# Survey Methodology 

## December 2008



## 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 about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

## Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

| Inquiries line | $1-800-263-1136$ |
| :--- | :--- |
| National telecommunications device for the hearing impaired | $1-800-363-7629$ |
| Fax line | $1-877-287-4369$ |
| Local or international calls: |  |
| Inquiries line | $1-613-951-8116$ |
| Fax line | $1-613-951-0581$ |
|  |  |
| Depository Services Program | $1-800-635-7943$ |
| Inquiries line | $1-800-565-7757$ |

## To access and order this product

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

This product, Catalogue no. 12-001-X, 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:

|  | Single issue | Annual subscription |
| :--- | ---: | ---: |
| United States | CAN $\$ 6.00$ | CAN $\$ 12.00$ |
| Other countries | CAN $\$ 10.00$ | CAN $\$ 20.00$ |

All prices exclude sales taxes.
The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail
infostats@statcan.gc.ca
- Mail

Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor

150 Tunney's Pasture Driveway
Ottawa, Ontario K1A OT6

- In person from authorized 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, Statistics Canada has developed standards of service that its employees observe. 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.gc.ca under "About us" > "Providing services to Canadians."

## Statistics Canada

# Survey <br> Methodology 

## December 2008

Published by authority of the Minister responsible for Statistics Canada
© Minister of Industry, 2008
All rights reserved. The content of this electronic publication may be reproduced, in whole or in part, and by any means, without further permission from Statistics Canada, subject to the following conditions: that it be done solely for the purposes of private study, research, criticism, review or newspaper summary, and/or for non-commercial purposes; and that Statistics Canada be fully acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form, by any means-electronic, mechanical or photocopy-or for any purposes without prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A OT6.

December 2008
Catalogue no. 12-001-XIE
ISSN 1492-0921
Catalogue no. 12-001-XPB
ISSN: 0714-0045
Frequency: semi-annual
Ottawa

Cette publication est disponible en français sur demande ( $n^{\circ} 12-001-X$ au catalogue).

## 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 <br> A Journal Published by Statistics Canada 

Volume 34, Number 2, December 2008

## Contents

In this issue ..... 127
Waksberg Invited Paper Series
Mary E. Thompson
International surveys: Motives and methodologies ..... 131
Regular Papers
Joanne Pascale and Alice McGee
Using behavior coding to evaluate the effectiveness of dependent interviewing ..... 143
Jing Xu, Jun Shao, Mari Palta and Lin Wang
Imputation for nonmonotone last-value-dependent nonrespondents in longitudinal surveys ..... 153
Robert G. Clark and Raymond L. Chambers
Adaptive calibration for prediction of finite population totals ..... 163
Lionel Qualité and Yves Tillé
Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added ..... 173
Inho Park
PSU masking and variance estimation in complex surveys ..... 183
Roberto Benedetti, Giuseppe Espa and Giovanni Lafratta
A tree-based approach to forming strata in multipurpose business surveys. ..... 195
Mohammad G. Mostafa Khan, Niraj Nand and Nesar Ahmad
Determining the optimum strata boundary points using dynamic programming ..... 205
José A. Díaz-García and Liliana Ulloa Cortez
Multi-objective optimisation for optimum allocation in multivariate stratified sampling ..... 215
Piero Demetrio Falorsi and Paolo Righi
A balanced sampling approach for multi-way stratification designs for small area estimation. ..... 223
Danny Pfeffermann, Bénédicte Terryn and Fernando A.S. Moura
Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries ..... 235
Acknowledgements. ..... 251

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences - Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" - "Permanence of Paper for Printed Library Materials", ANSI Z39.48-1984.

## In this issue

This issue of Survey Methodology opens with the eighth 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 - Sharon Lohr, chair, Bob Groves, Leyla Mohadjer and Wayne Fuller - for having selected Mