Var}}_{\text{rep}}(\hat{\theta}) = (R - 1)^{-1} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2$, where $\hat{\theta}^{(r)}$ is the estimate of $\theta$ based on the $r$th replicate sample. Note: the expression for replicate weights, $w_i^* = b_i^* w_i$, includes the Jackknife, Bootstrap and Balanced Repeated Replication as special cases.

As we can express $\hat{Y}_{\text{reg}}$ by a function $\hat{\theta}$, the variance of $\hat{Y}_{\text{reg}}$ can be calculated by the above steps where specifically steps (iii) and (iv) respectively become: (iii) calculate $\hat{Y}_{\text{reg}}^{(r)} = \sum_{s_i} \tilde{w}_i^{(r)} y_i$, where $\tilde{w}_i^{(r)} = w_i^{(r)} g_i$ and $g_i$ has the same form as $g_i$ but is calculated using the weights $w_i^{(r)}$ instead of the weights $w_i$ for $i \in s$; (iv) estimating variance by $\hat{\text{Var}}_{\text{rep}}(\hat{Y}_{\text{reg}}^{(r)}) = (R - 1)^{-1} \sum_{r=1}^R (\hat{Y}_{\text{reg}}^{(r)} - \hat{Y}_{\text{reg}})^2$.

The attractive feature of these replication methods is that only the selection of the replicate samples and the value $b_i^*$ is required to calculate unbiased variance estimates for many commonly used sample designs and for estimators that have good first order Taylor Series approximations. Also if the replicate weights, $\tilde{w}_i^{(r)}$ are stored the variance estimates of $\hat{Y}_{\text{reg}}$ require simple calculations that can be completed in a short time; this approach of storing replicate weights has been applied successfully by the ABS’ generalized estimation system for household surveys. Once replicate weights are available, calculation of variance for a variety of analysis, such as linear regression, involves simple calculations that require little time and does not require the analyst to have knowledge about the sample design. Next we consider the relative merits of some replicate variance estimators for implementation in ABEST.

The drop-one Jackknife forms replicate samples, $s^*$, by dropping one unit at a time. This implies that $R = n$. For large-scale surveys this storage requirement is excessive. The delete-a-group Jackknife, while reducing $R$ by dropping a group of units within a stratum at a time, would still have at least $R = 2H$ replicates- a minimum of two groups per stratum is required to calculate variance. Despite performing well in an empirical study where $n_h = 2$ (see Shao and Tu 1995, page 251), the Jackknife was rejected on the basis of its excessive storage requirement.

For stratified designs the scaled Balanced Repeated Replication (BRR) requires approximately $R = H$ replicate weights. Firstly, the replicate samples are formed by randomly splitting the stratum sample $s_h$ into two groups then allocating one of these groups to $s^*_h$ for each
4. Without replacement scaled bootstrap (WOSB) for point in time estimates

4.1 Method

For point-in-time GREG estimates, the Without Replacement Scaled Bootstrap (WOSB) variance estimator involves repeating the following \( R \) times:

(a) forming the set \( s' \) by selecting \( m_h \) units by SRSWOR from \( s_h \) independently within each stratum \( h = 1, ..., H \), where \( m_h = \lfloor n_h / 2 \rfloor \) and the operator \( \lfloor \cdot \rfloor \) rounds down its argument down to the nearest integer;

(b) calculating \( w_{hi}^* = w_h (1 - \gamma_h + \gamma_h m_h / m_h) \) for \( i \in s_h \), where \( \gamma_h = \sqrt{(1 - f_h) m_h / (n_h - m_h)} \), \( \delta_{hi}^* \) is 1 if \( i \in s_h' \) and 0 otherwise; and

(c) calculating \( \hat{w}_{hi} = w_{hi}^* g_{hi}^* \) for \( i \in s' \); and

(d) calculating the \( r^{th} \) Bootstrap estimate of \( Y, \hat{Y}_{reg} = \sum_{i=1}^{N} \hat{w}_{hi}^* y_i \). The justification \( m_h = \lfloor n_h / 2 \rfloor \) is given in section 4.2. The Bootstrap variance estimator is given by the Monte Carlo approximation, \( \hat{\text{Var}}(\hat{Y}_{reg}) = (R - 1)^{-1} \sum_{r=1}^{R} (\hat{Y}_{reg}^r - \hat{Y}_{reg})^2 \). The WSB method is the same as WOSB except that the replicate samples are selected by SRSWR and the scaling factor is instead \( \gamma_h = \sqrt{(1 - f_h) m_h / (n_h - 1)} \), where \( m_h \) is often set to \( n_h - 1 \) in the literature. Preston and Chipperfield (2002) found that WOSB was found to have significantly less replication error than the WSB- the error due to replicate sampling and conditional on the sample set.

It is easy to see that the WOSB and WSB estimators are unbiased estimators of \( \text{Var} (\hat{\theta}) \). The TS approximate variance is given by \( \hat{\text{Var}}(\hat{\theta}) = \nabla \hat{\theta} \hat{\nabla} \hat{\theta} \), where \( \hat{\nabla} \) is a \( P \times P \) matrix with elements

\[
\hat{\text{Cov}}(\hat{Y}_p, \hat{Y}_{p'}) = \frac{N}{n} (1 - f) \hat{s}_{p, p'}
\]

where

\[
\hat{s}_{p, p'} = \frac{1}{n - 1} \sum_{i=1}^{n} (y_{p_i} - \hat{\hat{\nabla}} p)(y_{p'_{i}} - \hat{\nabla} p');
\]

\[
\hat{\hat{\nabla}} p = \frac{1}{n} \sum_{i=1}^{n} y_{p_i};
\]

\[
\hat{\nabla} p = \sum_{i=1}^{n} y_{p_i}
\]

for \( p, p' = 1, ..., P \), and \( \nabla' = (\partial / \partial Y_1, ..., \partial / \partial Y_P) |_{\tilde{Y}} \). It is easy to see that

\[
E_r(\text{Var}(\hat{\theta})) = \nabla \hat{\theta} E_r(\hat{\nabla}(\hat{\theta})) \nabla \hat{\theta} = \nabla \hat{\theta} \hat{\nabla} (\tilde{Y}) \nabla \hat{\theta},
\]

by noting that

\[
E_r[\text{Cov}(\hat{\nabla}_p, \hat{\nabla}_{p'})] = \text{Cov}(\hat{\nabla}_p, \hat{\nabla}_{p'})
\]
where \( E \) denotes the expectation with respect to resampling. Note the scaling constants applied to \( w_{hi} \) to calculate the replicate weights are chosen so that the correct finite population correction factor is obtained. It therefore follows that the Monte Carlo approximation to the variance, 
\[
\overline{\text{Var}}_B(\hat{\theta}) = (R - 1) \sum_h R_i \left( \hat{\theta}^{(t)}_i - \hat{\theta} \right)^2,
\]

is unbiased for \( \text{Var}(\hat{\theta}) \).

### 4.2 A note on the relative efficiency of WSB and WOSB sampling

To simplify notation, let \( \hat{\theta}_{\text{boot}} = \overline{\text{Var}}_B(\hat{\theta}) \). The variance of the Bootstrap variance estimator can be written as
\[
\text{Var}(\hat{\theta}_{\text{boot}}) = \text{E}(\text{Var}(\hat{\theta}_{\text{boot}} | s)) + \text{E}(\text{Var}(\hat{\theta}_{\text{boot}} | s)),
\]

where \( s \) denotes the expectation with respect to the sample design. If \( \hat{\theta}_{\text{boot}} \) is unbiased (i.e., \( E(\hat{\theta}_{\text{boot}} | s) = \text{Var}(\hat{\theta}) \)) then \( \text{Var}(\hat{\theta}_{\text{boot}}) \) does not depend upon how the replicate samples are selected. The term \( \text{Var}(\hat{\theta}_{\text{boot}} | s) \) is the replication error conditional on the sample and is inversely proportional to \( R \). The value of \( R \) is chosen to be sufficiently large such that \( \text{Var}(E(\hat{\theta}_{\text{boot}} | s)) \) is small relative to \( \hat{\theta}_{\text{boot}} \), the estimated sample variance. The efficiency of two Bootstrap estimators can be compared by the size of \( \text{Var}(E(\hat{\theta}_{\text{boot}} | s)) \) when both estimators have the same value of \( R \). Next we summarise empirical results based on actual data that show the WOSB can be significantly more efficient than WSB. The benefits of efficiency are either reduced computation time and/or more accurate variance estimates.

Preston and Chipperfield (2002) compared the efficiency of WOSB with \( m_h = \lceil n_h/2 \rceil \) and WSB with \( m_h = n_h - 1 \) (see Rao and Wu 1984) for the Australian Quarterly Economic Activity Survey in March 2000. This survey has a stratum level sample size that varies from 4 and into the 100s. The results (derived from Preston and Chipperfield 2002, Table 1) show at the national level the size of \( \text{Var}(E(\hat{\theta}_{\text{boot}} | s)) \) was 54\% smaller for WOSB compared with WSB sampling when \( R = 100 \) (See Preston and Chipperfield 2002 for more empirical estimates of \( \text{Var}(E(\hat{\theta}_{\text{boot}} | s)) \) for WSB and WOSB). In other words, WOSB required about half the number of replicates to achieve the same replication error as WSB. This represents a significant efficiency gain. Another benefit of WOSB over WSB is that the computational time in selecting the replicate samples is considerably less.

From empirical investigations, the choice of \( m_h = \lceil n_h/2 \rceil \) for WOSB minimized \( \text{Var}(E(\hat{\theta}_{\text{boot}} | s)) \). As \( n \) increases, we suspect that the difference between WOSB and WSB will reduce to approximately zero. More work needs to be done to establish these properties.

### 5. Movement variance between single phase estimates

A key output requirement of many business surveys is the estimate of change between two time points. Denote the finite population at time \( t \) by \( U^{(t)} = \{U_1^{(t)}, U_2^{(t)}, \ldots, U_h^{(t)}\} \), where \( U_h^{(t)} \) is the stratum \( h \) population at time \( t \) that is made up of \( N_h^{(t)} \) units. The population total at time \( t \) is
\[
Y^{(t)} = \sum_h N_h^{(t)} y_h^{(t)}.
\]

Estimating the variance of \( \Delta^{(t)} = \hat{Y}^{(t)} - \hat{Y}^{(t-1)} \), the difference between two time periods, is the focus of this section. The terms corresponding to \( n_{hc}, n_{hc}^{(t)}, \) and \( n_{hc}^{(t-1)} \) to be the number of units in the following sets \( U_h^{(t)} \cap U_h^{(t-1)} = S_h^{(t-1)} \cap S_h^{(t)} \), \( S_h^{(t)} = S_h^{(t)} - S_h^{(t-1)} \), and \( S_h^{(t-1)} = S_h^{(t)} - S_h^{(t-1)} \) respectively. When sampling on two occasions define \( n_{hc}, n_{hc}^{(t)}, \) and \( n_{hc}^{(t-1)} \) units from \( S_h^{(t-1)} \) and an SRSWOR of \( U_h^{(t)} \). The time 2 sample is the union of the following two samples: an SRSWOR of \( n_{hc} \) units from \( S_h^{(t)} \) and an SRSWOR of \( n_{hc}^{(t-1)} \) units from \( U_h^{(t)} \cap U_h^{(t-1)} \). The time 2 sample is effectively an SRSWOR from \( U_h^{(t)} \). At the ABS, the size of the overlapping sample, \( n_{hc} \), is controlled by the Permanent Random Number method (see Brewer, Gross and Lee 1999).

The estimator of \( \text{Var}(\Delta) \) can be expressed as
\[
\text{Var}(\Delta) = \text{Var}(\hat{Y}^{(t)}) + \text{Var}(\hat{Y}^{(t-1)}) - 2 \text{Cov}(\hat{Y}^{(t)}, \hat{Y}^{(t-1)}),
\]

Consider the Horvitz-Thompson estimator \( \hat{\Delta} = \hat{Y}^{(t)} - \hat{Y}^{(t-1)} \), where \( t = 1, 2 \) and \( \hat{Y} \) is defined analogously to \( \hat{\Delta} \). Tam (1985) show that when \( U_h^{(t)} \neq U_h^{(t-1)} \), an unbiased estimator of \( \text{Var}(\Delta) \) under the above sampling scheme is

\[
\overline{\text{Var}}(\Delta) = \text{Var}(\hat{Y}^{(t)}) + \text{Var}(\hat{Y}^{(t-1)}) - 2 \text{Cov}(\hat{Y}^{(t)}, \hat{Y}^{(t-1)}),
\]

where
\[
\text{Var}(\hat{Y}^{(t)}) = \sum_h N_h^{(t)} \left( 1 - f_{ji} \right) s_h^{(t)}/n_h^{(t)},
\]
\[
\text{Cov}(\hat{Y}^{(t)}, \hat{Y}^{(t-1)}) = \sum_h N_h^{(t)} \left( 1 - f_{12,h} \right) s_h^{(1)(2)} n_{hc}/(n_h^{(1)} n_h^{(2)}),
\]
\[
\hat{Y}_t = n_{hc}^{(t)} \sum_{i \in G_{1t}} y_{it},
\]
\[
\hat{Y}_{t-1} = n_{hc}^{(t-1)} \sum_{i \in G_{1t-1}} y_{it},
\]
for \( t = 1, 2 \) and \( f_{12,h} = n_h^{(1)} n_h^{(2)}/n_h^{(1)} N_h \).

When \( U_h^{(t)} \neq U_h^{(t-1)} \) a more general form of Tam’s estimator is given by \( \overline{\text{Var}}(\Delta) \), except that
\[
\overline{\text{Var}}(\hat{Y}^{(t)}) = \sum_h N_h^{(t)} \left( 1 - f_{ji} \right) s_h^{(t)}/n_h^{(t)},
\]
\[ \text{Cov}(\hat{Y}^{(1)}, \hat{Y}^{(2)}) = \sum_h N_h^{(1)} N_h^{(2)} / (n_h^{(1)} n_h^{(2)}) n_h (1 - f_{12,h}) s_h^{(12)} \]

and

\[ f_{12,h} = n_h^{(1)} n_h^{(2)} N_h / n_h^{(1)} N_h^{(1)} N_h^{(2)} \]

For the reminder of this section we assume that \( \hat{\text{Var}}(\hat{\Delta}) \) is unbiased for \( \text{Var}(\Delta) \) when \( U_h^{(1)} \neq U_h^{(2)} \). (It is worthwhile noting that \( \hat{\text{Var}}(\Delta) \) can take negative values when \( U_h^{(1)} \neq U_h^{(2)} \). Nordberg (2000) gives an unbiased estimator of \( \text{Var}(\Delta) \) for the regression estimator when \( U_h^{(1)} \neq U_h^{(2)} \), but there is no obvious way in which it can be used with the Bootstrap as described in this paper.)

Estimating the variance of \( \hat{\Delta}_{\text{reg}} = \hat{Y}_{\text{reg}}^{(1)} - \hat{Y}_{\text{reg}}^{(2)} \), the movement between GREG estimates at times 1 and 2, using WOSB involves repeating the following \( R \) times:

(a) forming the set \( s^* \) by independently selecting \( m_{ch} = [n_{ch} / 2], m_{ch}^{(1)} = [n_{ch}^{(1)} / 2] \) and \( m_{ch}^{(2)} = [n_{ch}^{(2)} / 2] \) units by SRSWOR from the sets \( s_{ch}, s_{ch}^{(1)}, \) and \( s_{ch}^{(2)} \) respectively;

(b) for \( i \in s_{ch}^{(1)} \) calculate the replicate weights

\[ w_{hi}^{(1)} = N / n_h^{(1)} \left[ 1 - \gamma_{ch} n_{ch} / n_h^{(1)} - \gamma_{1ch} n_{1ch} / n_h^{(1)} + \gamma_{ch} m_{ch} \delta_{hi}^{(1)} \right] \]

for \( i \in s_{ch}, \)

\[ w_{hi}^{(2)} = \left[ 1 - \gamma_{ch} n_{ch} / n_h^{(2)} - \gamma_{1ch} n_{1ch} / m_{ch} + \gamma_{ch} m_{ch} \delta_{hi}^{(2)} \right] \]

for \( i \in s_{ch}^{(2)}, \) where

\[ \gamma_{1ch} = \sqrt{\left[ n_{ch} - f_{12,h} \right] m_{1ch} / \left[ m_{1ch} - n_{1ch} m_{ch} \right]}, \]

\[ \gamma_{ch} = \sqrt{\left[ 1 - f_{12,h} \right] m_{ch} / \left( n_{ch} - m_{ch} \right)} \]

and \( \delta_{hi}^{(1)} \) equals 1 if unit \( i \) is selected in the replicate group at time point 1 and zero otherwise;

(c) calculate weights defined analogously for \( i \in s_{ch}^{(2)}; \)

(d) calculate \( \hat{w}_{hi}^{(1)} = w_{hi}^{(1)} g_{hi}^{(1)} \) for \( i \in s_{ch}, \)

\( g_{hi}^{(1)} \) has the same form as \( g_{hi} \) but is calculated using the weights \( w_{hi}^{(1)} \) instead of \( w_{hi}^{(1)}; \)

(e) calculate \( \hat{\Delta}_{\text{reg}}^{*} = \hat{Y}_{\text{reg}}^{(1)} - \hat{Y}_{\text{reg}}^{(2)} \), where \( \hat{Y}_{\text{reg}}^{(1)} = \sum_{i \in s_{ch}} \hat{w}_{hi}^{(1)} y_i \). The WOSB variance estimator is given by

\[ \hat{\text{Var}}(\hat{\Delta}_{\text{reg}}) = (R - 1) I^{(2)} \sum_{i \in s_{ch}} (\hat{\Delta}_{\text{reg}}^{*} - \hat{\Delta}_{\text{reg}})^2, \]

where \( \hat{\Delta}_{\text{reg}} = \hat{Y}_{\text{reg}}^{(1)} - \hat{Y}_{\text{reg}}^{(2)} \) and \( \hat{Y}_{\text{reg}} = \sum_{i \in s_{ch}} \hat{w}_{hi}^{(1)} y_i. \)

The proof that \( \hat{\text{Var}}(\hat{\Delta}_{\text{reg}}) \) is unbiased is straightforward and is similar to the proof that \( \text{Var}(\hat{\theta}) \) is unbiased (see section 4).

The approach described above requires a separate set of replicate weights for movement and level variance estimates. Roberts, Kovaćević, Mantel and Phillips (2001) consider approximate Bootstrap variance estimators of movement that only use the level replicate weights, hence reducing computational costs and simplifying the method and its implementation in a computing system.

6. Simulation study

This section summarizes a simulation study for point-in-time and movement estimates carried out to empirically measure the bias and variability of WOSB and WSB over repeated sampling when \( R = 100 \). A population was generated at time points 1 and 2 from the following models, \( Y_{ij}^{(1)} = (0.75 x_{ij} + 0.25 x_{ij}) W(0, 2.5, 1) \) and \( Y_{ij}^{(2)} = 1.5 Y_{ij}^{(1)} W(0, 5, 1) \), where the auxiliary variables are given by \( x_{ij} \), \( W(0, 2.5, 1) \) and \( x_{ij} = 100L(0, 1, l) \) where \( W(\mu, \gamma, \alpha) \) and \( L(\mu, \gamma, \alpha) \) are the Weibull and Log-normal distributions with location, shape and scale parameters given by \( \mu, \gamma, \alpha \). These distributions reflect the long tails that are typical of economic survey data. The times 1 and 2 populations were of size 3,000, with 2,500 population units common to both time points. Each population unit, \( i, \) was assigned to one of 5 strata at both time points using \( z_i \), where \( z_i = x_i W(0, 2.5, 1) \) and the stratum boundaries were \( z_i = 50, 100, 150, 250. \) This resulted in stratum population sizes that ranged from 400 to 1,000.

A total of 3,000 simulated stratified SRSWOR were taken from the population at times 1 and 2, where \( n_h^{(1)} = 12, n_h^{(2)} = 4 \) and \( n_{ch} = 8 \) for all \( h \) and \( t = 1, 2 \). For WOSB the replicate sample sizes are given in sections 4 and 5. For WSB the replicate sample sizes for movements were \( m_{ch} = [n_{ch} - 1], m_{ch}^{(1)} = [n_{ch}^{(1)} - 1] \) and \( m_{ch}^{(2)} = [n_{ch}^{(2)} - 1] \) and for levels were \( m_{ch} = n_{ch}^{(1)} - 1 \). The WSB estimator for movements has the same form as WOSB but has a slightly different scaling factors and takes replicate samples with replacement.

From each of the 3,000 simulated samples \( \hat{Y}_{\text{reg}} \) is calculated, where \( \hat{Y}_{\text{reg}} \) is given by \( \hat{Y}_{\text{reg}} \) with \( x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)}) \), \( \sigma_j = 1 \) and \( j = 1, 2, ..., 3,000 \). The true standard error of \( \hat{Y}_{\text{reg}} \) is calculated by

\[ S = \sqrt{\frac{1}{3,000} \sum_{j=1}^{3,000} (\hat{Y}_{\text{reg}}^{(j)} - \bar{Y})^2}. \]
The Bootstrap’s estimated standard error of $\hat{Y}_{\text{reg}}$ from the $j^{th}$ sample is

$$
\hat{S}^j = \frac{1}{100} \sum_{r=1}^{100} (\hat{Y}_{\text{reg}}^{j(r)} - Y)^2,
$$

where $\hat{Y}_{\text{reg}}^{j(r)}$ is defined analogously to $\hat{Y}_{\text{reg}}^{r}$. The Relative Bias (RB) of the Bootstrap’s standard error is

$$
\text{RB}(\hat{S}) = \frac{1}{3,000} \sum_{j=1}^{3,000} (\hat{S}^j - \hat{S}).
$$

The Relative Root Mean Squared Error (RRMSE) of the Bootstrap’s estimated standard error is

$$
\text{RRMSE}(\hat{S}) = \frac{1}{3} \sqrt{\frac{1}{3,000} \sum_{j=1}^{3,000} (\hat{S}^j - \hat{S})^2}.
$$

Similar definitions for RRMSE and bias are used when estimating the movement variance. The 95% coverage probabilities, the percentage of 95% confidence intervals containing the true population total, of WOSB and WSB for levels and movement are also compared.

The results in Table 1 show that the RB and the RRMSE of the WOSB and WSB are both acceptably small. The bias of WSB’s time point 1 estimates are slightly higher than WOSB resulting in slightly worse coverage probabilities.

### Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>RB</th>
<th>RRMSE</th>
<th>C95%</th>
<th>Movement</th>
<th>RB</th>
<th>RRMSE</th>
<th>C95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOSB</td>
<td>0.7</td>
<td>17.3</td>
<td>94.7</td>
<td>21</td>
<td>20.7</td>
<td>95.3</td>
<td></td>
</tr>
<tr>
<td>WSB</td>
<td>-3.1</td>
<td>15.8</td>
<td>93.7</td>
<td>-1.3</td>
<td>19.6</td>
<td>94.6</td>
<td></td>
</tr>
</tbody>
</table>

7. Summary

From the simulation results, both the WOSB and WSB were considered to be reliably accurate over repeated sampling. Conditional on the sample, the WOSB was found to be significantly more efficient (up to 50%) than WSB for stratified sampling when the stratum sample size is sometimes small. As a result, the WOSB was implemented in ABEST.

### References


Bayesian estimation in small areas when the sampling design strata differ from the study domains

Jacob J. Oleson, Chong Z. He, Dongchu Sun and Steven L. Sheriff

Abstract

The purpose of this work is to obtain reliable estimates in study domains when there are potentially very small sample sizes and the sampling design stratum differs from the study domain. The population sizes are unknown as well for both the study domain and the sampling design stratum. In calculating parameter estimates in the study domains, a random sample size is often necessary. We propose a new family of generalized linear mixed models with correlated random effects when there is more than one unknown parameter. The proposed model will estimate both the population size and the parameter of interest. General formulae for full conditional distributions required for Markov chain Monte Carlo (MCMC) simulations are given for this framework. Equations for Bayesian estimation and prediction at the study domains are also given. We apply the 1998 Missouri Turkey Hunting Survey, which stratified samples based on the hunter’s place of residence and we require estimates at the domain level, defined as the county in which the turkey hunter actually hunted.

Key Words: Hierarchical Bayes; Markov chain Monte Carlo; Bayesian prediction; Random sample size; Spatial correlation; Stratification.

1. Introduction

Small sample sizes occur often when analyzing sample survey data. These small sample sizes arise frequently when studying subpopulations such as socio-demographic groups. We may also consider spatial regions and time periods as subpopulations or study domains. Due to small sample sizes, direct survey estimators could be highly unreliable. Estimation stemming from areas with small sample sizes has been termed small area estimation (SAE). Rao (2003) gives a nice review of many SAE techniques. Some recent small area review papers are found in Rao (2005) and Jiang and Lahiri (2006).

Appropriate models are needed in order to produce reliable small area statistics. Different model-based methods include the empirical best prediction method (see Prasad and Rao 1990, Jiang, Lahiri and Wan 2002, Das, Jiang and Rao 2004, Jiang and Lahiri 2006) and Bayesian methods (see Malec, Sedransk, Morairity and LeClere 1997, Ghosh, Natarajan, Stroud and Carlin, 1998, He and Sun 2000). For a good review on Bayesian small area estimation, the readers are referred to Rao (2003). This paper concerns a practical implementation of Bayesian methodology. One critical step in implementing a Bayesian model is selecting prior distributions. The propriety of the posterior distribution and the robustness of the priors should be carefully examined. When an MCMC simulation such as Gibbs sampler is used in the computation, the convergence of the Gibbs chain must be monitored. For more details, see Carlin and Louis (2000).

We consider a stratified cluster sampling design where within each stratum clusters are selected by simple random sampling without replacement. In our application, clusters are of unequal sizes and the cluster sizes are unknown at the time of designing the survey. We consider a domain estimation problem where domains cut across the design clusters. As a result, domain population size is unknown and domain sample size is random. In our application, realized sample sizes for the domains are small and as such standard design-based domain estimation techniques (see Cochran 1977, Lohr 1999) are unreliable. We propose a fully Bayesian hierarchical model to get around the problem.

We begin by obtaining estimates of success rates and population sizes simultaneously at the small area level for individuals from each of the design strata. The estimates are obtained by borrowing information from the neighboring small areas. This is done through a spatial structure built into the Bayesian model. Therefore, the resulting estimates are much more stable than the direct survey methods. We then compute a weighted average of the success rates from the design stratum for the final small area estimate. For example, if a county is the small area, we compute a county-specific success rate estimate for each of the design strata and average them for a single county estimate. To combine the sampling design strata, the individual stratum population sizes should be known. In the case where these population sizes are not known, they can be estimated using our proposed model. This work is motivated by applying Bayesian methods in estimating the turkey hunting success rates at the county level in Missouri. We propose a new
family of generalized linear mixed models with correlated random effects when there is more than one unknown parameter. We will call this a bivariate generalized linear mixed model (bivariate GLMM). A generalized linear mixed model (GLMM) with possible correlated random effects is often used when there is only one unknown parameter (Sun, Speckman and Tsutakawa 2000). The proposed model estimates both the population size and the parameter of interest and, hence, advances the current state of small area research.

The above-described scenario is present in the 1998 Missouri Turkey Hunting Survey (MTHS). This was a spring postseason mail survey that provided the Missouri Department of Conservation with information concerning the number of turkeys harvested by hunters on each day of the hunting season and in which county the harvest occurred. Also, the total number of trips made to the counties by these hunters on each hunting day was recorded. Hunting success rates were then calculated from this information.

The MTHS example is presented in detail in Section 2. It is followed in Section 3 by a summary of the proposed methodology and by general formulae. Although we estimate success rates, the methodology is generalizable. We also give general formulae to find the estimates and predictions for the small areas as well as full conditional distributions for use with MCMC simulations. Final comments are given in Section 4.

2. 1998 Missouri turkey hunting survey

2.1 Background to 1998 MTHS

The Missouri Department of Conservation began biennially in 1986 to track hunter tendencies with the MTHS. This survey asked the hunter what county he/she hunted in, on what day that occurred, and if the hunt was successful or not. It began as a simple random sample of all spring turkey hunting permit holders. He and Sun (1998) used the information from the 1996 survey to estimate turkey hunter success rates in all 114 counties of Missouri with a Bayesian Beta-Binomial model. He and Sun (2000) estimated county-specific hunting success rates per week of the hunting season. Only one harvested turkey was allowed per week in 1996 spring hunting season. They used a GLMM and estimated only success rates. Oleson and He (2004) extended this model in order to estimate hunting success rates for each day of the hunting season. They found significant auto-correlation among the days of the hunting season and among the counties of Missouri when estimating success rates.

In 1998, the MTHS sampling scheme was changed. The frame is still the list of all turkey hunters registered to hunt in Missouri and contains, among other things, information on each hunter’s county of residence. Simple random samples used previously put too much weight on the heavy population masses of Kansas City and St. Louis. Hence, counties near these metropolises received large samples and counties further away (e.g., the southern tiers) received insufficient samples. Ideally we would like to stratify by the county where hunters pursued turkeys to draw samples representative at this domain level. Information about where the hunters pursued turkeys is unavailable until after questionnaires are returned, meaning this type of stratification is not possible.

An alternative is to stratify by where the hunter lives since hunters tend to hunt near where they live (in locations with which they are familiar). This causes a problem in estimating the parameters of interest which are hunting success rates per county. We would like estimates based on hunting location, but the sampling design uses the hunter’s place of residence. The new sampling design of MTHS is a stratified simple random sampling of clusters with unequal sizes. In this case, a cluster represents a registered hunter and its elements are the hunting trips for that hunter. The hunter’s place of residence is used as a stratification factor. The design strata are: 1) Non-residents of Missouri, 2) Residents of Northern Missouri, 3) Residents of Southern Missouri, 4) Residents of St. Louis metro area, and 5) Residents of Kansas City metro area. Figure 1 shows the boundaries that were used in determining the four Missouri resident strata. These are based on the first three digits of the postal zip code. Proportional allocation was used to determine the number of sampled hunters in each stratum. As shown in Table 1, there were 110,691 total permit holders. A sample of 8,000 was proportionally allocated to each of the five strata. In Table 1 the column “% of Sample” refers to the percent of the overall sample that comes from that particular stratum (sample replied/total sample replied). The column “% of Strata Sampled” contains the percent of hunters who were sampled from that stratum (sample replied/total permits). The number of clusters in the population and the number of clusters in the sample for each stratum are known fixed numbers. The hunting trips taken by each hunter are the elements within the cluster. Throughout the remainder of the paper, population size is the number of hunting trips taken by all hunters in the frame, termed hunting pressure. For the MTHS example, the population size is unknown but not random; the sample size is best considered random since we do not know how many trips each hunter will take. Note that a hunter does not have to hunt in his/her design stratum. The hunter may hunt in many different counties during the season. For each study
domain \((i.e., \text{county})\), the population size is again unknown but not random; the sample size is again random where we do not know in which county a hunter is going to hunt. Small sample sizes can still result; for example, Dunklin County, Pemiscot County, and New Madrid County in southeast Missouri had zero trips reported to them throughout the 3-week season in 1998 and Lawrence County in southwest Missouri had zero trips on the first day and only five the remainder of the first week.

Table 1 Sample sizes and response rates from 5 strata for 1998 MTHS

<table>
<thead>
<tr>
<th>Design Stratum</th>
<th>Total Permits</th>
<th>Sample Sent</th>
<th>Sample Replied</th>
<th>Response Rate (%)</th>
<th>% of Sample</th>
<th>% of Strata Sampled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Resident</td>
<td>19,798</td>
<td>1,600</td>
<td>1,180</td>
<td>73.75</td>
<td>22.2</td>
<td>5.96</td>
</tr>
<tr>
<td>North</td>
<td>21,756</td>
<td>1,532</td>
<td>975</td>
<td>63.64</td>
<td>18.3</td>
<td>4.48</td>
</tr>
<tr>
<td>Missouri</td>
<td>34,375</td>
<td>2,421</td>
<td>1,509</td>
<td>62.33</td>
<td>28.4</td>
<td>4.39</td>
</tr>
<tr>
<td>South</td>
<td>19,959</td>
<td>1,405</td>
<td>969</td>
<td>68.97</td>
<td>18.2</td>
<td>4.85</td>
</tr>
<tr>
<td>St. Louis</td>
<td>14,803</td>
<td>1,042</td>
<td>688</td>
<td>66.03</td>
<td>12.9</td>
<td>4.65</td>
</tr>
<tr>
<td>Kansas City</td>
<td>14,803</td>
<td>1,042</td>
<td>688</td>
<td>66.03</td>
<td>12.9</td>
<td>4.65</td>
</tr>
<tr>
<td>Total</td>
<td>110,691</td>
<td>8,000</td>
<td>5,321</td>
<td>66.51</td>
<td>100</td>
<td>4.81</td>
</tr>
</tbody>
</table>

The MTHS is a post-season mail survey where the questionnaires are mailed to the hunter’s listed residence. The survey has questions specifying each day of the hunting season and asking hunters whether or not they hunted, in what county they hunted, and if their hunt was successful or not. In total 41% of the residents responded to the first mailing. Those who did not respond within two months were mailed the survey again, to which 26% responded from the second mailing. The survey was mailed a third time to those who had still not responded after two months with 20% of the third mailing responding. This resulted in an overall response rate of 66.5%. From Table 1 we see that non-residents of Missouri had the highest response rate at nearly 75%. The two metro areas of St. Louis and Kansas City then followed with response rates of 69% and 66%, respectively. The rural areas in northern Missouri and southern Missouri had the lowest response rates at approximately 63%.

The 1998 spring turkey hunting season in Missouri consisted of the 21 consecutive days beginning on Monday, 19 April. During the first week, a turkey hunter could harvest one bearded turkey. If successful during the first week, the hunter was allowed to harvest only one additional bearded turkey during the last two weeks of the season. If the hunter was unsuccessful during the first week, the hunter could harvest two bearded turkeys during the second two weeks, but only one bird per day could be taken. During this 3-week season, the turkey biologist wanted information concerning hunter success for the first day due to the opening day effect in the hunting season when hunting pressure normally exceeds that of any other day of the season. The remaining six days of the first week are combined into a second time period of interest, called week 1. Week 2 and week 3 are modeled separately to give a total of four time periods.
The small area of interest is the county-time period combination for turkey hunters from each stratum. Turkey population management focuses on the county level as the smallest areal unit of interest due to hunters’ ability to indicate where they have hunted in reporting their success. A hunting license is required to hunt, thus we know how many total hunters there are in Missouri and in which stratum they live. Spreading the data out across 114 counties produces very sparse data within each of the counties. For this reason, rather than looking at success rates on a daily basis, we will use four time periods as defined above. In calculating the county-specific success rates, however, we use the number of hunting trips in each county. We do not know how many trips each hunter will take or in which county they will hunt. Thus, while we know the number of hunters in each stratum, the number of trips is unknown. Furthermore, while we know the number of hunters who have been sampled from each stratum, the number of trips to a specific county is unknown and random because a different sample of hunters will yield a different number of trips. There are two principle parameters to estimate and one random variable to predict. The first parameter is the total number of hunting trips taken by all hunters, known as hunting pressure. This is important to wildlife managers who are concerned about the quality of the hunting experience. Too many hunters in an area have the tendency to interfere with each other’s hunt, which in turn lowers the quality of the hunting experience. The second parameter to estimate is the hunting success rate, or the proportion of turkeys harvested per hunting trip. If there are not enough turkeys in a county, wildlife managers will close the county for a number of years in order for the turkey population to recover. The random variable to predict is the total number of turkeys harvested. In Missouri, every turkey that is harvested must be reported. In 1998, turkey hunters were required to go to a check station where their turkeys were tallied. It was expensive to maintain these check stations. One of purposes of this project is to predict the total number of turkeys harvested at the county level and compare this to the actual number recorded at the check stations.

Next we present a model to estimate hunting success rate and hunting pressure, as well as predict total turkey harvest simultaneously. The model accounts for random sample sizes that the previous models of He and Sun (1998), He and Sun (2000), and Oleson and He (2004) did not.

2.2 The Model

Let \( p_{ijk} \) denote the success rate, \( y_{ijk} \) the number of successes, \( n_{ijk} \) the number of trips in the sample, and \( N_{ijk} \) the total number of trips, in county \( i \), at time \( j \), for individuals from stratum \( k \). We model turkey harvests with independent binomial distributions

\[
\begin{align*}
(y_{ijk} \mid n_{ijk}, p_{ijk}) & \sim \text{Binomial}(n_{ijk}, p_{ijk}), \\
(n_{ijk} \mid N_{ijk}) & \sim \text{Poisson}(R_k N_{ijk}).
\end{align*}
\]

where, \( i = 1, ..., I \) is the county (i.e., study domain), \( j = 1, ..., J \) is the time period, and \( k = 1, ..., K \) is the design stratum. Here \( I = 114, J = 4, K = 5 \) for MTHS. The previous analyses of the MTHS assumed a fixed sample size, \( n_{ijk} \), which we believe is best considered random. We don’t know the number of trips made to each county until after the survey has been collected. We also don’t know the total number of hunting trips taken in the small area, \( N_{ijk} \), only the number of potential hunters. Since hunters must stop after their second turkey, counties with higher success rates would be expected to have fewer hunting trips (or days) for the same number of harvested birds than a county with lower success rates. If there is a correlation, then the sample size must be considered random. Also, in Bayesian hierarchical modeling, if the distribution of a sample size is independent of the distribution of the response variable, then the estimates are identical for random and non-random (fixed) sample size \( n \), see Durbin (1969). The estimated success rate is smoother for a fixed \( n_{ijk} \) than when it is random (Woodard, He and Sun 2003). Malec et al. (1997) applied the Bayesian small area estimation to the National Health Interview Survey. There are two major differences between our model and that of Malec et al. (1997). First, the population sizes, \( N_{ijk} \), are known in their model but unknown in our model. This is the main reason that we introduce the bivariate GLMM. Secondly, the logit of the success rate is modeled as a linear function of covariates in their model. Therefore, the estimates depend on the values of covariates but not the spatial locations. We will add a spatial component to the logit of success rates in addition to the covariates so that the estimates depend on both covariates and spatial locations. This is necessary if some important covariates are not available.

To incorporate the randomness of the sample sizes, we model \( n_{ijk} \) with Poisson distributions

\[
(n_{ijk} \mid N_{ijk}) \sim \text{Poisson}(R_k N_{ijk}).
\]

The mean and variance of the Poisson distribution for \( n_{ijk} \) is a constant multiplied by the population size, \( N_{ijk} \). This constant \( R_k \) is the ratio of the number of hunters in the sample for stratum \( k \) to the total number of hunters from stratum \( k \). This ratio can be calculated from Table 1.

For the Poisson distribution, the overall sample size, \( n_k \) is considered random. We presented in the previous section why we consider this assumption to be appropriate. If \( n_k \) were fixed, then the multinomial distribution would be a more appropriate model. The likelihoods of these two
approaches are very similar and yield comparable results in either context (see Agresti 2002, pages 8-9).

We model $p_{ijk}$ using its logit function and $N_{ijk}$ using a logarithm transformation, i.e.,

$$\eta_{ijk} = \log \left( \frac{p_{ijk}}{1 - p_{ijk}} \right), \quad \omega_{ijk} = \log (N_{ijk}).$$

Linear mixed models are used for the priors on both $\eta_{ijk}$ and $\omega_{ijk}$ by assuming

$$\eta_{ijk} = \theta_{1i} + u_{1ik} + e_{ijk},$$

$$\omega_{ijk} = \theta_{0j} + u_{2jk} + e_{2jk}.$$  

Here for $a = 1, 2$, $\theta_{ai}$ denotes fixed effect due to the $a$th time in stratum $i$, $u_{ai}$ represents a random county effect, and random errors $e_{ai}$ are iid $N(0, \delta_{ai})$.

To complete the Bayesian hierarchical model, we need to specify the priors for $\theta_{0j} = (\theta_{0i1}, ..., \theta_{0iJ})$, $u_{ai} = (u_{ai1}, ..., u_{aiJ})$, and $\delta_{ai}$. He and Sun (2000) and Oleson and He (2004) show that there is significant spatial correlation among counties of Missouri in estimating the success rates. They use a conditional auto-regressive (CAR) structure to model spatial dependence between neighboring counties.

The joint density of $\{u_{ai}\}$ is

$$f(u_{ai}) = (2\pi)^{\frac{J}{2}} | I - \rho_{ai}C |^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2 \delta_{ai}} u_{ai}^t (I - \rho_{ai}C) u_{ai} \right\}.$$  

Here $\delta_{ai}$ and $\rho_{ai}$ are positive definite matrices and $I - \rho_{ai}C$ is a nonnegative definite symmetric matrix (Besag 1974). $I$ is the $J \times J$ identity matrix and $C$ is an adjacency matrix whose component $c_{ij}$ is 1 if areas $i$ and $j$ share a common boundary and $c_{ij} = 0$. We also define $\rho_{ai}$ and $\delta_{ai}$ to be the spatial correlation parameters. Let $\lambda_i \leq ... \leq \lambda_J$ be the eigenvalues of the matrix $C$.

These matrices, $I - \rho_{ai}C$, are positive definite if $\lambda_{1i} \leq \rho_{ai} \leq \lambda_{Ji}$ (Clayton and Kaldor 1987). For the Missouri data, $I = 114$ and the numerical values of $\lambda_i$ and $\lambda_{114}$ are -2.8931 and 5.6938, respectively. This means that the density of $u_{ai}$ exists if $\rho_{ai}$ is in (-0.3457, 0.1756).

For the remaining priors, we assume the following. $\theta_{ai}$ is normal with mean $\mu_{ai}$ and variance $\tau_{ai}^2$. Let $\rho_{ai}$ be uniformly distributed on the interval $[\lambda_{1i}, \lambda_{Ji}]$. Finally, a common prior for the variance components $\delta_{ai}$ and $\delta_{ai}^2$ is inverse gamma (Gelman, Carlin, Stern and Rubin 1995) whose densities are proportional to

$$\frac{1}{\delta_{ai}^{\delta_{ai}^2 + 1}} \exp \left( -\frac{\delta_{ai}^2}{\delta_{ai}^2} \right) \quad \text{and} \quad \frac{1}{\delta_{ai}^{\delta_{ai} + 1}} \exp \left( -\frac{\delta_{ai}}{\delta_{ai}} \right),$$

respectively.

To evaluate the posterior distribution, we apply MCMC methods such as Gibbs sampling (Gelfand and Smith 1990) to obtain samples from the posterior distribution. For an overview of MCMC methodologies see Gelman et al. (1995), Gilks, Richardson and Spiegelhalter (1996) and Robert and Casella (1999). The full conditional distributions required for this evaluation may be found directly from those given in Fact 1 in Section 3. Most of these conditional distributions ($\theta_{ai}, u_{ai}, \delta_{ai}, \delta_{ai}^2$) are of standard forms and are easily sampled from. Other full conditional distributions ($\eta_{ijk}, \omega_{ijk}$ and $\rho_{ai}$) have log-concave densities for the MTHS. The adaptive rejection sampling method of Gelks and Wild (1992) can be used to generate random samples from log-concave densities.

### 2.3 Bayesian estimation and prediction

At this point, we have obtained estimates of $(p_{ijk}, N_{ijk})$. We wish to pool the estimates from the stratum together in order to estimate $p_j$ and $N_j$. We will also predict the unobserved number of harvests in county $i$, denoted $h_{ij}$.

To obtain the estimates of $(p_{ij}, N_{ij})$, let $(p_{ij}^{(l)}, N_{ij}^{(l)})$, $l = 1, ..., L$ be the output from the sampled Gibbs chain after the burn-in sample. Define

$$p_{ij} = \frac{\sum_{l=1}^{k} p_{ij}^{(l)} N_{ij}^{(l)}}{\sum_{l=1}^{k} N_{ij}^{(l)}}, \quad l = 1, ..., L.$$  

The posterior mean and variance of $p_{ij}$ can then be approximated by

$$\hat{E}(p_{ij}) = \frac{1}{L} \sum_{l=1}^{L} p_{ij}^{(l)}$$

and

$$\hat{V}(p_{ij}) = \frac{1}{L-1} \sum_{l=1}^{L} (p_{ij}^{(l)})^2 - \frac{1}{L-1} \left( \hat{E}(p_{ij}) \right)^2,$$

respectively. Similarly define $N_{ij}^{(l)} = \sum_{k=1}^{K} N_{ijk}^{(l)}$. The posterior mean and variance of $N_{ij}$ can be approximated from the MCMC simulation output as well.

We now focus on Bayesian prediction of the unobserved harvest by their posterior predictive distributions. We have that $y_{ijk}$ represents the number of harvested turkeys in county $i$, of time period $j$, and of sampling stratum $k$ for those in the sample, whereas for those not in the sample, we let the number of harvested turkeys in county $i$, during time $j$, stratum $k$ be represented as $y_{ijk}^*$. Thus, the number of turkeys harvested in county $i$ at time $j$ for hunters from stratum $k$ is $h_{ijk} = y_{ijk} + y_{ijk}^*$. Here $y_{ijk}^*$ is a known value and we need only find $y_{ijk}^*$. We may think of $y_{ijk}^*$ given $(n_{ijk}, N_{ijk}, p_{ijk})$ as a binomial random variable in the form of (1). Thus

$$y_{ijk}^* | n_{ijk}, N_{ijk}, p_{ijk}) \sim \text{Binomial}(N_{ijk} - n_{ijk}, p_{ijk}).$$

Let $(p_{ijk}^{(l)}, N_{ijk}^{(l)}), l = 1, ..., L$, be the output from running a Gibbs chain after a burn-in sample. The predictive mean of $y_{ijk}^*$ given data $d = \{y_{ijk}, n_{ijk}\}$ : $i = 1, ..., I$, $j = 1, ..., J$, $k = 1, ..., K_i$ is then...
Finally we write
\[ \hat{h}_{jk} = y_{jk} + \hat{E}(y_{jk} | d), \]
\[ \hat{h}_i = \sum_{j=1}^{K} \sum_{k=1}^{K} \hat{h}_{jk} . \]  
(8)

The variability of \( h_i \) is given by
\[ \hat{V}(h_i | d) = \frac{1}{L - 1} \sum_{l=1}^{L} \left( \sum_{j=1}^{K} \sum_{k=1}^{K} (N_{ijk} - n_{ijk}) p_{ijk}^{(l)} (1 - p_{ijk}^{(l)}) \right) \]
\[ + \frac{1}{L - 1} \left( \sum_{j=1}^{K} \sum_{k=1}^{K} \left[ y_{ijk} + (N_{ijk} - n_{ijk}) p_{ijk}^{(l)} - \hat{h}_{ijk} \right] \right)^2 . \]

The proof is a straightforward application of conditional expectations. Note that \( h_{ijk} \) is defined as a random variable. It does not equal \( N_{ijk} p_{ijk} \) just as \( y_{ijk} \) does not equal \( n_{ijk} P_{ijk} \). Therefore, one should not simply use \( N_{ijk} P_{ijk} \) to estimate \( h_{ijk} \) although it might be very close if \( N_{ijk} \) is large enough.

2.4 Model fitting

Fifteen–thousand iterations were run for the Gibbs chain, of which 5,000 were discarded as a burn-in sequence. Thus, posterior estimates are based on 10,000 autocorrelated samples from the posterior distribution. In monitoring the convergence of Gibbs sampling, we have used the diagnostics of Heidelberger and Welch (1983) as well as graphical monitoring of the sample paths. Convergence diagnostics and posterior summaries were performed with the BOA software (Smith 2005). We let variance components \( \delta_{ijk}^{(u)} \) and \( \delta_{ijk}^{(v)} \) follow a non-informative IG(2,1) prior which gives a mean of one and infinite variance. We give data-dependent priors to \( \theta_{ijk} \sim N(-1.5,16) \) and \( \theta_{2,ijk} \sim N(4,25) \). To obtain a data-dependent prior, we modeled the following two steps because a completely arbitrary data-dependent prior might lead to unreliable posterior estimates (Wasserman 2000). First, we began with noninformative (but conjugate) priors to obtain the posterior means and standard deviations for each of these parameters through MCMC simulations. We then set data-dependent prior means to be close to the posterior means of \( \theta_{ijk} \) and \( \theta_{2,ijk} \) and the data-dependent prior standard deviations to be about ten times the posterior standard deviations, after the posterior estimates were obtained using the noninformative priors. The estimates using these two approaches were quite similar, but the model using the data-dependent priors gave smaller variances.

We fit many simplified models for comparison and model checking. As a model selection tool we use the Deviance Information Criterion (DIC) suggested by Spiegelhalter, Best, Carlin and van der Linde (2002). DIC is a generalization of the Akaike Information Criterion that measures the deviance over an MCMC run and includes a penalized fit measure taking into account the model dimension. Smaller values of DIC indicate a better-fitting model. Our model with data-dependent priors had a DIC = 13,910.5. No alternate models had a significantly reduced DIC value from this model. The model with non-informative priors on \( \theta_{ijk} \) and \( \theta_{2,ijk} \) had DIC = 14,700.4. A reduced model with common correlation parameters \( \rho_{11} = \rho_{12} = \rho_{13} = \rho_{14} = \rho_{15} = \rho_{21} = \rho_{22} = \rho_{23} = \rho_{24} = \rho_{25} \) had DIC = 13,896.9.

As another model check, we have calculated statewide averages of the estimates using both the simple naive design-based estimates and Bayesian estimates which follow in Table 2. At the statewide level, sample sizes are large enough to consider the design-based estimates reliable. The statewide design-based estimates and model estimates match closely. Thus, the Bayes estimator performs well in terms of the design consistency property (see You and Rao 2003 and Jiang and Lahiri 2006). We note that the first day estimates are slightly lower due to smoothing, but the success rate estimates are still much higher for the first day than any other time period.

<table>
<thead>
<tr>
<th>Design Stratum</th>
<th>Success Rate Estimates for 1998 MTHS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 1</td>
</tr>
<tr>
<td>Non-Resident</td>
<td>105</td>
</tr>
<tr>
<td>Week 1</td>
<td>271</td>
</tr>
<tr>
<td>Week 2</td>
<td>198</td>
</tr>
<tr>
<td>Week 3</td>
<td>75</td>
</tr>
<tr>
<td>North</td>
<td>82</td>
</tr>
<tr>
<td>Week 1</td>
<td>211</td>
</tr>
<tr>
<td>Week 2</td>
<td>161</td>
</tr>
<tr>
<td>Week 3</td>
<td>99</td>
</tr>
<tr>
<td>South</td>
<td>138</td>
</tr>
<tr>
<td>Week 1</td>
<td>224</td>
</tr>
<tr>
<td>Week 2</td>
<td>186</td>
</tr>
<tr>
<td>Week 3</td>
<td>90</td>
</tr>
<tr>
<td>St. Louis</td>
<td>54</td>
</tr>
<tr>
<td>Week 1</td>
<td>91</td>
</tr>
<tr>
<td>Week 2</td>
<td>85</td>
</tr>
<tr>
<td>Week 3</td>
<td>45</td>
</tr>
<tr>
<td>Kansas City</td>
<td>52</td>
</tr>
<tr>
<td>Week 1</td>
<td>95</td>
</tr>
<tr>
<td>Week 2</td>
<td>92</td>
</tr>
<tr>
<td>Week 3</td>
<td>55</td>
</tr>
<tr>
<td>Overall</td>
<td>431</td>
</tr>
<tr>
<td>Week 1</td>
<td>892</td>
</tr>
<tr>
<td>Week 2</td>
<td>722</td>
</tr>
<tr>
<td>Week 3</td>
<td>364</td>
</tr>
<tr>
<td>Total</td>
<td>2,409</td>
</tr>
</tbody>
</table>
2.5 Data analysis for MTHS

Posterior means and standard deviations for parameters under all design strata are listed in Table 3. We note that the mean and variance estimates of \( \hat{\theta}_{12k} \) and \( \hat{\theta}_{23k} \) as well as \( \hat{\theta}_{32k} \) and \( \hat{\theta}_{33k} \) for \( k = 1, \ldots, 5 \) are approximately equal. This gives reason to believe that success rate estimates and hunting pressure estimates are similar for week 1 and 2 of the hunting season. There remains a difference in the first and week 3 of the hunting season, though (Vangilder, Sheriff and Olsen 1990, Kimmel 2001).

Table 3 Posterior means (Standard deviations) of model parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Non-Res</th>
<th>North</th>
<th>South</th>
<th>St. Louis</th>
<th>K.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\delta}_{11} )</td>
<td>0.133 (0.034)</td>
<td>0.117 (0.030)</td>
<td>0.136 (0.036)</td>
<td>0.164 (0.051)</td>
<td>0.143 (0.040)</td>
</tr>
<tr>
<td>( \hat{\delta}_{12} )</td>
<td>0.096 (0.017)</td>
<td>0.044 (0.007)</td>
<td>0.041 (0.006)</td>
<td>0.061 (0.011)</td>
<td>0.060 (0.011)</td>
</tr>
<tr>
<td>( \hat{\delta}_{13} )</td>
<td>0.146 (0.040)</td>
<td>0.148 (0.041)</td>
<td>0.140 (0.038)</td>
<td>0.166 (0.053)</td>
<td>0.186 (0.059)</td>
</tr>
<tr>
<td>( \hat{\delta}_{14} )</td>
<td>0.570 (0.092)</td>
<td>0.511 (0.081)</td>
<td>0.633 (0.094)</td>
<td>0.373 (0.062)</td>
<td>0.356 (0.063)</td>
</tr>
<tr>
<td>( \hat{\theta}_{11} )</td>
<td>-1.209 (0.131)</td>
<td>-1.530 (0.130)</td>
<td>-1.244 (0.135)</td>
<td>-1.420 (0.159)</td>
<td>-1.472 (0.160)</td>
</tr>
<tr>
<td>( \hat{\theta}_{12} )</td>
<td>-1.613 (0.109)</td>
<td>-1.748 (0.111)</td>
<td>-1.867 (0.130)</td>
<td>-1.983 (0.147)</td>
<td>-1.928 (0.150)</td>
</tr>
<tr>
<td>( \hat{\theta}_{13} )</td>
<td>-1.801 (0.113)</td>
<td>-2.081 (0.113)</td>
<td>-2.152 (0.129)</td>
<td>-2.032 (0.148)</td>
<td>-2.007 (0.148)</td>
</tr>
<tr>
<td>( \hat{\theta}_{14} )</td>
<td>-1.830 (0.132)</td>
<td>-2.135 (0.123)</td>
<td>-2.354 (0.138)</td>
<td>-2.115 (0.156)</td>
<td>-2.008 (0.153)</td>
</tr>
<tr>
<td>( \hat{\theta}_{21} )</td>
<td>3.368 (0.098)</td>
<td>3.333 (0.098)</td>
<td>3.297 (0.087)</td>
<td>3.305 (0.088)</td>
<td>3.308 (0.098)</td>
</tr>
<tr>
<td>( \hat{\theta}_{22} )</td>
<td>4.582 (0.097)</td>
<td>4.361 (0.093)</td>
<td>4.371 (0.084)</td>
<td>4.349 (0.081)</td>
<td>4.238 (0.090)</td>
</tr>
<tr>
<td>( \hat{\theta}_{23} )</td>
<td>4.404 (0.097)</td>
<td>4.429 (0.093)</td>
<td>4.459 (0.088)</td>
<td>4.293 (0.083)</td>
<td>4.229 (0.093)</td>
</tr>
<tr>
<td>( \hat{\theta}_{24} )</td>
<td>3.693 (0.101)</td>
<td>4.093 (0.095)</td>
<td>4.039 (0.086)</td>
<td>3.934 (0.082)</td>
<td>3.891 (0.094)</td>
</tr>
<tr>
<td>( \hat{\rho}_{1} )</td>
<td>0.153 (0.019)</td>
<td>0.103 (0.071)</td>
<td>0.163 (0.011)</td>
<td>0.158 (0.016)</td>
<td>0.125 (0.068)</td>
</tr>
<tr>
<td>( \hat{\rho}_{2} )</td>
<td>0.168 (0.007)</td>
<td>0.172 (0.003)</td>
<td>0.171 (0.004)</td>
<td>0.172 (0.004)</td>
<td>0.172 (0.004)</td>
</tr>
</tbody>
</table>

We also point out the correlation parameter estimates of \( \rho_{12} \) and \( \rho_{23} \). The estimates for \( \rho_{23} \) for the hunting pressure are all near their upper boundary, suggesting a strong relationship between the counties when estimating the number of hunting trips taken. Most of the estimates for \( \rho_{1k} \) are more than two standard deviations away from zero as well.

The simple naive design-based success rate estimates for each county are in the first row of Figure 2. Design-based estimates in the counties range from 0 to 0.55 with some counties having no estimate because \( n_{ij} = 0 \). Alternatively, the Bayesian estimates for the success rate in county \( i \) for time \( j \) are plotted in the second row of Figure 2. The Bayesian model success rate estimates produce a much more sensible range from 0.03 to 0.30. We don’t expect any county to have a success rate estimate of 0. Also, turkeys are quite difficult to hunt and a high success rate is not sensible (He and Sun 1998, He and Sun 2000, Oleson and He 2004). Thus a lower value, such as that produced from the model, makes more intuitive sense. Standard deviations for the model success rate estimates are given in the third row of Figure 2.

Success rate estimates tend to decrease over the course of the hunting season. In addition, the highest success rate estimates are in the northern portion of Missouri. This was also shown to be true in previous analyses of the MTHS using 1996 data (He and Sun 2000, Oleson and He 2004). We note that the highest success rate estimates are not in the same counties, though. The highest estimated rates from 1998 have moved slightly to the east of where they were in 1996. This shift is due to a time trend as noted by conservationists.

We also produce estimates of the population size \( N_{ijk} \). The model-based estimates of \( N_{ij} \) are plotted in the first row of Figure 3. The standard errors for the model-based estimates are plotted in the second row of Figure 3. The values plotted for weeks 1, 2, and 3 are the daily averages for those weeks. It is apparent that more people were hunting on the first day than any other day of the hunting season.

We plot the actual check station data as the first map in Figure 4. Using formula (8) we predict the model-based total harvest \( \hat{h}_i \) in the second map of the first row. From these figures, the model-based predictions look to be reasonably accurate. We expect to see a small amount of overprediction as is shown by comparing the plots. Some hunters may underestimate how often they went out to hunt and overestimate the number of birds they harvested. The overcount may possibly also be attributed to turkeys harvested but not reported at a check station. One more explanation could be that the hunters returning the survey are those who were more successful, and those not returning the survey were less successful or did not hunt. Comparing the predictions from the model to the actual harvest numbers is a confirmation that the model is appropriate and gives accurate predictions. Also, many states do not have check stations and this shows that they could obtain appropriate estimates at the domain level even with a small sample at the domain level.
SUCCESS RATES

Day 1  
Week 1  
Week 2  
Week 3  

0.0 0.2 0.4 0.6  

No data  

STANDARD DEVIATIONS OF SUCCESS RATES

Day 1  
Week 1  
Week 2  
Week 3  

0.0 0.1 0.2 0.3  

STANDARD DEVIATIONS OF NUMBER OF TRIPS

Day 1  
Week 1  
Week 2  
Week 3  

0 400 800 1200  

Figure 2  Hunting success rates from stratified 1998 MTHS. Row 1 - design based estimates of success rates. Row 2 - Bayesian estimates of success rates. Row 3 - Standard deviations of Bayesian estimates of success rates.

NUMBER OF TRIPS

Day 1  
Week 1  
Week 2  
Week 3  

0 400 800 1200  

STANDARD DEVIATIONS OF NUMBER OF TRIPS

Day 1  
Week 1  
Week 2  
Week 3  

0 40 80 120  

Figure 3  Estimated number of hunting trips from stratified 1998 MTHS. Row 1 - Bayesian estimated number of trips. Row 2 - Standard deviation of Bayesian estimates of number of trips.
3. General formulae for a bivariate GLMM

3.1 A general model

We ignore the time component in the general model for simplicity. Let \( n_{ik} \) and \( Y_{ik} \) be the sample sizes and response variables of interest in study domain \( i \) and design stratum \( k \), respectively. We assume that \( \{Y_{ik}, n_{ik}\} \) given unknown parameters \( \{\eta_{ik}, \omega_{ik}, \phi_1, \phi_2\} \), with \( i = 1, \ldots, I \), and \( k = 1, \ldots, K \), are independent. Assume that the conditional density (mass) function of \( Y_{ik} \) given \( n_{ik} \) belongs to the following family of probability densities,

\[
g_1(y_{ik} \mid \eta_{ik}, n_{ik}, \phi_1) = \exp\left(A_1(\phi_1)\{y_{ik}\eta_{ik} - B_1(\eta_{ik}, n_{ik})\} + C_1(y_{ik}, n_{ik}, \phi_1)\right),
\]

where \( \eta_{ik} \) is an unknown parameter, and it is often assumed the scale parameter \( \phi_1 \) is known. The probability density function of \( n_{ik} \) is in the family,

\[
g_2(n_{ik} \mid \omega_{ik}, \phi_2) = \exp\left[A_2(\phi_2)\{n_{ik}\omega_{ik} - B_2(\omega_{ik})\} + C_2(n_{ik}, \phi_2)\right],
\]

where \( \omega_{ik} \) is an unknown parameter equal to a function of the population size \( N_{ik} \), and \( \phi_2 \) is often known. The joint density of \( Y_{ik} \) and \( n_{ik} \), a bivariate GLMM, is then

\[
p(y_{ik}, n_{ik} \mid \eta_{ik}, \omega_{ik}, \phi_1, \phi_2) = g_1(y_{ik} \mid \eta_{ik}, n_{ik}, \phi_1)g_2(n_{ik} \mid \omega_{ik}, \phi_2).
\]

The distribution family (10) is often called a generalized linear model, which includes binomial, Poisson, normal and

---

Figure 4 1998 check station data, bayesian prediction, scatterplot of check station by prediction, standard error of bayesian prediction

---
gamma distributions. See, for example, Sun et al. (2000). The distribution family (9) is a generalization of such a generalized linear model by including an additional parameter. Four special cases of (9) are the binomial, Poisson, normal and gamma distributions which are all a part of the exponential family.

The bivariate GLMM is applicable when estimates at the study domains are of interest and the sample size \( n_{ik} \) is considered random as well as being part of the observation. The bivariate GLMM is also useful when estimates of \( N_{ik} \) are required.

A linear mixed model can be used as a prior for \( \eta_{ik} \) to account for the variability in \( \eta_{ik} \). However, one might be interested in both \( \eta_{ik} \) and \( \omega_{ik} \) or a function of \( \eta_{ik} \) and \( \omega_{ik} \) where \( \omega_{ik} \) is often a function of the population size. Here we would need to model \( \eta_{ik} \) and \( \omega_{ik} \) simultaneously. A general class of GLMM for \( \eta_{ik} \) and \( \omega_{ik} \) could be

\[
\begin{align*}
  h_1(\eta_{ik}) &= X_{1i}^T \theta_{ik} + S_{u1i} + e_{1i}, \\
  h_2(\omega_{ik}) &= X_{2i}^T \theta_{2k} + S_{u2i} + e_{2i},
\end{align*}
\]

(12)

With \( a = 1, 2 \), \( X_a = \{x_{aik}\} \) and \( S_a = \{s_{aik}\} \) are known design matrices. The vector \( \theta_{ak} \) is the vector of fixed effects, \( u_{ak} \) is the vector of random effects, and \( e_{aik} \) are independent residual effects and \( e_{aik} \sim N(0, \delta^{(e)}_{aik}) \). In addition, \( u_{ak} \) and \( e_{aik} \) are assumed mutually independent.

### 3.2 Additional priors

To complete the Bayesian hierarchical model, we need to specify the priors for \( (\theta_{ak}, u_{ak}, \delta^{(e)}_{aik}) \), \( a = 1, 2, k = 1, \ldots, K \). The common prior for fixed effects \( \theta_{ak} \) is normal with a large variance or a constant prior. Random effects are often spatially correlated. The density may be of the CAR form whose joint density is given by equation (3) with \( B_{ak} = I - \rho_{aik} C \). Finally, a common prior for variance components \( \delta^{(e)}_{aik} \) is an inverse gamma distribution.

To evaluate the posterior distribution, MCMC methods, such as Gibbs sampling, may be used to obtain samples from the posterior distribution. We give full conditional distributions in the general case below.

**Fact 1** Let \( (\Omega | \cdot) \) represent the conditional distributions of \( \Omega \) given all other parameters and \( [\Omega | \cdot] \) represent the conditional density. The conditional posterior densities of \( \eta_{ik} \) and \( \omega_{ik} \) as follows.

i) \( [\eta_{ik} | \cdot] \propto \exp (A_{1}(\phi_{i})[v_{ik} \eta_{ik} - B_{i}(\eta_{ik}, n_{ik})] - 1/2 \delta^{(e)}_{ik} [h_{i}(\eta_{ik}) - x_{i}^T \theta_{ik} - S_{u} u_{ik}]) \) or equivalently \( v_{ik} = h_{i}(\eta_{ik}) \) has the conditional density,

\[
[v_{ik} | \cdot] \propto \exp \left\{ A_{1}(\phi_{i}) \left[ h_{i}^{-1}(v_{ik}) - B_{i}(h_{i}^{-1}(v_{ik}), n_{ik}) \right] - 1/2 \delta^{(e)}_{ik} (v_{ik} - x_{i}^T \theta_{ik} - S_{u} u_{ik})^2 \right\}.
\]

ii) \( [\omega_{ik} | \cdot] \propto \exp (A_{2}(\phi_{i})[n_{ik} \omega_{ik} - B_{2}(\omega_{ik})] - 1/2 \delta^{(e)}_{2k} [h_{2}(\omega_{ik}) - x_{2}^T \theta_{2k} - S_{u} u_{2k}]) \) or equivalently \( v_{2ik} = h_{2}(\omega_{ik}) \) has the conditional density,

\[
[v_{2ik} | \cdot] \propto \exp \left\{ A_{2}(\phi_{i}) \left[ h_{2}^{-1}(v_{2ik}) - B_{2}(h_{2}^{-1}(v_{2ik}), n_{ik}) \right] - 1/2 \delta^{(e)}_{2k} (v_{2ik} - x_{2}^T \theta_{2k} - S_{u} u_{2k})^2 \right\}.
\]

For \( a = 1, 2 \), we have the following conditional distributions.

iii) \( (\theta_{ak} | \cdot) \) are mutually independent and have conditional posterior distributions

\[
\begin{align*}
  &\left[ \frac{1}{\tau_{ak}} + \frac{1}{\delta^{(e)}_{aik}} X_{a}^T X_{a} \right]^{-1} \left[ \frac{1}{\tau_{ak}} \frac{1}{\delta^{(e)}_{aik}} X_{a}^T \left( v_{ak} - S_{a} \theta_{a} \right) + \frac{1}{\tau_{ak}} \frac{1}{\delta^{(e)}_{aik}} \right], \\
  &\left[ \frac{1}{\tau_{ak}} + \frac{1}{\delta^{(e)}_{aik}} X_{a}^T X_{a} \right]^{-1}
\end{align*}
\]

iv) \( (u_{aik} | \cdot) \) are mutually independent and have conditional posterior distributions

\[
\begin{align*}
  &\left[ \frac{1}{\delta^{(e)}_{aik}} S_{a} + \frac{1}{\delta^{(e)}_{aik}} B_{a} \right]^{-1} \left[ \frac{1}{\delta^{(e)}_{aik}} S_{a} \left( v_{aik} - X_{a} \theta_{a} \right) + \frac{1}{\delta^{(e)}_{aik}} B_{a} \right], \\
  &\left[ \frac{1}{\delta^{(e)}_{aik}} S_{a} + \frac{1}{\delta^{(e)}_{aik}} B_{a} \right]^{-1}
\end{align*}
\]

v) \( (\delta^{(e)}_{aik} | \cdot) \) have posterior distributions

\[
\begin{align*}
  &\left[ \alpha_{aik}^{(e)} + 1/2, \beta_{aik}^{(e)} \right], \\
  &\text{Inverse Gamma} + 1/2 (v_{aik} - X_{a} \theta_{a} - S_{a} u_{aik}) \right], \\
  &\left[ \alpha_{aik}^{(e)} + 1/2, \beta_{aik}^{(e)} + 1/2 u_{aik}^{(e)} + 1/2 B_{aik} u_{aik} \right).
\end{align*}
\]

vi) \( (\phi_{i} | \cdot) \) are independent and have posterior distributions

\[
\begin{align*}
  &\left[ \rho_{i} \right] \propto B_{i} | 1/2 \exp \left\{ u_{i}^T B_{i} u_{i} / 2\delta_{ik}^{(e)} \right\}, \\
  &\left[ \rho_{i} \right] \propto B_{i} | 1/2 \exp \left\{ u_{i}^T B_{i} u_{i} / 2\delta_{ik}^{(e)} \right\}.
\end{align*}
\]

viii) If \( \phi_{i} \) has the prior density \( p(\phi_{i}) \), then

\[
\begin{align*}
  &\left[ \phi_{i} \right] \propto p(\phi_{i}) \exp \left\{ -B_{i}(h_{i}^{-1}(v_{ik}), n_{ik}) \right\} + C_{i}(v_{ik}, n_{ik}, \phi_{i})
\end{align*}
\]

ix) If \( \phi_{2} \) has the prior density \( p(\phi_{2}) \), then

Statistics Canada, Catalogue No. 12-001-X
Examining parts i), ii), and vii) of Fact 1, the conditional densities of \( \eta_{ik}, \omega_{ik} \) and \( p_{ik} \) are often log-concave.

3.3 Estimating quantities of study domains

It is often of interest to estimate certain quantities related to the study domains. For instance, to estimate a quantity at domain \( i \), let

\[
\psi_i = f_i(\eta_{i1}, \omega_{i1}; \ldots; \eta_{ik}, \omega_{ik}).
\]

Bayesian estimates of \( \psi_i \) can be easily computed based on a random sample from the joint posterior. For example, let \( (\eta_{i1}^{(l)}, \omega_{i1}^{(l)}), l = 1, \ldots, L \) and \( k = 1, \ldots, K \) be the output from MCMC simulations, and define

\[
\psi_{i}^{(l)} = f_i(\eta_{i1}^{(l)}, \omega_{i1}^{(l)}; \ldots; \eta_{ik}^{(l)}, \omega_{ik}^{(l)}).
\]

Given \( y = \{Y_{ik}, i = 1, \ldots, I, k = 1, \ldots, K\} \), the posterior mean and variance of \( \psi_i \), the general forms of (4) and (5), can be approximated by

\[
\hat{E}(\psi_i | y) = \frac{1}{L} \sum_{l=1}^{L} \psi_{i}^{(l)},
\]

and

\[
\hat{V}(\psi_i | y) = \frac{1}{L-1} \sum_{l=1}^{L} (\psi_{i}^{(l)})^2 - \frac{L}{L-1} (\hat{E}(\psi_i | y))^2,
\]

respectively.

3.4 Predicting quantities of study domains

Now let \( N_{ik} \) be the population size and \( Y_{ik} \) the response value of interest for those not in the sample from study domain \( i \) and design stratum \( k \). We wish to predict the quantities \( Y_{ik} = \sum_{i=1}^{I} Y_{ik} \), the total response value in study domain \( i \). For given \( n_{ik}, Y_{ik} \) should be of the same family as \( Y_{ik} \) such that

\[
g_3(Y_{ik} | \eta_{ik}, n_{ik}, N_{ik}, p_i) = \exp \left[ A_i(\phi_i)(Y_{ik} - \eta_{ik}) \right],
\]

This is simply (9) with \( n_{ik} \) replaced by \( N_{ik} - n_{ik} \). To simplify notation, let \( \xi \) denote the parameters in the model, \( d \) represent the data. (In our case here \( d = \{y_{ik}, n_{ik}\}: i = 1, \ldots, I, k = 1, \ldots, K\} \) and \( \xi \) might include the parameters in modeling both \( y_{ik} \) and \( n_{ik} \). Under the prior \( \pi(\xi) \) (either proper or improper), we have the posterior density \( [\xi | d] \propto f(d | \xi) \pi(\xi) \).

Assume that a further observation \( y_{ik} \) follows \( g_3(y_{ik} | d, \xi) \) that may be dependent on \( d \). The predictive density of \( y_{ik} \) given \( d \) is written as

\[
[y_{ik} | d] = \int g_3(y_{ik} | \xi, d)[\xi | d]d\xi.
\]

Under the squared error loss function, the best predictor is then the predictive mean given by

\[
E(y_{ik} | d) = \int \left[ \int g_3(y_{ik} | \xi, d)[\xi | d]d\xi \right] y_{ik} | d d\xi.
\]

Similarly, the best predictor of \( h(y_{ik}^*, \ldots, y_{ik}^* | d) \) given \( d \) is then

\[
E\{h(y_{ik}^*, \ldots, y_{ik}^*) | d\} = \int E\{h(y_{ik}^*, \ldots, y_{ik}^*) | \xi, d\} [\xi | d]d\xi \tag{14}
\]

where

\[
E\{h(y_{ik}^*, \ldots, y_{ik}^*) | \xi, d\} = \int (y_{ik}^*, \ldots, y_{ik}^*)h(y_{ik}^*, \ldots, y_{ik}^*)\prod_{k=1}^{K}[y_{ik}^* | \xi, d]dy_{ik}^* \ldots dy_{ik}^* \tag{15}
\]

For the distribution family \( g_3(y_{ik} | \xi, d) \), the right hand side of (15) often has a closed form expression.

Bayesian predictions of (14) can be easily computed based on a random sample from the joint posterior of \( \xi \). For example, let \( \xi^{(l)}, l = 1, \ldots, L \), be the output from MCMC simulations. The posterior predictive mean (14) can then be approximated by

\[
\hat{E}\{h(y_{ik}^*, \ldots, y_{ik}^*) | d\} = \frac{1}{L} \sum_{l=1}^{L} E\{h(y_{ik}^*, \ldots, y_{ik}^*) | \xi^{(l)}, d\}.
\]

The posterior predictive variance of \( h(y_{ik}^*, \ldots, y_{ik}^*) \) given \( d \) may be written as

\[
V\{h(y_{ik}^*, \ldots, y_{ik}^*) | d\} = E[V\{h(y_{ik}^*, \ldots, y_{ik}^*) | \xi, d\} | d]
+ V[E\{h(y_{ik}^*, \ldots, y_{ik}^*) | \xi, d\} | d].
\]

This predictive variance can then be approximated by

\[
\hat{V}\{h(y_{ik}^*, \ldots, y_{ik}^*) | d\} = \frac{1}{L} \sum_{l=1}^{L} V\{h(y_{ik}^*, \ldots, y_{ik}^*) | \xi^{(l)}, d\}
+ \frac{1}{L} \sum_{l=1}^{L} [E\{h(y_{ik}^*, \ldots, y_{ik}^*) | \xi^{(l)}, d\} - \hat{E}\{h(y_{ik}^*, \ldots, y_{ik}^*) | d\}]^2.
\]

4. Comments

In this article, we developed a bivariate GLMM containing two unknown canonical parameters. This was necessary to obtain estimates when the sample size was random and estimates \( N_{ik} \) were required in addition to estimates of \( \eta_{ik} \). The model was built using two simultaneous GLMMs in a Bayesian hierarchical structure.
The proposed model has the advantage of being applicable to a wide array of problems. Introducing a random sample size and estimating the population sizes are useful techniques for many applications.

Naturally, we think that there is an inverse relationship between hunting success rates and the number of trips taken to a county. In modeling this, we may think of a two-fold CAR model to view the relationship between $\eta_{ik}$ and $N_{ik}$, such as that of Kim, Sun and Tsutakawa (2001). A multivariate CAR model may be another approach to address this situation.

For each of the spatial models, we have assumed a common correlation, $\rho_k$, across the entire state. It may be more reasonable to include additional correlation terms in different productivity regions defined by the Missouri Department of Conservation. This would be an interesting and complicated addition to the model. In addition, the spatial structure used in this paper is similar to that of He and Sun (2000) and Oleson and He (2004) where this spatial modeling worked well.

It may be useful to include the distance from hunter’s home to the hunting location to help estimate hunting pressure. Most hunters stay close to home when hunting and this information could be incorporated into the hierarchical framework.

Note that the estimated harvest is higher than the check station harvest. This is partly because more successful hunters tend to reply to the mail survey. We are conducting research adjusting for the nonresponse bias.

Acknowledgements

The work was partially supported by Federal Aid in Wildlife Restoration Project W-13-R. Sun’s research was supported by the National Science Foundation grant SES-0351523, and NIH grant R01-CA100760. The authors would like to thank Dr. Larry Vangilder and Mr. Jeff Beringer of the Missouri Department of Conservation for their helpful comments. The authors wish to express deep appreciation to the editors, an associate editor, and two referees for constructive comments on earlier versions of the paper.

References


Small area estimation of average household income based on unit level models for panel data

Enrico Fabrizi, Maria Rosaria Ferrante and Silvia Pacei

Abstract
The European Community Household Panel (ECHP) is a panel survey covering a wide range of topics regarding economic, social and living conditions. In particular, it makes it possible to calculate disposable equivalized household income, which is a key variable in the study of economic inequity and poverty. To obtain reliable estimates of the average of this variable for regions within countries it is necessary to have recourse to small area estimation methods. In this paper, we focus on empirical best linear predictors of the average equivalized income based on “unit level models” borrowing strength across both areas and times. Using a simulation study based on ECHP data, we compare the suggested estimators with cross-sectional model-based and design-based estimators. In the case of these empirical predictors, we also compare three different MSE estimators. Results show that those estimators connected to models that take units’ autocorrelation into account lead to a significant gain in efficiency, even when there are no covariates available whose population mean is known.

Key Words: European Community Household Panel; Average equivalized income; Linear mixed models; Empirical best linear unbiased predictor; MSE estimation.

1. Introduction
In recent years, the academic world has taken an increasing interest in the analysis of regional economic disparities that represent a serious challenge to the promotion of national economic growth, and thus to social cohesion. This is particularly true within the European Union, where regional disparities are a distinguishing feature of the economic landscape. This renewed interest in local economies has produced a growing demand for regional statistical information and has stimulated research on income distribution, poverty and social exclusion at the sub-national level.

In the 1990s, Eurostat (the EU’s Statistics Bureau) launched the European Community Household Panel (ECHP), an annual panel survey of European households conducted using standardised methods throughout the EU’s various member countries (Betti and Verma 2002; Eurostat 2002). The ECHP terminated in 2001, after eight waves. Currently, it is being replaced by the Survey on Income and Living Conditions in the Community (EU-SILC), which resembles the ECHP in many ways, but for which no data has yet been published. The ECHP panel survey covered a wide range of topics and, in particular, it made it possible to calculate disposable equivalized household income, which constitutes a key variable in the study of economic equity and poverty.

The ECHP was designed to provide reliable estimates for large areas within countries called NUTS1 (NUTS stands for the “Nomenclature of Territorial Units for Statistics” which is defined according to certain principles described on the EUROSTAT web site http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html ). Unfortunately NUTS1 correspond to areas (five groups of Administrative Regions in the Italian case) that are too large to effectively measure local area income disparity or to provide useful information for the purposes of regional governance. Therefore, to obtain estimates for a finer geographic detail, a small area estimation method has to be used and the problem is to select an appropriate and effective method.

In this paper, in order to combine information from past surveys, related auxiliary variables and small areas, we consider several possible extensions of the well-known unit level nested error regression model (see Battese, Harter and Fuller 1988) for the estimation of the average of household equivalized income. Using ECHP panel survey data, we illustrate how such model could be potentially useful in improving the efficiency of small area estimates by exploiting the correlation of individual household incomes over time.

In section 2, we present a general set-up for small area estimation using panel survey data and briefly review both design-based and model-based small area estimation methods. In this section, we develop empirical best linear unbiased predictors (EBLUP) and their mean squared error (MSE) estimators for selected unit level cross-sectional and time series models using the available theory on EBLUP for small area estimation (see Rao 2003, and Jiang and Lahiri 2006a, for details). We note that cross-sectional and time series models were considered in the small area literature,
but only in the context of area level modelling (see Rao and Yu 1994; Ghosh, Nangia and Kim 1996; Datta, Lahiri, Maiti and Lu 1999; Datta, Lahiri and Maiti 2002; Pfeffermann 2002; among others).

In section 3, we briefly review ECHP survey and describe how we use this survey data to conduct a Monte Carlo simulation study to compare different small area estimators and their MSE estimators. In section 4, we report results from the Monte Carlo simulation experiment. We note that the simulation experiment is aimed at evaluating design-based properties of all estimators, even if they are derived as model based predictors. We observed that the EBLUPs perform very well compared to the design-based estimators even though our pseudo-population exhibits signs of non-normality. The non-normality of the pseudo-population, however, seems to affect the efficiency of the MSE estimators. In our simulation, the Taylor series (see Prasad and Rao 1990; Datta and Lahiri 1999, among others) and the parametric bootstrap (see Butar and Lahiri 2003) MSE estimators are found out to be more sensitive to the non-normality than the jackknife method of Jiang, Lahiri and Wan (2002). We end the paper with a few concluding remarks.

2. The small area estimation methods considered

To describe sample data, let \( y_{dt} \) denote the value of a study variable for the \( i \)th unit belonging to the \( d \)th small area for time \( t \) \((d = 1, \ldots, m; t = 1, \ldots, T; i = 1, \ldots, n_{dt})\). Moreover, let \( x_{dt} \) be the vector of covariates’ values associated with each \( y_{dt} \) (and whose first element is equal to 1), and let \( X = \{x_{dt}'\} \) be the \( n \times p \) matrix of covariates’ values for the whole sample \((n = \sum_{d=1}^{m} n_{dt})\). Let us suppose that we are interested in predicting small area means for the target variable at final time \( T : \overline{Y}_{dT}(d = 1, \ldots, m)\). Let us also suppose that the vectors of mean population values of covariates are known for time \( T\); we denote these vectors by \( \overline{X}'_{dT}(d = 1, \ldots, m)\).

2.1 Design-based estimators

A first solution to the small area estimation problem is to use direct estimators, that is, estimators employing only \( y \) values obtained from the area (and time) which the parameter refers to. The simplest of direct estimators of the population mean is the weighted mean. We denote this direct estimator as \( \overline{Y}_{dT, DIR}(d = 1, \ldots, m) \) and we will be using it as a benchmark in the following sections.

Synthetic estimators may be generally defined as unbiased estimators for a larger area with acceptable standard errors. They are used to calculate estimates for small areas, under the hypothesis that small areas have the same characteristics as larger ones. Moreover, when information about auxiliary variables is available, a particular synthetic estimator, the regression estimator, may be obtained by fitting a regression model to all sample data. Note that the synthetic estimator is area specific with respect to the auxiliary variables but not with respect to the study variable.

For instance, if we consider only those observations from the last wave \((t = T)\), the simple regression model would be given by:

\[
y_{dt} = x_{dt}' \beta + e_{dt},
\]

\[E(e_{dt}) = 0, \quad E(e_{dt}^2) = \pi^2.\]

To account for the complexity of the sampling design, the weighted least squares estimate \( \hat{\beta} \) of \( \beta \) may be obtained, and thus the synthetic regression estimator will be given by:

\[
\overline{Y}_{dT, RSYN} = \overline{X}'_{dT} \hat{\beta}, d = 1, \ldots, m.
\] (1)

Synthetic estimators usually display very low variances, but they may be severely biased whenever the model holding for the whole sample does not properly fit area-specific data. Composite estimators are weighted averages of a direct and a synthetic estimator. We consider the composite estimator:

\[
\overline{Y}_{dT, COMP} = \phi_{dT} \overline{Y}_{dT, DIR} + (1 - \phi_{dT}) \overline{Y}_{dT, RSYN},
\] (2)

where

\[
\phi_{dT} = \frac{\text{MSE}_D(\overline{Y}_{dT, RSYN})}{\text{MSE}_D(\overline{Y}_{dT, DIR}) + \text{MSE}_D(\overline{Y}_{dT, RSYN})}
\]

and \( \text{MSE}_D \) signifies that the mean square error is evaluated in relation to the randomization distribution. This choice of \( \phi_{dT} \) leads to composite estimators \( \overline{Y}_{dT, COMP} \) that are approximately optimal in terms of \( \text{MSE}_D \) (see Rao 2003, section 4.3). In practice, the quantities in the formula for \( \phi_{dT} \)’s are unknown and may be estimated from the data. Unbiased and consistent estimators can be obtained for \( \text{MSE}_D(\overline{Y}_{dT, DIR}) = V_D(\overline{Y}_{dT, DIR}) \) using standard formulas. An approximately design unbiased estimator of \( \text{MSE}_D(\overline{Y}_{dT, RSYN}) \) can be obtained using the formulas discussed in Rao (2003, section 4.2.4). In particular, we calculate the approximation:

\[
\text{mse}_D(\overline{Y}_{dT, RSYN}) \approx (\overline{Y}_{dT, RSYN} - \overline{Y}_{dT, DIR})^2 - v_D(\overline{Y}_{dT, DIR}),
\]

where \( \text{mse}_D \) and \( v_D \) stand for the estimators of the corresponding \( \text{MSE}_D \) and \( V_D \). In particular, \( v_D \) is the ordinary design unbiased estimator of \( V_D \). We then take its average over \( d \), as usual, in order to obtain a more stable estimator. In fact, one problem with \( \text{mse}_D \) is that it can even be negative.
Moreover, a modified direct estimator borrowing strength over areas for estimating the regression coefficient can be used to improve estimator reliability. If auxiliary information is available, the generalized regression estimator (GREG),

$$\tilde{Y}_{dt},_{GREG} = \bar{X}_{dt} \hat{p} + \sum_{j \in x_{dt}} \frac{w_j e_j}{\sum_{j \in x_{dt}} w_j},$$

(3)

approximately corrects the bias of the synthetic estimator by means of the term $((\sum_{j \in x_{dt}} w_j) \times \sum_{j \in x_{dt}} w_j e_j)$, based on regression residuals $e_j$.

### 2.2 Model-based estimators

The model-based estimators we have considered are based on the specification of explicit models for sample data which approximate a hypothetical data-generating process. As a consequence, the problem of estimating $\tilde{Y}_{dt}$ comes down to one of prediction. Moreover, mean square errors and other statistical properties of estimators are usually evaluated with respect to the data-generating process. We have focused here on “unit level” models based on models relating $y_{dti}$ to a vector of covariates $x_{dti}$. The use of explicit models has several advantages, the most important of which being the opportunity to test underlying assumptions.

In the estimation of the small area means or totals of continuous variables, linear mixed models are very often used. A general linear mixed model can be described as follows:

$$y = X\beta + Z_1 v_1 + ... + Z_s v_s + e,$$

(4)

where $y = \{y_{dti}\}$ is the $n$-vector of sample observations, $\beta$ a $p \times 1$ vector of fixed effects, $v_j$ a $q_j \times 1$ vector of random effects ($j = 1, ..., s$), $e = \{e_{dti}\}$ a vector of errors; $X$ is assumed of rank $p$, $Z_j = \{Z_{j,dti}\}$ is a $n \times q_j$ matrix of incidence of the $j$th random effect. We assume that $E(v_j) = 0, V(v_j) = G_j, E(e) = 0, V(e) = R$ (all expectations are wrt. model (4)) and that $v_1, ..., v_s, e$ are mutually independent.

As a consequence, the variance-covariance matrix of $y$ is given by:

$$V = V(y) = \sum_{j=1}^{s} Z_j G_j Z_j' + R = ZGZ' + R,$$

where $Z = [Z_1 | ... | Z_s]$. It is usually assumed that matrices $G, R$ depend on a $k$-vector of random components $\psi$, and so we can write $V(\psi) = ZG(\psi)Z' + R(\psi)$.

Note that at the level of individual observations, the model (4) can be rewritten as $y_{dti} = x_{dt} \beta + v_{dti} + e_{dti}$.

We consider different specifications for linear mixed models, all of which can be viewed as special cases of the general model (4). For the sake of simplicity, we have adopted a unit level notation when describing the models considered. The first model:

$$MM1: \quad y_{dti} = x_{dti} \beta + v_{dti} + e_{dti},$$

(5)

may be obtained from formula (4) setting $s = 2, q_1 = m, v_2 = T, G_1 = \sigma_1^2 I_m, G_2 = \sigma_2^2 I_r, R = \sigma_3^2 I_s$. It includes independent area and time effects, and therefore area effects are assumed not to evolve over time. This random effects structure corresponds to the assumption of a constant covariance between units that belong to the same area, observed at two different points in time.

The second model:

$$MM2: \quad y_{dti} = x_{dti} \beta + \delta_{dti} + e_{dti},$$

(6)

corresponds to the particular case in which $s = 1, q_1 = mq, G_1 = \sigma_1^2 I_m, \delta = \sigma_2^2 I_q$. The effects of interaction between area and time are introduced, that is, we assume there are area effects which are not constant over time.

The third model:

$$MM3: \quad y_{dti} = x_{dti} \beta + v_{dti} + \alpha^* + e_{dti},$$

(7)

is obtained setting $s = 2, q_1 = m, q_2 = T, G_1 = \sigma_1^2 I_m, R = \sigma_3^2 I_r, \alpha^* = \sigma_3^2 \rho_{h, k, q}, h, k = 1, ..., T$. There are independent area and time effects, just as in MM1, but the time effects are assumed to follow an AR(1) process.

The fourth model:

$$MM4: \quad y_{dti} = x_{dti} \beta + \delta_{dti} + e_{dti},$$

(8)

is similar to model MM2 in that it is characterized by time varying area effects, but the further assumption that such effects follow an AR(1) process is also introduced. Thus, provided we order observations by area, with respect to the general formula (4) we have $s = 1, q_1 = mq, G_1 = \text{diag}(G_{id}), R = \sigma^2 I_n$ where $G_{id}, d = 1, ..., m$, is a $T \times T$ matrix the generic element $g_{id}(h, k) = \sigma^2 \rho_{h, k, q}, h, k = 1, ..., T$.

The last specification:

$$MM5: \quad y_{dti} = x_{dti} \beta + v_{dti} + \alpha^* + e_{dti},$$

(9)

may be obtained by (4) setting $s = 2, q_1 = m, q_2 = T, G_1 = \sigma_1^2 I_m, G_2 = \sigma_2^2 I_r$. Provided we order observations by household and time, $R = \text{diag}(R_{di})$ where $R_{di}$ is a $T \times T$ matrix whose generic element is given by $r_{di}(h, k) = \sigma^2 \rho_{h, k, q}, h, k = 1, ..., T$. There are independent area and time effects like in MM1, but errors are assumed to be autocorrelated according to an AR(1) process.
In order to evaluate the impact that using past survey waves has on the efficiency of estimator, a cross-sectional linear mixed model (SMM) using data from the last wave $T$ only, has been taken as the benchmark:

$$SMM: \quad y_{dji} = x_{dji}^\prime \beta + \varrho_{d} + e_{dji} \quad (10)$$

with $\varrho_{d} \sim N(0, \sigma^2_{\varrho})$, $e_{dji} \sim N(0, \sigma^2_e)$.

This is also a particular case of (4) obtained for $s = 2$, $q_i = m$, $G_i = \sigma^2_{\varrho_i} I_m$ and $R = \sigma^2_e I_n$. Note that (10) is the standard nested error regression model of Battese et al. (1988).

We also consider the corresponding random error variance linear models (see Rao 2003; section 5.5.2) obtained by replacing $variance$ linear models (see Rao 2003; section 5.5.2) (1988).

standard nested error regression model of Battese et al. (1988).

One popular method is based on the Taylor series approximation of $MSE$ under normality (Prasad and Rao 1990; Datta and Lahiri 1999). More recently, due to the advent of high-speed computers and powerful software, resampling methods have been proposed. For instance, Butar and Lahiri (2003) introduce a parametric bootstrap method based on the assumption of normality, but analytically less onerous than the Taylor series method. Jiang et al. (2002) discuss a general jackknife method, which requires a distributional assumption weaker than normality (posterior linearity). We aim to empirically compare the performance of these three estimators within a context where the number of areas is moderate and the assumption of normality may not hold perfectly true. The following is a short description of the three estimation approaches.

Let us define $MSE[\hat{\eta}_{EBLUP}(\psi)] = E[\hat{\eta}_{EBLUP}(\psi) - \eta]^2$, where expectation refers to model (4). It is possible to show that, under normality,

$$MSE[\hat{\eta}_{EBLUP}(\psi)] = g_1(\psi) + g_2(\psi) + E[\hat{\eta}_{EBLUP} - \eta_{EBLUP}]^2 \quad (12)$$

where $g_1(\psi) = k(G - GZV^{-1}ZG)k'$ and $g_2(\psi) = d'(XV^{-1}X)^{-1}d$, with $d = m^t - kGZV^{-1}X$ (see Rao 2003, chapter 3). Using the following approximation, based on a Taylor series argument

$$E[\hat{\eta}_{EBLUP} - \eta_{EBLUP}]^2$$

$$\approx tr[(\partial^2 b'/\partial \psi')V(\partial^2 b'/\partial \psi')\nabla(\psi)] = g_3(\psi)$$

where $b' = kGZV^{-1}$, a second order approximation to (12) can be found:

$$MSE[\hat{\eta}_{EBLUP}(\psi)] \approx g_1(\psi) + g_2(\psi) + g_3(\psi). \quad (13)$$

Note that here $\approx$ means that the omitted terms are of order $o(m^{-1})$. An asymptotically unbiased estimator of (13), based on Prasad and Rao (1990), is given by

$$mse_{PR}[\hat{\eta}_{EBLUP}(\psi)] = g_1(\psi) + g_2(\psi) + 2g_3(\psi). \quad (14)$$

Datta and Lahiri (1999) show that, under normality and REML or ML estimation of $\psi$, $mse_{PR}[\hat{\eta}_{EBLUP}(\psi)]$ estimates $MSE[\hat{\eta}_{EBLUP}(\psi)]$ with a bias of order $O(m^{-1})$.

Butar and Lahiri (2003) propose a parametric bootstrap estimation (for bootstrap) under the assumption of normality. We adapt their estimator to the models we are analysing, assuming the following bootstrap model:

$$i) \quad y^* \sim N[X\hat{\beta} + Zv^*, R(\psi)] \quad (15)$$

$$ii) \quad v \sim N(0, G(\psi))$$

where $v = (v_1, ..., v_p)'$. The parametric bootstrap is then used twice, once to estimate the first two terms of (13), thus...
correcting the bias of \( g_1(\hat{\psi}) + g_2(\hat{\psi}) \), and once to estimate \( g_3(\hat{\psi}) \).

The following estimator of (13) is proposed:

\[
\text{mse}_{BL}(<\text{EBLUP}>_j) = 2[g_1(\hat{\psi}) + g_2(\hat{\psi})] - E_g[g_1(\hat{\psi}^*) + g_2(\hat{\psi}^*)] + E_g[\eta(\hat{y}, \hat{\beta}(\hat{\psi}^*), \hat{\psi}^*) - \eta(\hat{y}, \hat{\beta}(\hat{\psi}), \hat{\psi})]
\]  

(16)

where \( \hat{\psi}^* \) is the same as \( \hat{\psi} \) except that it is calculated on \( y^* \) instead of \( y \), and \( E_g \) is the expected value with regard to the bootstrap model (15).

The bootstrap estimator (16) does not require the analytical derivation of \( g_3(\hat{\psi}) \) which can be rather laborious when \( G \) and \( R \) have complicated structures.

Jiang et al. (2002) introduced a general jackknife estimator for the variance of empirical best predictors in linear and non-linear mixed models with \( M \)-estimation. In the problem we are investigating here, the estimator they propose can be written as:

\[
\text{mse}_{JL,W}(<\text{EBLUP}>_j) = g_1(\hat{\psi}) - \frac{m-1}{m} \sum_{j=1}^{m} [g_1(\hat{\psi}_{-j}) - g_1(\hat{\psi})] + \frac{m-1}{m} \sum_{j=1}^{m} (<\text{EBLUP}>_{-j} - <\text{EBLUP}>_j)^2
\]

(17)

where \( \hat{\psi}_{-j} \) is the estimate of \( \psi \) calculated by using all data except those from the \( j \)-th area. Similarly, \( <\text{EBLUP}>_{-j} = <\text{EBLUP}>_j(\hat{y}_{-j}, \hat{\beta}(\hat{\psi}_{-j}), \hat{\psi}_{-j}) \).

It is worth pointing out that, on the basis of the simulation results reported in Jiang et al. (2002), \( \text{mse}_{JL,W} \) is deemed to be more robust than \( \text{mse}_{PR} \) with regard to departures from the assumption of normality, which can also be expected to be crucial for \( \text{mse}_{BL} \).

3. The simulation study based on the European Household Community Panel data

The target population of the ECHP survey consists of all resident households of a large subset of the EU member countries. Although general survey guidelines were issued by Eurostat, a certain degree of flexibility was allowed, so there are some differences in the sampling design across countries. As far as Italy is concerned, the survey is based on a stratified two stage design, in which strata were formed by grouping the PSUs (municipalities) according to geographic region (NUTS2) and demographic size: For more details of the survey, see Eurostat (2002).

The ECHP deals with unit non-response, sample attrition and new entries using weighting and imputation. As attrition could lead to biased estimates of income if it does not appear at random, the effect of poverty on dropout propensity has been investigated (Rendtel, Behr and Sisto 2003; Vandecasteele and Debels 2004), and the results of these studies show that in the case of some countries, including Italy, this effect disappears under the control of weighting variables.

We have focused our attention on the eight ECHP waves available for Italy (1994-2001). Given that our aim is to assess whether the use of several successive observations of the same household could be profitable for the purposes of small area estimation, we have overlooked the problem of attrition and only considered those households that participate to the survey for all waves.

Our target variable is disposable, post-tax household income at the time of the last wave (2001). In studies of poverty and inequality, income is often equivalized according to an equivalence scale in order to avoid comparison problems caused by differences in the composition of households. We consider the widely-used modified OECD scale, also adopted by Eurostat (2002) in its publications on income, poverty and social exclusion. According to this scale, equivalized income is calculated by dividing disposable household income by the number \( k \) of "equivalent adults", defined as \( k = 1 + 0.5a + 0.3c \), where \( a \) is the number of adults other than the "head of the household" and \( c \) is the number of children aged 13 or less. In general, the equivalized income can be perceived as the amount of income that an individual, living alone, should dispose of in order to attain the same level of economic wellbeing he/she enjoys in his/her household.

Of the many covariates available in the bountiful ECHP questionnaire, we have chosen only those for which area means were available from the 2001 Italian Census results. Thus the chosen covariates are: the percentage of adults; the percentage of employed; the percentage of unemployed; the percentage of people with a high/medium/low level of education in the household; household typology (presence of children, presence of aged people, etc.); the number of rooms per-capita and the tenure status of the accommodation (rented, owned etc.).

As we have said, the aim of this paper is to compare the performance of different estimators in the controlled environment of a simulation exercise. A number of works in the literature have compared small area estimators using Monte Carlo experiments in which samples are drawn from synthetic populations based either on Censuses (Falorsi, Falorsi and Russo 1994; Ghosh et al. 1996) or on the replication of sample units’ records (Falorsi, Falorsi and Russo 1999; Lethtonen, Särndal and Veijanen 2003; Singh, Mantel and Thomas 1994). Since household income is not measured by the Italian Census (nor is it given by the results of other Censuses conducted by EU countries), we treated the ECHP survey data as the pseudo-population and then draw samples using stratified probability proportional to
size sampling, the size variable being given by survey weights. This solution may not be as good as that of using data from a real Census population, but it is hopefully more realistic than generating population values of household income from a parametric model.

Monte Carlo samples of 1,000 (roughly 15% of the actual ECHP sample size) were drawn from the synthetic population by stratified random sampling without replacement, with strata given by the 21 NUTS2 regions. Thus these regions are treated as planned domains (as in the ECHP) for which sample size in the small areas is established beforehand, so that the sampling fractions reflect the over-sampling of smaller regions exactly as they do in the actual ECHP sampling design. The region-specific sample sizes we obtained range from 14 to 112, being on average equal to 48. Therefore in our simulation we obtained range from 14 to 112, being on average equal to 48.

The difference between mean and median is 9% of the mean. An interesting feature is given by the large coefficient of variation (ranging from 0.28 to 0.84), skewness (γ ranging from -0.7 to 32.9) showing that the distribution of our target variable is quite a bit different from the distribution of our target variable is quite a bit different.

Moreover the residuals show a degree of autocorrelation, the average of the autocorrelation coefficient calculated over all individual residual histories being 0.27. Even though this autocorrelation level is not very high, for the sake of completeness we decided to also take into consideration models with autocorrelated errors or random effects. After having tested various different autocorrelation structures (ARMA(p, q), General Linear, etc.), we found that the autoregressive process of order 1 provides the best fit to our data.

To motivate the selected specifications of the random effects part of the considered linear mixed models (see section 2.2), an approach often recommended in textbooks (see Verbeke and Molenberghs 2000, chapter 9) has been followed: first we fit a standard OLS regression to our data using all available covariates; then we analyse the resulting residuals as a guide to identifying the random effects. This preliminary analysis has been conducted separately on several random samples of size 1,000, drawn up according to the replication design described above.

The adjusted $R^2$ of the OLS regression is close to 0.35 in every observed sample. This rather low figure is the result of the nature of the phenomenon under study (household income is not easy to predict), the information contained in the survey and the constraint represented by the need to include only those covariates for which the population total can be obtained from the Census.

Figure 1 contains “box and whiskers” plots of the residuals by area and wave constructed for one of the Monte Carlo sample (very similar findings may be observed in every sample). Analysis of the plots suggests that there is within-area and within-wave correlation, and thus the need to specify models including area and wave effects. From an analysis of residuals, it is less clear whether the inclusion of interaction effects (that is time varying area effects) would be beneficial or not.

Statistics Canada, Catalogue No. 12-001-X
The apparent skewness of residuals also suggests that the normality assumption for errors does not hold exactly. We maintain this assumption for all the models we specify, and we use REML estimators for variance components. In fact, we may expect departures from normality to have a slight impact on point values of predictors. BLUP formulas can be derived without normality; moreover, there are sound reasons for us to expect REML (and ML) estimators of $\psi$ to perform well even if normality does not hold (see Jiang, 1996, for details). Departures from normality may have more a serious impact on MSE estimation, and this is a problem we are going to be looking at in section 4.2 below.

4. Results

4.1 Point estimators

All computations involved in the simulation exercise described in section 3 were carried out using SAS version 9.1 for Windows. EBLUP estimators are obtained using Proc MIXED, and the generation of samples is based on Proc SURVEYSELECT.

Given that the primary goal of Small Area Estimation is the precise estimation of area-specific parameters, we first evaluated how well the described estimators perform when predicting individual area values. Moreover, we also evaluated the amount of over-shrinkage connected with each estimator. In fact, small area estimates should reflect (at least approximately) the variability in the underlying area parameters taken as a whole.

We note that our simulation experiment is aimed at evaluating design-based properties of the estimators, that is, the population from which the random samples are generated is held fixed.

For the evaluation of the estimators’ performance, we adopted an approach that is commonly found in the literature (see Rao 2003; section 7.2.6), using two indicators, the Average Absolute Relative Bias (AARB) and the Average Relative Mean Square Error (ARMSE):

$$\text{AARB} = m^{-1} \sum_{d=1}^{m} \left( R^{-1} \sum_{r=1}^{R} \left( \frac{\bar{y}_{d,r}(T) - \bar{y}_{d,T}}{\bar{y}_{d,T}} - 1 \right) \right)$$

$$\text{ARMSE} = m^{-1} \sum_{d=1}^{m} \left( R^{-1} \sum_{r=1}^{R} \left( \frac{\bar{y}_{d,r}(T)}{\bar{y}_{d,T}} - 1 \right) \right)^2$$

(18)

where $\bar{y}_{d,r}(T)$ is the estimate for area $d$, time $T$ and replicated sample $r$, while $\bar{y}_{d,T}$ is the population mean being estimated. Note that AARB measures the bias of an estimator, whereas ARMSE measures its accuracy. The number of replications $R$ is set at 500, a figure large enough to obtain stable Monte Carlo estimates of expected values and variances, frequently used in simulation studies on small area estimation (Heady, Higgins and Ralphs 2004; EURAREA Consortium 2004).

The gain in efficiency connected to each small area estimator is evaluated using the ratio of its ARMSE to the ARMSE of certain estimators we use as benchmarks. In particular, all estimators are compared with the weighted mean $\bar{y}_{d,T,\text{Dir}}$ and we denote this ratio as $\text{AEFF}_{\text{Dir}}$. Moreover, EBLUP estimators associated with models (5) - (9), which use data from previous waves, are compared with the EBLUP estimator associated with the cross-sectional model (10), in order to assess the gain in efficiency deriving from the use of past waves. In this case the ratio is denoted as $\text{AEFF}_{\text{Seq}}$.

As far as the evaluation of the degree of shrinkage is concerned, we have compared the empirical standard deviation of population area values:

$$\text{ESD} = \sqrt{m^{-1} \sum_{d=1}^{m} \left( \frac{\bar{y}_{d,T} - \bar{y}_{r}}{\bar{y}_{r}} \right)^2}$$

where $\bar{y}_{r}$ is the mean of the population values of the $m$ areas at time $T$, with the empirical standard deviation of the estimated area values, which in the case of a simulation study is given by:

$$\text{esd} = R^{-1} \sum_{r=1}^{R} \sqrt{m^{-1} \sum_{d=1}^{m} \left( \frac{\bar{y}_{d,r}(T) - \bar{y}_{r}}{\bar{y}_{r}} \right)^2}$$

where $\bar{y}_{r}$ is the mean of the estimated values for the $m$ areas at time $T$ in the simulation run $r$. The comparison is carried out using the indicator

$$\text{RESD} = \frac{\text{esd}}{\text{ESD}} - 1$$

(19)

which tells us how the empirical standard deviation associated with one estimator differs from that of the population.

Table 1 contains the percentage values of AARB, ARMSE, AEFF and RESD obtained for the direct estimator, the design-based estimators given in (2) and (3) and the EBLUP estimators derived from models (5) - (10).

All estimators perform significantly better than $\bar{y}_{d,T,\text{Dir}}$ in terms of ARMSE, leading to less than 100% $\text{AEFF}_{\text{Dir}}$ values. We can also see that design-based estimators are worse than EBLUP estimators in terms of ARMSE, and that the gain in efficiency demonstrated by $\text{AEFF}_{\text{Seq}}$ is particularly high in some cases (in excess of 50%). This result highlights the superior accuracy of the model-based estimators in question.
The most reliable EBLUP estimator is the one associated with the MM5 model, with independent area and time effects and residuals autocorrelated according to an AR(1) process, leading to a gain in efficiency of about 60% compared with the direct estimator. This is followed by the EBLUP estimator associated with model MM1, which differs from the previous one only because of the absence of autocorrelated residuals.

In terms of bias, the GREG estimator gives the smallest value of AARB, as would be expected (Särndal, Swensson and Wretman 1992, chapter 7; Veijanen, Lehtonen and Särndal 2005). This is followed by the remaining estimators, all of which reveal a similar value for AARB. Of the EBLUP estimators, those associated with the MM1 and MM5 models are more efficient in terms of ARMSE, but they are slightly more biased than the one associated with the SMM. This is probably due to the fact that we limit our evaluation of performance to the last wave; for this data subset we would expect the fit of the regression underlying SMM, based on the last wave only, to be better than the one based on the whole data set. As far as EBLUP estimators are concerned, the AEFFsect column shows how the gain in efficiency of the predictors, based on borrowing strength over time, is positive in some cases and negative in others. Models MM2 and MM4 (see formulas (6) and (8)), where effects of interaction between area and time are present, are apparently inadequate because the predictors associated with both models perform rather poorly. The performance of the predictor associated with MM3 (see (7)) is also slightly worse than that of the predictor associated with the cross-sectional model: this rather surprising result is probably due to the low number of waves, which does not allow for an effective estimation of the correlation coefficient between consecutive time effects.

As we have already said, the estimator associated with model MM5 is the one that performs the best: it is considerably more efficient than the one associated with SMM, with an AEFFsect of roughly 85% representing a gain in efficiency of about 15% due to consideration of more than one wave. The EBLUP estimator associated with MM1 also turns out to be more efficient than the one associated with SMM, but in this case the gain is one of only 5%.

These results confirm the fact that household level data at several consecutive points in time may be employed, via certain kinds of longitudinal model, to produce more efficient estimates.

Moving on to the indicator for shrinkage reported in the last column of the table, we can see that the direct estimator overestimate the standard deviation of the population of area means, by 15%. The same effect, albeit somewhat attenuated, is observed for the GREG estimator, whose standard deviation is over-inflated by 10%. On the contrary, the COMP estimator tends to "shrink" the estimates towards the centre of the distribution, leading to a reduction in the standard deviation of area means of about 10% with respect to the population. These results are in line with those obtained by other authors comparing the same kinds of estimator (Heady et al. 2004; Spjotvoll and Thomsen 1987). The results obtained for EBLUP estimators are more encouraging, as the calculated percentage difference is always less than 10% in absolute terms. Hence, in this respect all EBLUP estimators seem to be acceptable. Moreover, we may expect that the BLUP estimators are under-dispersed compared to the corresponding population parameters. In this case, the indicator RESD assumes positive values for some longitudinal EBLUP estimators because it is calculated only on the last wave, while longitudinal models are aimed to predict $m \times T$ parameters.

Table 2 summarizes the results regarding those EBLUP estimators associated with random error variance models, as described in the last paragraph of section 2. When no auxiliary variables are included in the models, the advantage of “borrowing strength” over time and area is singled out independently of the advantage associated with covariates.

As expected, the improvements in efficiency measured by AEFFdir are smaller than those shown in Table 1, although the reductions in ARMSE remain significant. The ranking of those predictors associated with the various random effects specification remains the same as the one presented in Table 1, the predictor associated with the MM5 model resulting the most efficient, as shown by ARMSE%. The gain in efficiency associated with this latter estimator compared with the direct estimator is about 43%.
With regard to bias, the EBLUP estimators obtained from those models with no covariates tend to be more biased than the corresponding ones with covariates.

The analysis of the AEFF$_{sect}$ column shows that the reduction in ARMSE allowed for by some of those models borrowing strength over time, is larger than in the case where covariates are included, as it reaches 22% in the best example of the MM 5$^*$ model.

This last result is really encouraging. In fact, within the context of Small-Area Estimation, the absence of any known totals of covariates in the population can be very limiting when trying to obtain reliable estimates. The observed ARMSE reduction connected to the consideration of more waves in a panel survey show that estimates may be improved “borrowing strength” over time, when it is not possible to exploit auxiliary information.

With regard to the results of shrinkage, they may be considered acceptable also in this case, and one can see a relationship between the results obtained for EBLUP estimators derived from analogous models with or without covariates.

### 4.2 Comparing different estimators of the MSE of EBLUP estimators

In section 2.3 we reviewed three different estimators of the MSE associated with EBLUP estimators. In this section we are going to compare the performances of these three estimators using our simulation exercise. Given that we are focusing on MSE estimation rather than a comparison of EBLUP estimators derived from different models, we only consider the predictor associated with model MM5, which emerged as the best performer in the previous section.

Let us denote the predictor of $\hat{Y}_{dt}$ with $\hat{\eta}_{dt}^{EBLUP}$ and its mean square error as $\text{MSE}(\hat{\eta}_{dt}^{EBLUP})$. The following estimator:

$$
\text{mse}_\text{ACT}(\hat{\eta}_{dt}^{EBLUP}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\eta}_{dt}^{EBLUP} - \hat{\eta}_{dt}^{EBLUP})^2 + (\hat{\eta}_{dt}^{EBLUP} - \bar{Y}_{dt})^2,
$$

where $\hat{\eta}_{dt}^{EBLUP}$ is $\hat{\eta}_{dt}^{EBLUP}$ calculated on the $r$th replicated sample and $\hat{\eta}_{dt}^{EBLUP} = R^{-1} \sum_{r=1}^{R} \hat{\eta}_{dt}^{EBLUP}$, will be used as benchmark for the comparison of the performance of the mean square error estimators described in section 2.3, because the true mean squared error is not known.

As in the case of point estimators, all computations are done using SAS. To determine the Prasad-Rao estimator (14), the output of Proc MIXED’s ESTIMATE statement is used with the option KENWARDROGER activated. The sum $G_1(\psi) + G_2(\psi)$ is obtained from the output of Proc MIXED. The KENWARDROGER option allows for the calculation of an MSE inflation factor, described in Kenward and Rogers (1986), which is equivalent to $2g_1(\psi)$ (see also Rao 2003, section 6.2.7).

The estimator $\text{mse}_{BL}(\hat{\eta}_{dt}^{EBLUP})$ is re-sampling based. Hence the evaluation of its performance with respect to a Monte Carlo exercise requires the implementation of two nested simulations: for each $r (r = 1, \ldots, R)$, we run the $R_{BOOT}$ replications needed to approximate expectations with respect to the bootstrap model. To limit the computational burden, we set $R_{BOOT} = 150$. Butar and Lahiri (2003) propose an analytical approximation of $\text{mse}_{BL}$, but only for models that are not as complex as the one in question.

For both $\text{mse}_{BL}(\hat{\eta}_{dt}^{EBLUP})$ and $\text{mse}_{ILW}^*(\hat{\eta}_{dt}^{EBLUP})$, we have prepared ad-hoc SAS codes using the output of Proc MIXED as inputs.

In order to compare the three MSE estimators, we employ the same measures used to evaluate the performance of point estimators, AARB and ARMSE. As there is usually some concern about the under-estimation of MSE estimators, we are also interested in the sign of any bias associated with the estimators in question. Therefore, in the case of MSE estimators we do not only calculate the average of the absolute values of the estimates obtained for the bias in each region (AARB), but also the average of these estimates without the absolute value (AARB$^*$), so as to better understand whether the given estimators indeed tend to under-evaluate the MSE or not. Hence the calculated measures are:

<table>
<thead>
<tr>
<th>Model</th>
<th>AARB%</th>
<th>ARMSE%</th>
<th>AEFF$_{De}$%</th>
<th>AEFF$_{sect}$%</th>
<th>RESD%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMM$^*$</td>
<td>2.7</td>
<td>0.575</td>
<td>72.8</td>
<td>100.0</td>
<td>-7.6</td>
</tr>
<tr>
<td>MM1$^*$</td>
<td>2.9</td>
<td>0.556</td>
<td>70.3</td>
<td>96.6</td>
<td>7.5</td>
</tr>
<tr>
<td>MM2$^*$</td>
<td>2.8</td>
<td>0.639</td>
<td>80.8</td>
<td>111.0</td>
<td>-3.0</td>
</tr>
<tr>
<td>MM3$^*$</td>
<td>3.7</td>
<td>0.574</td>
<td>72.6</td>
<td>99.7</td>
<td>8.6</td>
</tr>
<tr>
<td>MM4$^*$</td>
<td>3.5</td>
<td>0.691</td>
<td>87.2</td>
<td>119.8</td>
<td>-6.7</td>
</tr>
<tr>
<td>MM5$^*$</td>
<td>3.0</td>
<td>0.445</td>
<td>56.2</td>
<td>77.2</td>
<td>-6.3</td>
</tr>
</tbody>
</table>

Table 2 Performance indicators - auxiliary information is unavailable
where the symbol * refers to the considered estimation procedures, that are PR, BL, JLW. Results of the comparisons based on $R = 500$ MC iterations are reported in Table 3.

| Table 3: Performance of MSE estimators of $\hat{\eta}_{d,t}$ under model MM5 |
|-----------------|--------|--------|--------|
| Estimator       | AARB   | AARB'  | ARMSE  |
| $\text{mse}_{\text{PR}}$ | 0.378  | -0.383 | 0.238  |
| $\text{mse}_{\text{BL}}$   | 0.377  | -0.318 | 0.228  |
| $\text{mse}_{\text{JLW}}$  | 0.337  | 0.036  | 0.261  |

In terms of ARMSE and AARB, the three estimators behave similarly, with no particular one emerging as clearly better than the other two. Nonetheless, the AARB' column behaves similarly, with no particular one emerging as clearly better than the other two. Normality is a crucial assumption in the derivation of $\text{mse}_{\text{PR}}$ and $\text{mse}_{\text{BL}}$, while $\text{mse}_{\text{JLW}}$ could be expected to be more robust in this respect. Our findings are consistent with the theory predictions and simulation results described in Jiang et al. (2002). Although Bell (2001) noted that $\text{mse}_{\text{JLW}}$ may be negative for some data set because of the bias correction, this never happens in our simulations. For all replicated data set we have that the second term in (17) gives a positive, and in most cases substantial contribution to the estimate of the MSE. A discussion of modifications of (17) when it returns negative values can be found in Jiang and Lahiri (2006b).

To conclude then, in the case of the present problem, $\text{mse}_{\text{JLW}}$ emerges as the most appropriate of the three measures for estimating $\text{MSE}(\hat{\eta}_{d,t})$. This finding could be of importance for any application of normality-based linear mixed models theory to data set in which normality assumptions for error terms do not hold exactly.

We replicated the simulation exercise also for the cross-section model without covariates $\text{SSM}^*$, that is often considered in simulations aimed at the comparison of different estimation methods. To this end we note that for this model the ratio $\hat{\sigma}_e^2 / \hat{\sigma}_\gamma^2$ is around 12, leading to a EBLUP predictor characterized by $\gamma_1 = \hat{\sigma}_e^2 n_i / (\hat{\sigma}_\gamma^2 n_i + \hat{\sigma}_e^2)^{-1}$ ranging from 0.54 to 0.9. We note also that some, but not all, areas are characterized by the presence of outliers (skewness coefficient $\gamma_1$ ranges from 0.1 to 4.6).

In this setting MSE estimators show a behavior quite different form that illustrated in the case of model MM5. Results are shown in Table 4.

| Table 4: Performance of MSE estimators of $\hat{\eta}_{d,t}$ under model SSM* |
|-----------------|--------|--------|--------|
| Estimator       | AARB   | AARB'  | ARMSE  |
| $\text{mse}_{\text{PR}}$ | 0.449  | 0.262  | 0.503  |
| $\text{mse}_{\text{BL}}$   | 0.376  | 0.213  | 0.376  |
| $\text{mse}_{\text{JLW}}$  | 0.354  | 0.149  | 0.335  |

All estimators overestimate the actual MSE, although $\text{mse}_{\text{JLW}}$ overestimates less than the other two. From a detailed analysis of results related to individual areas, we have the values of AARB' (that represents the most apparent difference with the results of Table 3) is driven by severe overestimation of actual MSE in areas characterized by the lowest levels of skewness and kurtosis. For these areas $\hat{\sigma}_e^2$ largely overstates actual variation in the data, thus leading to overestimation of $g_1(\hat{\sigma}_e^2, \hat{\sigma}_\gamma^2)$. This is likely to be due to the fact that the failure of normality (the excess of kurtosis) causes the overestimation of $\hat{\sigma}_\gamma^2$. This problem did not appear in the case of model MM5 because of the presence of covariates and the AR(1) modeling of individual residuals.

5. Concluding remarks and further developments

The results obtained show that, in general, EBLUP estimators derived from unit level linear mixed model specifications that “borrow strength over time”, as well as over areas, provide a significant gain in efficiency compared with both the direct estimator and with other commonly-used design based estimators such as the optimal composite estimator and the GREG estimator. Moreover, the mean squared error of some of the longitudinal EBLUP estimators in question is considerably lower, on average over the areas, than that of the analogous cross-sectional EBLUP estimators. Among the model specifications used to derive EBLUP estimators, those with independent time and area effects, whether inclusive of the autocorrelation of residuals or not, appear the most efficient, offering a gain in efficiency of about 55-60% compared with the direct estimator. These results also hold when covariates are removed; in fact, they offer the chance to obtain reliable small area estimates even in the absence of covariates, provided that repeated observations of the same unit at several points in time are available. Besides the shrinkage
effect connected to EBLUP estimators appears moderate, reducing the need for ensemble or multiple estimation (Rao 2003, Chapter 9). With regard to estimation of the MSE of the small area estimators in question, we noted that the jackknife estimator provides the best results being correct, on average, over the areas and thus more robust to any departure from the standard assumptions of linear mixed models. This finding may be of importance to all applications of normality-based linear mixed models to data set in which normality assumptions do not hold exactly, as in the case of income data.

Acknowledgements

Research was partially funded by Miur-PRIN 2003 “Statistical analysis of changes of the Italian productive sectors and their territorial structure”, coordinator Prof. C. Filippucci. The work of Enrico Fabrizi was partially supported by the grants 60FABR06 and 60BIFF04, University of Bergamo.

We thank ISTAT for kindly providing the data used in this work.

References


Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results

Anne Renaud

Abstract
Coverage deficiencies are estimated and analysed for the 2000 population census in Switzerland. For the undercoverage component, the estimation is based on a sample independent of the census and a match with the census. For the overcoverage component, the estimation is based on a sample drawn from the census list and a match with the rest of the census. The over- and undercoverage components are then combined to obtain an estimate of the resulting net coverage. This estimate is based on a capture-recapture model, named the dual system, combined with a synthetic model. The estimators are calculated for the full population and different subgroups, with a variance estimated by a stratified jackknife. The coverage analyses are supplemented by a study of matches between the independent sample and the census in order to determine potential errors of measurement and location in the census data.

Key Words: Census; Coverage errors; Dual system; Multi-stage sampling plan; Measurement errors.

1. Introduction

In any census, some people are not enumerated and should be, while others are counted twice or should not have been enumerated. There is both undercoverage and overcoverage, and quite often, the combined result is net undercoverage. For example, net undercoverage is estimated at 1.6% in the United States in 1990 (Hogan 1993), 2.2% in the United Kingdom in 1991 (Brown, Diamond, Chambers, Buckner and Teague 1999) and 3% in Canada in 2001 (Statistics Canada 2004). By contrast, in the United States in 2000, there is estimated to be net overcoverage of 0.5% (Hogan 2003). Coverage deficiencies may vary greatly between subgroups of the population. In the United States in 2000, blacks were found to have a net undercoverage of 1.8%, while whites had an overcoverage of 1.1%. Also, values often vary between age classes and regions, for example. These coverage deficiencies, and other errors such as measurement errors, result in a biased picture of the population. They are therefore studied in order to obtain information on the quality of the available data and to find ways to improve censuses of the population.

The 2000 population census in Switzerland gives a picture of the population on December 5, 2000. In this article, coverage deficiencies in a Swiss census are estimated for the first time. Undercoverage, overcoverage and net coverage resulting from the 2000 census are all analysed. Undercoverage is estimated from a sample of individuals \( S_p \), independent of the census, on which a coverage survey was organized a few months after the census (collection took place in April and May 2001). The data from the survey are matched with data from the census to determine whether persons in \( S_p \) were enumerated.

Overcoverage is estimated from a sample of individuals \( S_E \) drawn from census records. A search for duplicates and other erroneous records then serves to determine whether a given record corresponds to a real person to be enumerated. Net coverage is estimated on the basis of a capture-recapture model known as the dual system (Wolter 1986, Fienberg 1992). The dual estimator is applied in homogeneous cells, and the results are recombined using a synthetic model to obtain results for different domains of the population (Hogan 2003). The purpose of the project is not to adjust the census figures but rather to obtain information on the quality of the 2000 census and potential improvements for future censuses.

This article describes the different steps followed in obtaining estimates, then presents the results. Sections 2 and 3 describe the data sets and the coverage estimators. Section 4 provides the details on constructing the different statuses used in the estimators. Section 5 describes the approach used to compare the values collected in the census and in the survey for the matched persons from \( S_p \). Sections 6 and 7 present the numerical results and the conclusion.

2. The three data sets

2.1 Census

The 2000 census was conducted under the auspices of the Federal Statistical Office, with the reference date of December 5, 2000. Information was collected for 7.3 million inhabitants, 3.1 million households, 3.8 million dwellings and 1.5 million buildings. The different levels were then linked by common identifiers when the data were processed.
The collection of information on persons and households was the responsibility of Switzerland’s 2,896 political communes. The latter had a choice between different methods of collection:
- TRADITIONAL: use of census agents;
- SEMI-TRADITIONAL: pre-printed questionnaires based on the communal register of inhabitants are mailed out and then collected by census agents;
- TRANSIT: pre-printed questionnaires are mailed out and mailed back;
- FUTURE: identical to TRANSIT except that links between households and dwellings are supplied by the commune;
- TICINO: similar to TRANSIT but limited to the canton of Tessin.

Most of the SEMI-TRADITIONAL, TRANSIT, FUTURE and TICINO communes also offered the option of completing questionnaires online. The 2,208 SEMI-TRADITIONAL, TRANSIT, FUTURE and TICINO communes that used the pre-printing of questionnaires based on communal registers of inhabitants account for nearly 96% of the population. For most of these communes, the tasks of mailing out questionnaires and controlling their return were organized at a national centre.

The data set for individuals contains 7,452,075 entries. One feature of this data set is that it contains two records for the same person if that person has two residences (2.3% of the population; for example, a student who both resides with his parents and has a residence close to his school). In the case of two residences, one is coded as the economic residence and the other as the civil residence. The economic residence is the place where the person spends the most time per week and the civil residence is where the person’s official papers are kept (birth certificate for Swiss citizens, residence permit for foreigners). Where there is just one residence, it is both the economic and the civil residence. Switzerland is considered to have a resident population of 7,280,010 based, on the set of records showing the economic residence.

Households are classified as private, collective or administrative. Examples of private households are families, couples and persons living alone. Examples of collective households are groups of occupants of a home for the aged or a boarding school or the inmates of a prison. Administrative households group together people with no fixed residence, travellers and persons - by building or commune - who could not be assigned to private or collective households (2.4% of the resident population).

Census data contain no imputation at the record level, since communes sent basic information for non-respondents (unit non-response). However, values are imputed in the case of missing data or inconsistency in questionnaires (item non-response).

The population of interest for coverage estimates is the resident population (based on economic residence) in private and administrative households. Collective households, which account for 2.3% of the enumerated resident population, are excluded from the estimates.

2.2 \( S_p \) sample, coverage survey and matching (undercoverage)

The size objective for the \( S_p \) sample is set at approximately 50,000 people. In the absence of existing frames in Switzerland, this value was determined approximately, based on experiences in other countries. In particular, the Australian results for 1996 were used, since the sampling plan for Australia’s coverage survey was similar to the one for Switzerland in 2000 (ABS 1997).

The \( S_p \) sample, which is independent of the census, is constructed in two parts: the canton of Tessin (TICINO) and the rest of Switzerland (NORD). Both parts use a multi-stage draw. The first stage consists in selecting 303 primary units - these are political communes for TICINO and postal codes for NORD - according to a stratified plan with a draw proportional to the number of buildings. The second stage consists in a simple random draw of a fixed number of 60 buildings per primary unit. In the NORD plan, these buildings are allocated to a maximum of three mail delivery routes, based on an intermediate sampling stage. The sampling is thus constructed so as to consolidate the field work while limiting the variability of the weights. For practical reasons and in light of available resources, postal codes that include a large proportion of buildings lacking complete postal addresses or coded as unoccupied are selected with a lower probability than other postal codes. These tend mainly to be postal codes in rural areas or industrial zones, which are unlikely to exhibit major coverage deficiencies. With the assistance of postal employees, complete lists of households are drawn up in the field within the sample of approximately 16,000 buildings. A sub-sample of buildings is then drawn so as to obtain a total of approximately 27,000 households. For more information on the sampling and survey procedure, see Renaud (2001) and, in greater detail, Renaud and Eichenberger (2002).

The coverage survey consists in contacting the 27,000 households - by telephone if a telephone number is found and in person if not. The variables collected are those that lend themselves to matching with the census and defining subgroups of interest for the coverage study (socio-demographic variables, addresses). The collection operation covers all members of all households in the selected buildings.
The final sample \( S_p \) contains \( n_p = 49,883 \) people in the population of interest (persons listed at their economic residence and residing in a private household). Of the households contacted, 88% were reached by telephone and 12% in person. The weighting depended on the sampling and an adjustment for non-response. The adjustment for non-response was based on a homogeneity model in cells constructed on the basis of the sampling strata and whether or not a telephone number was known to exist (interviews conducted by telephone or in person). It also incorporated an estimate of the proportion of true households among the households to be contacted, since a sizable portion of the households to be contacted actually consisted of vacant dwellings, stores or businesses. No calibration was applied, since the auxiliary data available were not independent of the census. There was no partial non-response. The weighting details are documented in Renaud and Potterat (2004).

Based on the questions asked in the survey and various plausibility controls, we hypothesize that the \( S_p \) data are correct and usable for matching with the census. The quality criteria used are as follows:

- completeness: the record is sufficient to identify the person;
- appropriateness: the person should have been enumerated;
- uniqueness: the person is listed only once;
- belonging to population of interest: the person is listed at his/her economic residence and in a private household;
- correctness of location: the person is listed at the correct address on Census Day.

The matching between the \( S_p \) sample and the census serves to determine the matching status \( P_j \) of each element \( j \) of \( S_p \). Status \( P_j \) is equal to 1 if the element is matched in the census (enumerated person) and 0 if this is not the case (person not enumerated). In our case, the data collected in the coverage survey, the final census data and images of the census questionnaires are used for automatic matching, supplemented by an analysis of suspect cases. An element \( j \) is considered unique if no duplicate or triplicate is detected in the census. There is no supplementary interview for \( S_E \). There is therefore no information supplementary to the census for \( S_E \) persons (actual location? actual type of residence or household?). The search for duplicates/triplicates and suspect cases results in an enumeration status \( E_i \) for each element \( j \) of \( S_E \). Status \( E_i \) is equal to 1 if the element should indeed have been enumerated in the census (default value) and 0 if it should not have been enumerated. In practice, it can take on values between 0 and 1 if the case is not determined precisely. Thus, duplicates and triplicates receive respectively the values 1/2 and 1/3 if there is no information allowing the correct record to be determined from among the records detected. These cases, which are rare, consist of persons who completed more than one questionnaire in the census without any link having been made between those questionnaires during the processing of the data.

The \( S_E \) sample was selected from the census data using a two-stage draw. Only elements included in the population of interest were eligible (records at the economic residence, without members of collective households). The primary units of \( S_E \) were identical to the primary units of \( S_p \) (postal codes and communes). However, the list of postal codes in the NORD plan that was used for \( S_p \) did not correspond exactly to the list of postal codes that were present in the census data. Census records found in postal codes that did not exist in the list used for \( S_p \) were therefore reallocated to existing codes, taking geographic location into account (this involved assigning fictitious postal codes for the sampling). In the second stage, records were drawn from the population of interest using a simple random plan, without intermediate stages. The allocation was done in such a way as to obtain constant weights in the sampling strata of the primary units. In the end, the sample contained \( n_E = 55,375 \) records (Renaud 2003).

We hypothesize that \( S_E \) records are sufficient to identify persons (completeness), since there is little imputation in the census data and most questionnaires were pre-printed based on registers of inhabitants. Appropriateness and uniqueness were determined in a matching between \( S_E \) and the rest of the census using a procedure similar to the matching between \( S_p \) and the census. In our case, this involves a search for duplicates or triplicates of elements of \( S_E \), supplemented by an analysis of suspect cases in \( S_E \). An element \( j \) is considered appropriate if it is not considered erroneous in the analysis of suspect cases (e.g., a note on the questionnaire indicating that the person has gone abroad). An element \( j \) is considered unique if no duplicate or triplicate is detected in the census. There is no supplementary interview for \( S_E \). There is therefore no information supplementary to the census for \( S_E \) persons (actual location? actual type of residence or household?).
3. Coverage estimators

3.1 Undercoverage and overcoverage

The undercoverage rate is estimated by \( \hat{R}_{\text{under}} = 1 - \hat{R}_m \), where \( \hat{R}_m \) is the estimate of the correct matches rate based on the \( S_p \) sample. Similarly, the overcoverage rate is defined as \( \hat{R}_{\text{over}} = 1 - \hat{R}_c \), where \( \hat{R}_c \) is the estimate of the correct records rate based on the \( S_p \) sample. The correct matches rate and the correct records rate are estimated by the weighted means of matching status \( P_j \) and enumeration status \( E_j \), as follows:

\[
\hat{R}_m = \frac{\sum_{j \in S_p} w_{p,j} P_j}{\sum_{j \in S_p} w_{p,j}} \quad \text{and} \quad \hat{R}_c = \frac{\sum_{j \in S_p} w_{E,j} E_j}{\sum_{j \in S_p} w_{E,j}},
\]

where \( w_{p,j} \) is the weight of element \( j \) of sample \( S_p \) and \( w_{E,j} \) is the weight of element \( j \) of sample \( S_E \). We note that the denominator of \( \hat{R}_m \) is the sum of the weights \( w_{E,j} \) of \( S_E \) and not the number \( C \) of known records in the census, so as to have a potentially less biased estimator.

The estimate of the undercoverage and overcoverage rates in a domain \( d \) is given by \( \hat{R}_{\text{under},d} = 1 - \hat{R}_{m,d} \) and \( \hat{R}_{\text{over},d} = 1 - \hat{R}_{c,d} \), with

\[
\hat{R}_{m,d} = \frac{\sum_{j \in S_p} w_{p,j} P_{j,d}}{\sum_{j \in S_p} w_{p,j}} \quad \text{and} \quad \hat{R}_{c,d} = \frac{\sum_{j \in S_p} w_{E,j} E_{j,d}}{\sum_{j \in S_p} w_{E,j}},
\]

Identifiers \( I_{j,d} \) and \( J_{j,d} \) take the value 1 if element \( j \), respectively of \( S_p \) and \( S_E \), is found in domain \( d \); otherwise their value is 0.

3.2 Net coverage

The net undercoverage rate is estimated by \( \hat{R}_{\text{net under}} = 1 - \hat{R}_{\text{net}} \) where \( \hat{R}_{\text{net}} = C / \hat{N} \) is the estimate of the net coverage rate, \( C \) is the number enumerated in the population of interest and \( \hat{N} \) is the estimate of the true total in the population of interest. If \( \hat{R}_{\text{net under}} \) is negative, there is net overcoverage.

The estimate of the true total \( \hat{N} \) is based on the dual model (Wolter 1986). This model is built on the principle of capture (census) and recapture (coverage survey). It is applied in estimation cells \( k = 1, \ldots, K \) in order to best satisfy the assumptions of the model; see discussion below. Thus, the estimate of the true total \( \hat{N} \) is composed of the sum of the estimated true totals \( \hat{N}_k \) in disjoint estimation cells covering the population of interest \( k = 1, \ldots, K \):

\[
\hat{N} = \sum_{k=1}^{K} \hat{N}_k.
\]

The estimated totals \( \hat{N}_k \) have the form given by the dual model:

\[
\hat{N}_k = [N_{1r,k} \left[ \frac{N_{l+1,k}}{N_{l+1,k}} \right]]
\]

where \( N_{1r,k} \) is the total of records correctly counted in cell \( k \) during capture (census), \( N_{l+1,k} \) is the total in \( k \) during recapture (estimated from sample \( S_p \)) and \( N_{l+1,k} \) is the number of records common to the two lists (estimated from matches between \( S_p \) and the census).

The different terms of equation (4) are estimated using undercoverage and overcoverage estimates. This is an extension of the model in Wolter (1986), similar to the one used by Hogan (2003). Thus, the total of the records correctly counted in the census \( N_{1r,k} \) is estimated by the enumerated total \( C_k \) multiplied by the correct records rate \( \hat{R}_{c,k} \) to take account of overcoverage. Also, the ratio between the total in the recapture \( N_{l+1,k} \) and the number of records common to the two lists \( N_{l+1,k} \) is estimated by the inverse of the rate of matching \( \hat{R}_{m,k} \) between the coverage survey and the census in order to take account of undercoverage. We obtain

\[
\hat{N}_k = [C_k \hat{R}_{c,k}^{-1}] = C_k [\hat{R}_{c,k}^{-1} \hat{R}_{m,k}^{-1}] = C_k \hat{F}_k
\]

where \( \hat{F}_k = \hat{R}_{c,k}^{-1} \hat{R}_{m,k}^{-1} \) is the coverage correction factor in cell \( k \). Factor \( \hat{F}_k \) combines the effects of overcoverage and undercoverage of cell \( k \) estimated from samples \( S_p \) and \( S_E \). We note that undercoverage in one domain may be offset by overcoverage in the same domain. Thus, nil net undercoverage in a domain does not mean that no coverage deficiency exists in it.

The proposed estimates are based on the assumptions of the dual model, the choice of estimation cells, and the choice of the statuses defining the estimators \( \hat{R}_{c,k} \) and \( \hat{R}_{m,k} \). The dual model is useful since it takes into account the fact that some persons are reached neither by the census (capture) nor by the coverage survey (recapture). However, a series of conditions must be met to avoid estimation biases. The coverage survey and the census must be totally independent. The matching must be of very high quality. The model must be applied in cells with persons who have the same probability of being enumerated in the census and the survey respectively; see Section 3.3. Lastly, the population must not change too much between Census Day and the day of the survey. As to the estimators \( \hat{R}_{c,k} \) and \( \hat{R}_{m,k} \), they are based on the quality of the matching and the search for erroneous records. Also, it is necessary to ensure that the definition of a correct match in \( S_p \) and the definition of a correct record in \( S_E \) are identical, i.e., there is a balance between overcoverage and undercoverage; see Section 4. All those elements are taken into consideration insofar as possible in the present estimates.

The estimate of net undercoverage in a domain \( d \) has the form \( \hat{R}_{\text{net under},d} = 1 - \hat{R}_{\text{net},d} = 1 - C_d / \hat{N}_d \), where \( C_d \) is
the enumerated number in the domain and $\hat{N}_d$ is the estimate of the true total. The estimate of the true total $\hat{N}_d$ is based on a synthetic model that assumes that the correction factor is fixed in each cell $k = 1, ..., K$:

$$\hat{N}_d = \sum_{k=1}^{K} \hat{N}_{k,d} = \sum_{k=1}^{K} C_{k,d} \hat{F}_k.$$  \hfill (6)

$C_{k,d}$ is the number enumerated in the population of interest in the intersection between cell $k$ and domain $d$, and $\hat{F}_k$ is the correction factor for the coverage in cell $k$. The hypothesis of the synthetic model is satisfied if the behaviour of any subset in the cell is identical to that of the entire cell. This homogeneity is best controlled by the choice of cells. Here we are using the homogeneous cells defined by the dual model.

### 3.3 Estimation cells

The estimation cells $k = 1, ..., K$ are constructed in such a way as to group together elements that have homogeneous probabilities of enumeration in the census and the survey respectively (dual hypothesis) and homogeneous net coverage rates (synthetic hypothesis). We want a minimum of 100 persons per cell in $S_p$ and $S_E$ in order to control the variance and limit the estimation bias. The variables defining the cells are selected using a logistic regression model and a discrimination method applied to the data from $S_p$ (binary variable: $P_i$). The three most influential variables are cross-tabulated: nationality in two categories, marital status in two categories and size of commune in three categories. The other variables are then successively integrated. Groupings are created when the cell sizes are too small (official language of commune in two categories, age class in seven categories and sex in two categories). In the end, 121 estimation cells are obtained; see Renaud (2004) for more details.

### 3.4 Variance of coverage estimators

The variance of the estimators is estimated by a stratified jackknife applied to the (identical) primary units of $S_p$ and $S_E$. We note that the variance of the estimated under-coverage $\hat{R}_{under} = 1 - \hat{R}_m$ is equal to the variance of the estimated matching rate $\hat{R}_m$. Similarly, the variance of the overcoverage $\hat{R}_{over} = 1 - \hat{R}_c$ is equal to that of the correct record rate $\hat{R}_c$, and the variance of the net undercoverage $\hat{R}_{netunder} = 1 - \hat{R}_{net}$ is equal to that of the net coverage rate $\hat{R}_{net}$.

Let $\theta$ be the parameter of interest taking the form of a weighted mean of statuses in the case of undercoverage and overcoverage, and the form of a linear function of quotients between two weighted means in the case of net undercoverage. Its estimator is $\hat{\theta}$.

Let $h = 1, ..., H$ be the stratum used in the first stage of sampling, $i = 1, ..., m_h$ the number of the primary unit in stratum $h$ (postal code for NORD or commune for TICINO), and $j = 1, ..., n_{hi}$ the number of the person in primary unit $i$ of $h$. For the needs of the jackknife method, samples $S_p$ and $S_E$ are partitioned, in each stratum $h$, in $m_h$ subsets corresponding to the persons in primary units $\alpha = 1, ..., m_h$.

Let $\hat{\theta}_{(ha)}$ be the estimator having the same form as $\hat{\theta}$ but calculated on the sample from which primary unit $\alpha$ of stratum $h$ has been removed. We note that estimators $\hat{R}_{m(ha),k}$ and $\hat{R}_{c(ha),k}, k = 1, ..., K$ are combined to form $\hat{R}_{net(ha)}$:

$$\hat{R}_{net(ha)} = C \left[ \sum_{k=1}^{K} C_k \hat{R}_{m(ha),k} \right]^{-1}.$$  \hfill (7)

The corrected weights $w'_{hi}$ used to calculate values $\hat{R}_{m(ha)}$ and $\hat{R}_{c(ha)}$ have the following form:

$$w'_{hi} = \begin{cases} 0 & \text{if } i = \alpha \\ w_{hi} / m_h & \text{if } \alpha \in h \text{ and } i \neq \alpha \\ w_{hi} & \text{if } \alpha \notin h. \end{cases}$$  \hfill (8)

This form of correction is preferred to the quotient between the sum of the weights of the elements in the stratum and the sum of the weights without primary unit $\alpha$ since it allows us to take account of the variability due to the unknown number of elements in the stratum.

The jackknife estimator becomes:

$$\hat{\theta}_{JK} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_{(ha),k},$$  \hfill (9)

with pseudo values $\hat{\theta}_{(ha)} = m_h \hat{\theta} - (m_h - 1) \hat{\theta}_{(ha)}$. The estimator of its variance can take different forms; see the example of Shao and Tu (1995). We apply the following form:

$$v(\hat{\theta}_{JK}) = \sum_{h} \frac{m_h - 1}{m_h} \sum_{\alpha \in h} \left( \hat{\theta}_{(ha)} - \hat{\theta}_{(h)} \right)^2,$$  \hfill (10)

with $\hat{\theta}_{(h)} = \sum_{\alpha \in h} \hat{\theta}_{(ha)}/m_h$. Lastly, we use $v(\hat{\theta}_{JK})$ as an estimator of the variance of $\hat{\theta}$. The estimates in the subgroups use the same form of estimator with integration of a domain indicator in the construction of $\hat{\theta}_{(ha)}$. No correction for the finite population is applied in the estimates. Also, other variabilities are not taken into account, such as the variability induced by the weighting model for non-response in $S_p$.

Problems, such as the lack of stability of estimation in strata with few primary units, appeared in the course of applying this approach. However, tests on the sharing of some primary units and a comparison with the Taylor
linearization or a simple jackknife suggest that the estimators of variance by stratified jackknife that are presented in this document are fairly conservative.

4. Choice of correct matching statuses and enumerations

A key element of coverage estimates is the definition of the correct matching status for the elements of $S_P$ and the correct enumeration status for the elements of $S_E$. These correct statuses are defined on the basis of frames $P_j$ and $E_j$, determined during the matchings.

Is a match with a census element that is part of a collective household accepted as a correct match for an element of $S_P$, or is this a case of undercoverage of the population of interest? Is a duplicate outside the population of interest for an element of $S_E$ really considered a duplicate, and hence an instance of overcoverage, or should it be excluded? A clear definition is needed. Also, the statuses used in estimates of net undercoverage must be chosen in such a way as to satisfy the balance between over- and undercoverage; see the concept of “balancing,” as, for example, in Hogan (2003). A match $(P_j = 1)$ with an element outside the population of interest may, for example, be rejected as a correct match (correct match status $= 0$, no undercoverage) only if the search for correct records would also detect this element as incorrect because it is out of scope (correct enumeration status $= 0$, no overcoverage).

The criteria for defining correct statuses are constructed using information available for elements of $S_P$ and $S_E$. As regards $S_P$, we start with the assumption that census records that were matched with elements of $S_P$ serve to identify persons (completeness) and these persons should indeed have been enumerated (appropriateness). We also consider that they are unique, since uniqueness, while controlled by matches, is achieved in the great majority of cases controlled in $S_E$. The criteria of belonging to the population and correctness of location are controlled by comparison with the information collected in the coverage survey, considered as reference information. No supplementary data collection was organized to resolve ambiguous cases. As regards $S_E$, we have the criterion of completeness considered as having been met in the census data and the results concerning uniqueness and appropriateness obtained in the matching with the rest of the census. For duplicates and triplicates, we define $E_j = 1/d' \times d'$, with $d'$ = number of duplicates/triplicates in the population of interest according to the census. The criteria of belonging to the population of interest and correctness of location for the elements of $S_E$ cannot be controlled, since we have no reference data supplementing the census.

For estimates of net undercoverage, it is important to meet the balancing requirement. The criteria used in defining correct statuses are thus completeness, appropriateness and uniqueness. The criteria of belonging to the population of interest and correctness of location cannot be considered, since they are not usable in defining the correct enumeration status. The criteria of completeness, appropriateness and uniqueness are already integrated into the construction of the basic statuses $P_j$ and $E_j$. Thus, estimates are made with basic statuses $P_j$ and $E_j$.

For estimates not using the dual system and the need for balancing, it is possible to use other criteria to define correct statuses. Other types of correct match statuses are used in the analysis of potential measurement errors in Section 5 and the more detailed analyses of matches and enumerations presented in Renaud (2004).

5. Comparison of matches

5.1 Potential measurement errors

Measurement errors or classification errors are related to coverage errors. A person who is classified in domain $d$ according to the census (e.g., a person between 10 and 19 years of age) but who in reality is outside the domain (e.g., a person 60 years of age) would end up as a case of overcoverage in domain $d$ and an undercoverage case outside that domain. This misclassification does not cause a coverage error at the overall level, but it causes an error at the level of subgroups of the population.

The reasons for differences between the values collected in two surveys such as the census and the coverage survey may be quite varied and difficult to dissociate. It is to be expected that there will be matching errors, differences resulting from collection methods (paper questionnaire or telephone/faceto-face interview) and data processing methods, or real differences due to the time lag between the collection periods (December 2000 and April-May 2001). Also, it is difficult to determine the correct response if there are two different values. What is the correct choice - the census? the survey? another value not collected?

Potential measurement errors with respect to census data are analysed on the basis of a set of matches between the independent sample $S_P$ and the census. We choose to determine which variables show respectively few or many potential classification problems, without making a judgment on the quality of either data collection. One use of this information is to evaluate the choice of estimation cells for the dual system and select subgroups for which the estimates of coverage deficiencies are soundest.

For the category variable $X$, we define the matching rate in the good domain $R_X$ as follows:
\[ \hat{R}_d = \frac{\sum_{j \in \text{S}_p \text{match}} w_{p,j} X_{j,i}}{\sum_{j \in \text{S}_p \text{match}} w_{p,j}}, \]  

(11)

where \( w_{p,j} \) is the weight of element \( j \) of matched sample \( S_p \) (\( S_p \text{match} \)) and the classification status \( P_{x,i} \) is equal to 1 if element \( j \) appears in the same class in the census and the survey, and 0 otherwise. The value of \( \hat{R}_d \) is estimated with the set of matched elements and with the subgroup of elements without imputation in the census.

We also define a measure of asymmetry \( \varphi_X(d, d') \) for classes \( d \) and \( d' \) of variable \( X \):

\[ \varphi_X(d, d') = \frac{\sum_{j \in \text{S}_p \text{match}} w_{p,j} I_j(d, d')}{\sum_{j \in \text{S}_p \text{match}} w_{p,j} I_j(d', d)}, \]  

(12)

where \( I_j(d, d') = 1 \) if element \( j \) appears in domain \( d \) according to the survey and in domain \( d' \) according to the census, and 0 otherwise. The factor \( \varphi_X(d, d') \) is equal to 1 if there is a balance in the classification errors - in other words, if the number of elements in \( d \) according to the survey and \( d' \) according to the census is equal to the number in \( d' \) according to the survey and in \( d \) according to the census. The further the factor lies from 1, the less balance there is.

### 5.2 Potential location errors

Comparisons between the census and the survey can also be used to study people’s geographic location. In the census data, we have a unique address if the person has a single residence and two addresses - principal and secondary - if the person has two residences. In the survey data, we have one or two addresses on Census Day, one or two addresses on the day of the survey and information on a possible move between the two dates. If a person has a single residence and has not moved, that person’s principle address on Census Day and his/her principle address on the day of the survey are identical. The person does not have secondary addresses.

Different measures of distance are considered in order to determine potential location errors in the census. For practical reasons, including the data available, we define geographic areas around the person’s principle address collected in the survey for Census Day (reference address). The areas are sets of political communes. They are defined on the bases of postal codes identified in the survey. The person’s basic area is defined by the set of communes that have buildings within the postal code of the person’s reference address. The definition of this area uses data from the Swiss building register, since the latter has information on buildings’ postal address and the commune within which they are located. The extended area includes the communes within the basic area and the set of communes adjacent to them; see Renaud (2004) for examples.

Like classification errors, location errors do not cause coverage errors at the overall level but they cause errors at the level of subgroups such as regions or types of communes. Different rates may be defined. We will retain the basic location rate and the extended location rate, both weighted by \( w_p \), the weight of element \( j \) of matched sample \( S_p \). The location status takes on the value 1 if the element lies within the basic area or the extended area, as the case may be; otherwise it equals 0. In particular, we will study the correctness of location of persons who have moved, in order to detect possible problems relating to the time lag between Census Day and the actual day of collection of census data.

### 6. Results

#### 6.1 Estimates of coverage deficiencies

The overall net undercoverage rate is estimated at 1.41% with a standard deviation of 0.12%. The overcoverage rate is 0.35% (standard deviation = 0.03%) and the undercoverage rate is 1.64% (standard deviation = 0.11%). These results are of the same order of magnitude as those of other countries, although they are in the lower range; see Table 1.

Overcoverage is minor in the great majority of the domains studied. The highest rate is observed for persons between 20 and 31 years of age (0.93% with a standard deviation of 0.09%); see Table 2. However, undercoverage is high in several domains. For example, a rate of 8.03% (standard deviation = 0.85%) is observed for foreigners with temporary settlement permits (“other permits”) and a rate of 3.50% (standard deviation = 0.50%) is observed for 20-31-year-olds. Also, an undercoverage rate of 2.4% is observed in the Italian-speaking region of the country (language of commune: Italian; NUTS region: Ticino, and collection method: TICINO). However, the results are related to relatively great variability (standard deviation of approx. 0.5%), since samples \( S_p \) and \( S_e \) include only 1,500 and 1,700 persons respectively in this region.

Net undercoverage is positive in all the domains studied. There is therefore no net overcoverage. The highest values are observed for foreigners with permanent or temporary permits (2.89% and 3.48%, standard deviations = 0.32% and 0.39%) as well as for 20-31-year-olds (2.84%, standard deviation = 0.36%). No significant difference is observed between males and females, between languages or between NUTS regions. Because of the small size of the sample with the collection variant TICINO, this method cannot be differentiated from the others used in the country. On the other hand, significant differences are observed between marital statuses, as well as between types and sizes of communes.
We note that the net undercoverage rate is greater than the undercoverage rate in the case of permanent settlement permits. This effect, which is unrealistic, is due to the choice of estimation cells and the resulting smoothing. The construction of the cells made it necessary to group foreigners with permanent and temporary permits into a single category for aggregates so as to obtain the minimum size of 100 persons per cell. By making this grouping, we are treating foreigners as a homogeneous group, whereas this is not the case. This shows the limitations of the method and the difficulty of satisfying the assumptions of the models used in applying the approach. In the case of foreigners, we note, however, that the confidence intervals of the undercoverage and net undercoverage rates overlap. The consequences of the weaknesses of the application are therefore limited.

It should also be noted that the results are presented in domains defined by variables for which low levels of potential measurement errors were observed. The fact is that results for groups as defined by household or labour market characteristics would not be very reliable; see Section 6.2.

The precision of the results obtained is generally better than the objective set at the beginning of the project. That objective was to have a standard deviation of 0.3% for subgroups of 10,000 individuals in SP. In the case of, for example, age classes 32-44 and 45-59, which have between 10,000 and 12,000 persons, the standard deviations are 0.19% and 0.14%.

### 6.2 Potential measurement and classification errors

Of the 49,107 elements matched between the coverage survey and the census, 96% exhibit no difference in sex, the seven age classes, the three marital status classes and the three settlement permit classes (Swiss, permanent, temporary). The matching rate in the good domain $R_g$ is 99.3 % for sex (with and without imputations), 98.3% for marital status (98.4% for non-imputed values) and 98.7% for settlement permits (98.8% for non-imputed values). The $R_g$ rate is 99.5% for age classes (with or without imputations). However, it should be noted that date of birth, along with surname and given name, was one of the main variables in the matching. Age differences are therefore possible only in the case of a non-automatic (computer-assisted or manual) matching. Three variables exhibit a matching rate in the good domain that is markedly lower than that observed for sex, age, permit and marital status. These are the variables for labour market status (in the labour market, unemployed, not in the labour market), position in household (alone, spouse, common-law union, person with child or children, other head of household, related to head of household, other; results limited to private households), and size of the person’s household (according to economic residence and in private households). The $R_g$ rate is 90.4% for labour force status (91.1% for non-imputed values), 91.4% for position in household (94.9% for non-imputed values) and 88.3% for household size.

The measure of asymmetry $\phi_{d,d'}(d, d')$ takes on the value 1.33 for sex ($d = \text{male}$ and $d' = \text{female}$). There are more persons coded as males according to the survey who are coded as females in the census than there are females according to the survey who are males according to the census. The proportion of males is slightly higher in the survey. However, these results must be interpreted with caution, since they based on very few cases; see Table 3. A McNemar test is just significant at the 5% level without taking the design into account, but it is no longer significant at that level if the design is factored in. On the other hand, quite substantial asymmetries are observed for marital status. There are fewer single persons in the survey who are married in the census than the reverse (factor 0.33 for $d = \text{single}$ and $d' = \text{married}$). Similarly, there are fewer married persons in the survey who are widowed in the census than the reverse (factor 0.42 for $d = \text{married}$ and $d' = \text{other}$). Asymmetry is also observed for the settlement permit variable. The tendency is to have more Swiss persons in the survey who are described as foreigners in the census than the reverse, and to have more permanent permits in the survey and temporary permits in the census than the reverse (factors 5.22 for $d = \text{Swiss}$ and $d' = \text{foreigner}$ with permanent permit and 3.83 for $d = \text{foreigner}$ with permanent permit and $d' = \text{foreigner}$ with temporary permit). The factors calculated are based on few cases. However, they give an insight into the potential differences between data collection via the census questionnaire and a survey conducted mainly by telephone. The labour force status variable includes more divergent cases; see Table 4. Thus, for example, we observe fewer persons employed force in the survey and fewer persons not in the labour force census than the reverse (factor of 0.46 for $d = \text{in labour
force and \( d' \) = not in labour force). There are also fewer unemployed persons in the survey and persons not in the labour force in the census than the reverse (factor of 0.26 for \( d = \) unemployed and \( d' = \) not in labour force). The position-in-household variable also exhibits asymmetries, but these are based on few elements, since the dispersion of the elements in the boxes \( (d, d') \) is sizable. The census variables at the household level (position in household and size of household) are influenced by the complex process of household formation. They are less reliable than those concerning persons. The values at the household level are more reliable in the survey.

| Table 2 | Enumerated number \( C \) and estimated rates of overcoverage \( \hat{R}_{\text{over}} \), undercoverage \( \hat{R}_{\text{under}} \) and net undercoverage \( \hat{R}_{\text{netunder}} \) for different domains [%], with corresponding estimated standard deviations (SDs) |
|---------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Variable | Category          | \( C \)          | \( \hat{R}_{\text{over}} \) | SD              | \( \hat{R}_{\text{under}} \) | SD              | \( \hat{R}_{\text{netunder}} \) | SD              |
| Overall  |                  | 7,121,626        | 0.35             | 0.03            | 1.64             | 0.11            | 1.41             | 0.12            |
| Sex      | Male             | 3,497,940        | 0.37             | 0.04            | 1.74             | 0.13            | 1.46             | 0.13            |
|          | Female           | 3,623,686        | 0.33             | 0.03            | 1.55             | 0.10            | 1.37             | 0.13            |
| Age class| \( \leq 9 \)     | 810,373          | 0.26             | 0.05            | 1.46             | 0.21            | 1.34             | 0.26            |
|          | 10-19            | 833,185          | 0.27             | 0.05            | 1.30             | 0.19            | 1.04             | 0.22            |
|          | 20-31            | 1,115,804        | 0.93             | 0.09            | 3.50             | 0.34            | 2.84             | 0.36            |
|          | 32-44            | 1,544,721        | 0.33             | 0.05            | 1.65             | 0.16            | 1.43             | 0.19            |
|          | 45-59            | 1,431,771        | 0.22             | 0.04            | 1.18             | 0.14            | 1.04             | 0.14            |
|          | 60-79            | 1,146,709        | 0.10             | 0.03            | 0.91             | 0.13            | 0.82             | 0.12            |
|          | \( \geq 80 \)    | 239,063          | 0.11             | 0.06            | 1.20             | 0.31            | 1.03             | 0.27            |
| Settlement permit | Swiss            | 5,674,266        | 0.33             | 0.03            | 1.28             | 0.09            | 0.98             | 0.10            |
|          | Foreigner, permanent | 1,020,242 | 0.33             | 0.06            | 1.85             | 0.29            | 2.89             | 0.32            |
|          | Foreigner, temporary | 427,118 | 0.56             | 0.11            | 8.03             | 0.85            | 3.48             | 0.39            |
| Marital status | Single | 2,975,643 | 0.50             | 0.05            | 2.07             | 0.18            | 1.72             | 0.19            |
|          | Married          | 3,377,223        | 0.23             | 0.04            | 1.27             | 0.11            | 1.25             | 0.12            |
|          | Widowed          | 369,339          | 0.25             | 0.08            | 1.23             | 0.26            | 0.79             | 0.13            |
|          | Divorced         | 399,421          | 0.24             | 0.08            | 1.95             | 0.35            | 1.02             | 0.10            |
| Commune language | German + Romansh | 5,128,353 | 0.33             | 0.04            | 1.50             | 0.11            | 1.28             | 0.12            |
|          | French           | 1,680,062        | 0.35             | 0.06            | 1.89             | 0.25            | 1.79             | 0.27            |
|          | Italian          | 313,211          | 0.53             | 0.12            | 2.35             | 0.49            | 1.56             | 0.19            |
| NUTS region | Région lémanique | 1,296,464 | 0.37             | 0.07            | 2.19             | 0.38            | 1.84             | 0.28            |
|          | Espace Mittelland | 1,640,489 | 0.35             | 0.09            | 1.39             | 0.15            | 1.25             | 0.10            |
|          | Nordwestschweiz  | 976,699          | 0.18             | 0.04            | 1.50             | 0.27            | 1.32             | 0.12            |
|          | Zurich           | 1,221,014        | 0.31             | 0.05            | 1.58             | 0.19            | 1.46             | 0.13            |
|          | Ostschweiz       | 1,020,897        | 0.40             | 0.07            | 1.29             | 0.23            | 1.24             | 0.12            |
|          | Zentralschweiz   | 665,904          | 0.36             | 0.06            | 1.57             | 0.25            | 1.19             | 0.12            |
|          | Ticino           | 300,159          | 0.54             | 0.12            | 2.38             | 0.52            | 1.57             | 0.19            |
| Commune size | Small           | 1,372,958        | 0.34             | 0.05            | 1.50             | 0.15            | 1.12             | 0.14            |
|          | Medium           | 2,398,256        | 0.41             | 0.07            | 1.32             | 0.16            | 1.07             | 0.19            |
|          | Large            | 3,350,412        | 0.31             | 0.03            | 2.01             | 0.19            | 1.77             | 0.19            |
| Type     | City/town        | 2,078,780        | 0.35             | 0.04            | 1.96             | 0.17            | 1.82             | 0.20            |
|          | Agglomeration    | 3,145,541        | 0.36             | 0.06            | 1.49             | 0.19            | 1.34             | 0.12            |
|          | Rural            | 1,897,305        | 0.32             | 0.04            | 1.56             | 0.17            | 1.07             | 0.12            |
| Collection method | TRADITIONAL | 265,607          | 0.39             | 0.05            | 1.91             | 0.28            | 1.07             | 0.12            |
|          | SEMI-TRADITIONAL | 174,501 | 0.37             | 0.08            | 1.07             | 0.24            | 1.16             | 0.13            |
|          | TRANSIT + FUTURE | 6,381,359 | 0.33             | 0.03            | 1.62             | 0.11            | 1.42             | 0.12            |
|          | TICINO           | 300,159          | 0.54             | 0.12            | 2.38             | 0.52            | 1.57             | 0.19            |
### Table 3  Comparison of values collected in the survey and the census for the sex variable

<table>
<thead>
<tr>
<th>Sex</th>
<th>Survey</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census</td>
<td>Total</td>
<td>393</td>
<td>383</td>
<td>776</td>
</tr>
<tr>
<td>Matched</td>
<td>Total</td>
<td>24,171</td>
<td>24,936</td>
<td>49,107</td>
</tr>
<tr>
<td>Matched</td>
<td>Male</td>
<td>23,967</td>
<td>166</td>
<td>24,133</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>204</td>
<td>24,770</td>
<td>24,974</td>
</tr>
<tr>
<td>Matched</td>
<td>(imputed value)</td>
<td>6</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Matched</td>
<td>(imputed value)</td>
<td>0</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>24,564</td>
<td>25,319</td>
<td>49,883</td>
</tr>
</tbody>
</table>

### Table 4  Comparison of values collected in the survey and the census for the labour force status variable

<table>
<thead>
<tr>
<th>Labour force status</th>
<th>Survey</th>
<th>Employed</th>
<th>Unemployed</th>
<th>Not in labour force</th>
<th>≤ 15 years of age</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census</td>
<td>Not matched</td>
<td>424</td>
<td>23</td>
<td>217</td>
<td>112</td>
<td>776</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>25,163</td>
<td>498</td>
<td>14,501</td>
<td>8,945</td>
<td>49,107</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>Employed</td>
<td>23,953</td>
<td>188</td>
<td>2,007</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Unemployed</td>
<td>300</td>
<td>221</td>
<td>323</td>
<td>1</td>
<td>845</td>
</tr>
<tr>
<td></td>
<td>Not in labour force</td>
<td>901</td>
<td>89</td>
<td>12,143</td>
<td>18</td>
<td>13,151</td>
</tr>
<tr>
<td>≤ 15 years of age</td>
<td>9</td>
<td>0</td>
<td>28</td>
<td>8,913</td>
<td>8,950</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matched (imputed variable)</td>
<td>564</td>
<td>22</td>
<td>312</td>
<td>6</td>
<td>904</td>
</tr>
<tr>
<td></td>
<td>Employed</td>
<td>14</td>
<td>8</td>
<td>26</td>
<td>1</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Unemployed</td>
<td>92</td>
<td>15</td>
<td>881</td>
<td>5</td>
<td>993</td>
</tr>
<tr>
<td>≤ 15 years of age</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>25,587</td>
<td>521</td>
<td>14,718</td>
<td>9,057</td>
<td>49,883</td>
</tr>
</tbody>
</table>

### 6.3 Potential location and time lag errors

Of the 49,107 elements matched between the coverage survey and the census, 97.7% are found within the basic area around the reference address collected in the survey. The corresponding value is 98.1% for persons who did not indicate any move between Census Day and the day of the survey. It is 83.9% for those who indicated a move (1,512 persons); see absolute numbers in Table 5.

It is worth noting that 9.4% of the persons in NORD who did not move are found close to their reference address but not in exactly the same building. While these problems of exact location have a negligible effect on the census data, they show the difficulty of identifying the buildings sampled when constructing lists of households in the field during the survey, as well as the difficulty of assigning persons to buildings during the processing of the census data. However, a supplementary survey would be needed to evaluate the respective effects of these two difficulties.

Efforts to locate persons who moved indicate that 151 = 145 + 6 persons were located near their address reported on the day of the survey and not near their Census Day address (9%, weighted). Also, a set of 688 persons in NORD, among the 922 located in the two basic areas, were actually found to be residing in the building on the day of the survey. During the coverage survey, special care was taken regarding questions on addresses on Census Day and on the day of the survey. We therefore believe that the addresses of persons who moved are of better quality in the survey data than in the census data. On this basis, we deduce that out of the 1,512 persons who moved, at least 151 + 688 = 839 are enumerated in the census at an address that they did not have on the official day of data collection but at an address that they had some time after that date. The exact time lag is not known, since the moving date was not collected in the survey.
Table 5
Comparison of the location of matched persons. The areas are defined for the address on Census Day (according to information collected in the survey) and for the address on the day of the survey (also according to information collected in the survey). Presence in the basic area, the extended area (outside the basic area) or outside the extended area for persons who did not move (stayed) and persons who moved (moved) between the census and the survey.

<table>
<thead>
<tr>
<th>Day of survey</th>
<th>Stayed</th>
<th>Moved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basic area</td>
<td>Extended area</td>
</tr>
<tr>
<td>Census</td>
<td>46,689</td>
<td>922</td>
</tr>
<tr>
<td>Day</td>
<td></td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>258</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>648</td>
<td>15</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>1,124</td>
</tr>
<tr>
<td>Total</td>
<td>47,595</td>
<td></td>
</tr>
</tbody>
</table>

7. Conclusion

Overall coverage deficiencies in the 2000 census of population in Switzerland are of the same order of magnitude as those for the censuses of other countries. However, differences are noted for subgroups (e.g., regions). Of the three components, undercoverage is of great interest, since it not only serves to detect groups of persons not well enumerated, but it also lends itself to analysing location and measurement errors. As to overcoverage estimates, these are limited by the lack of information supplementary to the census for $S_e$. In the future, they could be improved by collecting supplementary information on characteristics reported on Census Day in a survey of persons in that sample (e.g., location and household type). Net undercoverage estimates are based on several assumptions. The results in large domains seem reliable, but certain risks, notably related to the choice of estimation cells, exist when domains are smaller. For future estimates, we propose to evaluate the model approach applied in the United Kingdom instead of the estimation cells traditionally used in the United States.

An important element to review for future estimates is the choice of the population of interest. The decision to limit that population to persons in private households and the economic residence led to a few problems in the estimates, since it was difficult to delimit that population precisely. In a future estimation, collective households could be excluded so as to avoid practical problems relating to collection but retain all types of residences. The set of records for the economic residence would then be treated as a domain.

Estimating the coverage deficiencies of a census is an ambitious project that has proved to be worthwhile. The results provide information on the quality of the data from the 2000 census and the different coverage problems. Upcoming censuses will essentially be based on registers. Coverage estimates will be based on the experience acquired in making the 2000 estimates, with probable adaptations to take account of the new data collection system.

Acknowledgements

I wish to thank Philippe Eichenberger of the statistical methods unit of the Office fédéral de la statistique for the productive discussions throughout the project. Warm thanks are also extended to Dr. Rajendra Singh and his colleagues in the Decennial Statistical Studies Division of the U.S. Census Bureau for their assistance in developing methods and estimates. I also wish to thank all census personnel who performed tasks and provided information needed to carry out the project, and to Paul-André Salamin of the statistical methods unit for his careful rereading of the article.

References


Use of a web-based convenience sample to supplement a probability sample

Marc N. Elliott and Amelia Haviland

Abstract

In this paper we describe a methodology for combining a convenience sample with a probability sample in order to produce an estimator with a smaller mean squared error (MSE) than estimators based on only the probability sample. We then explore the properties of the resulting composite estimator, a linear combination of the convenience and probability sample estimators with weights that are a function of bias. We discuss the estimator’s properties in the context of web-based convenience sampling. Our analysis demonstrates that the use of a convenience sample to supplement a probability sample for improvements in the MSE of estimation may be practical only under limited circumstances. First, the remaining bias of the estimator based on the convenience sample must be quite small, equivalent to no more than 0.1 of the outcome’s population standard deviation. For a dichotomous outcome, this implies a bias of no more than five percentage points at 50 percent prevalence and no more than three percentage points at 10 percent prevalence. Second, the probability sample should contain at least 1,000-10,000 observations for adequate estimation of the bias of the convenience sample estimator. Third, it must be inexpensive and feasible to collect at least thousands (and probably tens of thousands) of web-based convenience observations. The conclusions about the limited usefulness of convenience samples with estimator bias of more than 0.1 standard deviations also apply to direct use of estimators based on that sample.

Key Words: Bias; Composite estimator; Calibration.

1. Introduction

Web-based surveys have steadily increased in use and take a variety of forms (Couper 2000). For instance, web-based probability samples use a traditional sampling frame and provide web-mode as one response option or the only response option. Web-based probability samples can have high response rates and produce estimators with minimal non-response bias (Kypri, Stephenson and Langley 2004). In contrast, web-based convenience samples are based on “inbound” hits to web pages obtained from anyone online who finds the site and chooses to participate (sometimes as a result of advertising to a population that is not specifiable) or based on volunteerism from recruited panels that are not necessarily representative of the intended population.

The primary appeal of web-based convenience samples lies in the potentially very low marginal cost per case. Visits to a web site do not require expensive labor (as for phone calls) or materials (as for mailings) for each case, combined with rapid data collection and reductions in marginal data processing costs per case. Even with some fixed costs, the total costs per case are potentially very low, especially for large surveys. The disadvantage of these samples is also clear: potentially large and unmeasured selection bias.

Most discussions of web-based convenience samples of which we are aware have either argued that probability samples are unimportant in general, tried to delineate the circumstances under which convenience samples may be useful, or dismissed the use of convenience samples entirely. We explore a different avenue by investigating the possibility of integrating web-based convenience samples into the context of probability sampling.

In this paper we describe a methodology for combining a convenience sample with a probability sample to produce an estimator with a smaller mean squared error (MSE) than estimators that employ only the probability sample. We then explore the properties of the resulting composite estimator, a linear combination of the convenience and probability samples with weights determined by bias. This leads to recommendations regarding the usefulness of supplementing probability samples with web-based convenience samples. Because the marginal costs of web-based convenience samples are very low, we focus on identifying situations in which the increase in effective sample size (ESS) attributable to the inclusion of the convenience sample may be sufficient to justify a dual-mode approach. We demonstrate that there are limited circumstances under which a supplemental web-based convenience sample may meaningfully improve MSE. While we focus on web-based convenience samples, the discussion that follows applies to other low-cost data collection methods with poor population coverage.

2. Problem context

2.1 Initial conditions

For the combined probability/convenience sample, we propose that the same survey be administered simultaneously to a traditional probability sample (with or
without a web-based response mode) and a web-based convenience sample. We envision a multi-purpose survey with a number of survey outcomes. In this paper we will focus on the estimation of means, but future work might extend these results to other parameters, such as regression parameters. Although we will initially consider cases where the bias of convenience sample estimates is known, we will later consider the extent to which the probability sample provides a means of measuring the unknown bias in each parameter estimate from the convenience sample.

With known bias, one may combine the convenience and probability samples in a manner that minimizes MSE. If estimates from the convenience sample are very biased, the convenience sample will accomplish little. This possibility requires that the probability sample be large enough to stand on its own. Thus, one approach would be to set aside a small portion of the probability sample budget to create a large convenience sample supplement.

For example, consider a survey for which the primary interest is in estimates for the population as a whole, but for which subpopulations estimates would also be desirable if a sample size supporting adequate precision were affordable. Suppose further that one could draw 4,000 probability observations and 10,000 convenience observations for the cost of the probability sample of 5,000. For a given outcome, if bias is large, standard errors increase moderately through a small proportionate loss in sample size; if bias is small overall and within each subpopulation, there might be a "precision windfall," allowing acceptably precise subpopulation analyses.

2.2 Initial bias reduction

We will demonstrate that the bias of convenience sample estimators must be quite small for the sample to be useful, suggesting that it may be best to focus on estimating parameters that are typically subject to less bias than overall unadjusted population estimates of proportions or means, such as regression coefficients (Kish 1985).

Additionally, one might reduce bias by calibrating the convenience sample to known population values (Kalton and Kasprzyk 1986) or by applying propensity score weights that model membership selection between the two samples to observations from the convenience sample (Rosenbaum 2002; Rosenbaum and Rubin 1983). A small set of items can be included to allow the use of either approach. These items might include both items that predict differences between respondents to web surveys and other survey modes, as well as items tailored to the content of the particular survey. The design effect from the resulting variable weights will reduce the ESS for convenience sample estimators, but the low costs of these observations makes compensating for moderate design effects affordable.

We then can estimate the remaining bias for a given parameter as the difference between the estimate in the probability sample and the weighted estimate in the convenience sample.

3. Efficiency considerations

3.1 Linear combinations of biased and unbiased estimators of a population mean

The most efficient estimator that is a linear combination of the (weighted) convenience and probability samples is a special case of an estimator given in a result by Rao (2003, pages 57-58). The properties of this estimator lead to general recommendations regarding the conditions of probability sample size, convenience sample size, and convenience sample estimator bias under which the convenience sample meaningfully improves the ESS of the probability sample.

We begin by asking: What is the most efficient estimator of this form when the magnitude of the bias is known? We will later consider relaxing the assumption of known magnitude of the bias.

Let \( n_1 \) and \( n_2 \) be the effective sample sizes of the probability sample and convenience samples, respectively, after dividing nominal sample sizes by design effects associated with the sample design and non-response adjustments. This includes propensity score or other weighting in the case of the convenience sample. The former population has mean \( \mu \), and variance \( \sigma_1^2 \); the latter has mean \( \mu + \varepsilon \) and variance \( \sigma_2^2 \), where \( \varepsilon \) is the known bias remaining after weighting and \( \mu \) is the unknown parameter of interest. The corresponding sample means have expectation \( \mu \) and \( \mu + \varepsilon \) and variance \( \sigma_i^2 / n_i \) for \( i = 1, 2 \) under an infinite population sampling model. We assume these two estimators are uncorrelated, as they come from independent samples.

From Rao (2003, pages 57-58), the most efficient composite estimator of \( \mu \) takes the form

\[
\hat{\mu} = \frac{\bar{x}_1 (\sigma_1^2 / n_1) + \bar{x}_2 (\varepsilon^2 + \sigma_2^2 / n_2)}{\varepsilon^2 + \sigma_1^2 / n_1 + \sigma_2^2 / n_2},
\]

with remaining bias

\[
\varepsilon_c = \varepsilon \left( \frac{\sigma_i^2 / n_i}{\varepsilon^2 + \sigma_i^2 / n_i + \sigma_j^2 / n_j} \right)
\]

and

\[
\text{MSE}_c = \frac{(\sigma_i^2 / n_i) (\varepsilon^2 + \sigma_j^2 / n_j)}{\varepsilon^2 + \sigma_i^2 / n_i + \sigma_j^2 / n_j}.
\]
As can be seen, the composite estimator is a convex combination of the convenience sample and probability sample means. The influence of the former is determined by the ratio of the MSE (here variance) of the probability sample mean to the sum of that term and the MSE of the convenience sample mean. Similarly, the remaining bias is the original bias multiplied by this same ratio, whereas the resultant $MSE_c$ is the product of the two MSEs divided by their sum. Note that bias approaches zero both as $E \rightarrow 0$ (no selection bias in the convenience sample estimate) and as $E \rightarrow \infty$ (no weight given to the convenience sample).

### 3.2 Quantifying the contributions of the convenience sample

We now can evaluate the contributions of the convenience sample based on the known remaining bias in its associated estimators. To this end, we will define several quantities.

Let

$$ESS_i = \frac{\sigma_i^2}{n_i} = \left( \frac{\varepsilon_i^2 + \sigma_i^2}{\varepsilon_i^2 + \sigma_i^2/n_i} \right) n_i$$

be the effective sample size needed for an unbiased sample mean with the same MSE as the composite estimator. To further simplify this expression, let us define the remaining standardized bias, $E = \varepsilon / \sigma$, and consider the case in which the observations from the convenience and probability populations have equal variance, $(\sigma_1 = \sigma_2 = \sigma)$. In this case, the increment to $ESS_i$ attributable to the convenience sample, the difference between $ESS_i$ with and without the convenience sample, is

$$\frac{1}{n_2 + E^2} = n_2 \left( \frac{1}{1 + n_2 E^2} \right).$$

### 3.3 Maximum contribution of the convenience sample

As $n_2 \rightarrow \infty$, the increment to $ESS_i$ approaches $1/E^2$. This limit, the inverse of the squared standardized bias, is the maximum possible incremental contribution of the convenience sample to the $ESS_i$ (abbreviated MICCS). If the MICCS is small, then a convenience sample of any size cannot meaningfully improve MSE. If the MICCS is large enough to be meaningful, we then need to consider what convenience sample sizes are needed to achieve a large proportion of the MICCS.

To develop intuition for the magnitude of $E$ (standardized bias) we consider the important case of a dichotomous outcome, for which $E = \varepsilon/\sqrt{P(1-P)}$ where $P$ is the population probability of the outcome. Table 1 below translates bias for a dichotomous outcome from percentage points to standardized bias and then to the corresponding MICCS for $P = 0.1$ and $P = 0.5$.

<table>
<thead>
<tr>
<th>$E$ (Standardized Bias)</th>
<th>Overall Prevalence of Outcome</th>
<th>MICCSa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>0.01</td>
<td>0.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td>0.02</td>
<td>0.6%</td>
<td>1.0%</td>
</tr>
<tr>
<td>0.05</td>
<td>1.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>0.10</td>
<td>3.0%</td>
<td>5.0%</td>
</tr>
<tr>
<td>0.20</td>
<td>6.0%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

*a Of estimators of means using only the convenience sample

a ESS added with an infinitely large convenience sample relative to no use of a convenience sample.

For a proportion near 50%, a bias of 2.5 percentage points limits the potential increment of $ESS_i$ to 400. The minimum increment to $ESS_i$ that offsets the fixed cost of setting up the web-based response mode will vary by user, but we suspect increments of less than 100 will rarely be cost-effective. Table 1 then implies that convenience samples for which the standardized biases of estimators restricted to the convenience sample generally exceed 0.1 standard deviations will rarely prove cost-effective. For a dichotomous variable with $P$ between 0.1 and 0.5 this corresponds to a bias of 3 to 5 percentage points.

How easily are biases of this magnitude achieved with adjusted estimates from convenience samples? Several studies compared propensity-weighted web-based convenience samples to RDD surveys. One (Taylor 2000) advocated the stand-alone use of such convenience samples despite differences of as much as five percentage points in a number of estimates for dichotomous outcomes regarding political attitudes, with standardized bias of 0.05 to 0.10 if one treats RDD as a gold standard. Another (Schonlau, Zapert, Simon, Sanstad, Marcus, Adams, Spranca, Kan, Turner and Berry 2003) does not report magnitudes of differences, but does report that 29 of 37 items regarding health concerns exhibit differences that are statistically significant at $p < 0.01$. Given the reported sample sizes (and optimistically ignoring any DEFF from weighting), it can be shown that significance at that threshold implies point estimates of standardized bias exceeding 0.05 for estimators of 78% of items. The key outcome in a Slovenian comparison of a probability phone sample and a Web-based convenience sample (Vehovar, Manfreda and Batagelj 1999) would be estimated with a standardized bias of more than 0.1 from the convenience sample even after extensive weighting adjustments. It should be noted that there may also be mode effects on responses for the Web mode when compared to a telephone mode among subjects randomized to response mode (Fricker, Galesic, Tourangeau and Yan 2005), so that not all differences between Web convenience samples and non-Web probability samples may result from selection.
3.4 Actual contribution of the convenience sample

While the maximum possible increment (MICCS) is $1/E^2$, the actual increment to ESS can be expressed as $(k/k+1)$ MICCS where $k=n_1E^2$. The shortfall of the actual increment to ESS from the MICCS can then be expressed as MICCS-ESS = $1/[(E^2)(1+n_1E^2)]$. This implies that the returns to ESS diminish with increasing size of the convenience sample, more quickly with large bias since the bias eventually dominates any further variance reduction. Half of the MICCS noted is achieved when the ESS of the convenience sample is equal to the MICCS. For example, if bias is 0.01 standard deviations and a convenience sample has an ESS of 10,000, then the MICCS is 10,000, but the actual incremental contribution to ESS will be 5,000. This suggests that convenience samples with ESS 2-20 times as large as MICCS will suffice for most purposes, which correspond to 67%-95% of the potential gain in ESS. Such heuristics in turn imply collecting 200 - 4,000 such cases when $E$ is relatively large ($E = 0.05$ to 0.10) and 5,000 - 200,000 such cases when $E$ is relatively small ($E = 0.01$ to 0.02). Table 2 provides illustrative examples of the ESS achieved at several combinations of sample sizes and bias.

Table 2 Examples of ESS1 at several sample sizes and levels of standardized bias

<table>
<thead>
<tr>
<th>$n_1$ (Probability Sample Size)</th>
<th>$n_2$ (Convenience Sample Size)</th>
<th>$E$ (Standardized Bias)</th>
<th>ESS1 for the Composite Estimate</th>
<th>ESS1 / $n_1^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>1,000</td>
<td>0.01</td>
<td>1,909</td>
<td>1.909</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000</td>
<td>0.10</td>
<td>1,091</td>
<td>1.091</td>
</tr>
<tr>
<td>1,000</td>
<td>10,000</td>
<td>0.01</td>
<td>6,000</td>
<td>6.000</td>
</tr>
<tr>
<td>1,000</td>
<td>10,000</td>
<td>0.10</td>
<td>1,099</td>
<td>1.099</td>
</tr>
<tr>
<td>1,000</td>
<td>100,000</td>
<td>0.01</td>
<td>10,091</td>
<td>10.091</td>
</tr>
<tr>
<td>1,000</td>
<td>100,000</td>
<td>0.10</td>
<td>1,100</td>
<td>1.100</td>
</tr>
<tr>
<td>10,000</td>
<td>1,000</td>
<td>0.01</td>
<td>10,909</td>
<td>1.091</td>
</tr>
<tr>
<td>10,000</td>
<td>1,000</td>
<td>0.10</td>
<td>10,091</td>
<td>1.091</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
<td>0.01</td>
<td>15,000</td>
<td>1.500</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
<td>0.10</td>
<td>10,099</td>
<td>1.010</td>
</tr>
<tr>
<td>10,000</td>
<td>100,000</td>
<td>0.01</td>
<td>19,091</td>
<td>1.909</td>
</tr>
<tr>
<td>10,000</td>
<td>100,000</td>
<td>0.10</td>
<td>10,100</td>
<td>1.010</td>
</tr>
</tbody>
</table>

Number of estimators of means using only the convenience sample

Of estimators of means using only the convenience sample

ESS relative to no use of a convenience sample.

3.5 Precision for estimating bias

Heretofore, we have assumed a known bias in convenience sample estimators; in practice, the bias will need to be estimated using information from both samples. We next explore the extent to which the size of the probability sample also constrains the usefulness of the convenience sample through the need to precisely estimate the remaining bias.

We can estimate $\epsilon$ as the difference between the sample mean of the probability sample and the weighted mean of the convenience sample. The true standard error for the estimate of bias is $\sigma_\epsilon = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. If $\sigma_1 = \sigma_2 = \sigma$, the true standard error for the estimate of standardized bias (E) is $\sigma_\epsilon = \sqrt{1/n_1 + 1/n_2}$. No matter how large the convenience sample, this term can never be less than the inverse of the square root of the probability sample size.

It has been demonstrated that the relative error in MSE for a composite estimator is relatively insensitive to small errors in the estimates of bias (Schaible 1978), which is encouraging for well-estimated biases. Unfortunately, unless both the probability and convenience ESS are large, the standard error of the estimate for $E$ is impractically large relative to the values of $E$ that make the convenience supplement useful ($E < 0.10$). For example, suppose that a probability sample of ESS 1,000 and a convenience sample of ESS 5,000 yielded a point estimate of standardized bias of 0.02. If the point estimate were correct, the convenience sample would increase the ESS by 1,667. But this estimate could also have a true bias of 0.088 standard deviations (95% upper confidence limit), which would imply that the increment would be less than 130.

If we assume that the convenience sample size will always be at least twice the probability sample size, these results imply that practical applications of this technique must have a minimum sample size of 1,000-10,000 for the probability sample if they are to address the uncertainty in the magnitude of bias in convenience sample estimators (standard errors of $E$ in the 0.01 to 0.04 range).

4. Discussion

We describe a composite estimator that is a linear combination of unbiased sample mean estimates from a probability sample and a biased (propensity-score weighted) sample mean estimate from a web-based convenience sample. We use the MSE of this composite estimator to characterize the contributions of the convenience sample to an estimator based only on the probability sample in terms of ESS. We then calculate the maximal contribution of the convenience sample, the role of the convenience sample size in approaching this limit, and the roles of both sample sizes in estimating bias with sufficient precision.

Practitioners sometimes assume that small probability samples are sufficient to estimate the bias in estimates from corresponding convenience samples. Our results suggest otherwise. We demonstrate that the standardized bias of web-based convenience sample estimators after initial adjustments to reduce bias must be quite small (no more than 0.1 standard deviations, and probably less than 0.05 standard deviations) for the MSE of the overall estimate to be meaningfully smaller than it would be without use of the convenience sample. We further demonstrate that convenience sample sizes of thousands or tens of thousands...
are also needed to realize practical gains. Finally, we demonstrate that a large probability sample size (1,000-10,000) is also needed for reasonably precise estimates of the remaining bias in initially bias-adjusted convenience sample estimators. Because the bias of estimates in an application to a multipurpose survey is likely to vary by outcome, the global decision to substitute a large number of inexpensive surveys for fewer traditional surveys must be made carefully.

The greatest opportunity in cost savings may be in large surveys, simply as a function of their size. On the other hand, the greatest proportionate gains in precision are likely to occur for samples of intermediate size. Gains might also be substantial for large samples in which the main inferences are smaller subgroups. For example, a national survey of 100,000 individuals might make inference to 200 geographic subregions, with samples of 500 for each. If one supplemented this national sample with a very large web-based convenience sample, estimated the bias nationally, and elected to assume that the bias did not vary regionally, one might decrease the MSE of the sub-region estimates substantially through the use of such a composite estimator.

As a final caveat, the conclusions about the limited usefulness of convenience samples with estimator bias of more than 0.1 standard deviations are not limited to attempts to use a composite estimator. The same approach can be applied to show that an estimator based only on a convenience sample of any size with a standardized bias of 0.2 (e.g., ten percentage points for a dichotomous variable with $P = 0.5$) will have an MSE greater than or equal to that of an estimate from a probability sample of size 25.

**Acknowledgements**

The authors would like to thank John Adams, Ph.D., for his comments on the manuscript, Matthias Schonlau, Ph.D., for his comments on an earlier version of the manuscript, and Colleen Carey, B.A., Kate Sommers-Dawes, B.A., and Michelle Platt, B.A. for their assistance in the preparation of the manuscript.

Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056); the views expressed in this article do not necessarily reflect the views of the Centers for Disease Control and Prevention.

**References**


ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2007.

R.C. Bailey, George Mason University
M. Bankier, Statistics Canada
J.-F. Beaumont, Statistics Canada
Y. Bélanger, Statistics Canada
W. Bell, U.S. Bureau of the Census
Y. Berger, University of Southampton
C. Boudreau, Medical College of Wisconsin
M. Brick, Westat, Inc.
J. Chen, University of Waterloo, Canada
R. Clark, University of Wollongong, Australia
E. Dagum, University of Bologna
A. Dessertaine, EDF, R&D-OSIRIS-CLAMART
F. Dupont, INSEE
M. Elliott, University of Michigan
L.R. Ernst, Bureau of Labour Statistics, U.S.A.
L. Fattorini, Università di Siena
R. Fisher, U.S. Department of Treasury
W. Fuller, Iowa State University
J. Gambino, Statistics Canada
D. Garriguet, Statistics Canada
C. Girard, Statistics Canada
E. Gois, Instituto Nacional de Estatistica, Portugal
B. Graubard, National Cancer Institute
R. Griffin, US Bureau of the Census
D. Haziza, Statistics Canada
S. Heeringa, ISR, Michigan
M. Hidiroglou, Statistics Canada
J. Jiang, University of California at Davis
W. Kalsbeek, U. of North Carolina
J.-K. Kim, Department of Applied Statistics, Seoul, Korea
P. Knottnerus, Statistics Netherlands
E. Korn, National Institute of Health
P. Kott, National Agricultural Statistics Service
D. Ladiray, INSEE
P. Lahiri, JPSM, University of Maryland
N. Laniel, Statistics Canada
M. Latouche, Statistics Canada
P. Lavallée, Statistics Canada
S. Lee, University of California at Los Angeles
S. Lele, University of Alberta
C. Léon, Statistics Canada
P.S. Lévy, RTI International
S. Lohr, Arizona State University
W.W. Lu, Acadia University
L. Mach, Statistics Canada
T. Maiti, Iowa State University
D. Malec, United States Bureau of the Census
H. Mantel, Statistics Canada
A. Matei, Université de Neuchâtel, Switzerland
S. Merad, Office for National Statistics, UK
J. Montequilla, Westat, Inc.
R. Munnich, University of Tubingen
P.L. Nascimento Silva, University of Southampton
G. Nathan, Hebrew University of Jerusalem
J. Opsomer, Iowa State University
S.M. Paddock, Rand corporation, U.S.A.
S. Pramanik, JPSM, University of Maryland
N. Prasad, University of Alberta
L. Qualité, Université de Neuchâtel, Switzerland
T.J. Rao, ISI
J.N.K. Rao, Carleton University
L.-P. Rivest, Université Laval
S. Rubin-Bleuer, Statistics Canada
M. del Mar Rueda, University of Granada
P. Saavedra, ORC Macro, U.S.A.
F. Scheuren, NORC
E. Schindler, US Bureau of the Census
K. Sinha, University of Illinois at Chicago, U.S.A.
C.J. Skinner, University of Southampton
P.J. Smith, Center for Disease Control and Prevention
M. Thompson, University of Waterloo
Y. Tillé, Université de Neuchâtel
F. Verret, Statistics Canada
S.G. Walker, University of Kent, U.K.
K. Wolter, NORC
C. Wu, University of Waterloo
W. Yung, Statistics Canada

Acknowledgements are also due to those who assisted during the production of the 2007 issues: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Michel Pelchat and Isabelle Poliquin (Dissemination Division), Sheri Buck (Systems Development Division), François Beaudin (Official Languages and Translation Division) and Sophie Chartier and Gayle Keely (Household Survey Methods Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.
JOURNAL OF OFFICIAL STATISTICS
An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 23, No. 1, 2007

Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965
Margo Anderson and William Seltzer ................................................................. 1

Efficient Stratification Based on Nonparametric Regression Methods
Enrico Fabrizi and Carlo Trivisano ....................................................................... 35

A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption
Barry Schouten ..................................................................................................... 51

Imputing for Late Reporting in the U.S. Current Employment Statistics Survey
Kennon R. Copeland and Richard Valliant .......................................................... 69

Incentives in Random Digit Dial Telephone Surveys: A Replication and Extension
Richard Curtin, Eleanor Singer and Stanley Presser ............................................. 91

Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience
Peter Lynn, Sabine Häder, Siegfried Gabler and Seppo Laaksonen ......................... 107

Book and Software Reviews .................................................................................. 125

Contents
Volume 23, No. 2, 2007

Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents
Jaak Billiet, Michel Phillippens, Rory Fitzgerald and Ineke Stoop ................................ 135

Measuring Disability in Surveys: Consistency over Time and Across Respondents
Sunghee Lee, Nancy A. Mathiowetz and Roger Tourangeau .................................. 163

Quantifying Stability and Change in Ethnic Group
Ludi Simpson and Bola Akinwale ........................................................................ 185

Seasonal Adjustment of Weekly Time Series with Application to Unemployment Insurance Claims and Steel Production
William P. Cleveland and Stuart Scott ................................................................ 209

Finite Population Small Area Interval Estimation
Li-Chun Zhang .................................................................................................... 223

Predicting Natural Gas Production in Texas: An Application of Nonparametric Reporting Lag Distribution Estimation
Crystal D. Linkletter and Randy R. Sitter ............................................................ 239

Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update
Laura Zayatz ...................................................................................................... 253

Book and Software Reviews .................................................................................. 267
## Contents

**Volume 23, No. 3, 2007**

The Morris Hansen Lecture 2006 Statistical Perspectives on Spatial Social Science  
Michael F. Goodchild...................................................................................................................... 269

Using Geospatial Information Resources in Sample Surveys  
Sarah M. Nusser ........................................................................................................................... 285

Discussion  
Linda Williams Pickle.................................................................................................................. 291

Optimizing the Use of Microdata: An Overview of the Issues  
Julia Lane ......................................................................................................................................... 299

Benchmarking the Effect of Cell Adjustment on Tabular Outputs: The Shortcomings of Current Approaches  
Paul Williamson ............................................................................................................................ 319

Summary of Accuracy and Coverage Evaluation for the U.S. Census 2000  
Mary H. Mulry ............................................................................................................................... 345

Resampling Variance Estimation in Surveys with Missing Data  
A.C. Davison and S. Sardy ........................................................................................................... 371

Nonresponse Among Ethnic Minorities: A Multivariate Analysis  
Remco Feskens, Joop Hox, Gerty Lensvelt-Mulders and Hans Schmeets ........................................ 387

Procedures for Updating Classification Systems: A Study of Biotechnology and the Standard Occupational Classification System  
Neil Malhotra and Jon A. Krosnick ............................................................................................ 409

---

All inquiries about submissions and subscriptions should be directed to jos@scb.se
Volume 35, No. 1, March/mars 2007

Paul GUSTAFSON
Editorial ....................................................................................................................................................................................1

Report from the previous Editor..............................................................................................................................................................5

Acknowledgement of referees’ services...................................................................................................................................................7

Renato ASSUNÇÃO, Andréa TAVARES, Thais CORREA & Martin KULLDORFF
Space-time cluster identification in point processes .....................................................................................................................9

Elvan CEYHAN, Carey E. PRIEBE & David J. MARCHETTE
A new family of random graphs for testing spatial segregation ..................................................................................................27

Aurélie LABBE & Mary E. THOMPSON
Multiple testing using the posterior probabilities of directional alternatives, with application to genomic studies.........................51

H.a.o.l.a.n. LU, James S. HODGES & Bradley P. CARLIN
Measuring the complexity of generalized linear hierarchical models..........................................................................................69

Aurore DELAIGLE
Nonparametric density estimation from data with a mixture of Berkson and classical errors .......................................................89

Ozlem ILK & Michael J. DANIELS
Marginalized transition random effect models for multivariate longitudinal binary data............................................................105

Richard A. LOCKHART, John J. SPINELLI & Michael A. STEPHENS
Cramér-von Mises statistics for discrete distributions with unknown parameters ........................................................................125

P.W. FONG, W.K. LI, C.W. YAU & C.S. WONG
On a mixture vector autoregressive model ......................................................................................................................................135

Jung Wook PARK, Marc G. GENTON & Sujit K. GHOSH
Censored time series analysis with autoregressive moving average models ................................................................................151

Abdessamad SAIDI
Consistent testing for non-correlation of two cointegrated ARMA time series ...........................................................................169

Complete online access to The Canadian Journal of Statistics .....................................................................................................189

Volume 35, No. 2, June/juin 2007

Pierre DUCHESNE
On consistent testing for serial correlation in seasonal time series models ........................................................................................193

Jaakko NEVALAINEN, Denis LAROCQUE & Hannu OJA
On the multivariate spatial median for clustered data ..................................................................................................................215

Liqun WANG
A unified approach to estimation of nonlinear mixed effects and Berkson measurement error models ...........................................233

Juan Carlos PARDO-FERNÁNDEZ, Ingrid VAN KEILEGOM & Wenceslao GONZÁLEZ-MANTEIGA
Goodness-of-fit tests for parametric models in censored regression ..............................................................................................249

Song Xi CHEN & Tzee-Ming HUANG
Nonparametric estimation of copula functions for dependence modelling ...........................................................................................265

Victor DE OLIVEIRA
Objective Bayesian analysis of spatial data with measurement error ............................................................................................283

Gonzalo GARCÍA-DONATO & Dongchu SUN
Objective priors for hypothesis testing in one-way random effects models ................................................................................303

Hui LI, Guosheng YIN & Yong ZHOU
Local likelihood with time-varying additive hazards model .........................................................................................................321

Forthcoming papers/Articles à paraître ...............................................................................................................................................338

Complete online access to The Canadian Journal of Statistics .....................................................................................................339
David F. ANDREWS
Robust likelihood inference for public policy ................................................................. 341

Zeny Z. FENG, Jiahua CHEN, Mary E. THOMPSON
Asymptotic properties of likelihood ratio test statistics in affected-sib-pair analysis ........... 351

Hwashin H. SHIN, Glen TAKAHARA & Duncan J. MURDOCH
Optimal designs for calibration of orientations ............................................................... 365

Jiancheng JIANG, Haibo ZHOU, Xuejun JIANG & Jianan PENG
Generalized likelihood ratio tests for the structure of semiparametric additive models .......... 381

Runze LI & Lei NIE
A new estimation procedure for a partially nonlinear model via a mixed-effects approach ......... 399

Axel MUNK, Matthias MIELKE, Guido SKIPKA & Gudrun FREITAG
Testing noninferiority in three-armed clinical trials based on likelihood ratio statistics .......... 413

Inkyung JUNG & Martin KULLDORFF
Theoretical properties of tests for spatial clustering of count data .................................. 433

QiQi LU & Robert B. LUND
Simple linear regression with multiple level shifts ....................................................... 447

Forthcoming papers/Articles à paraître ............................................................................ 459

Volume 36 (2008): Subscription rates/Frais d’abonnement ............................................ 460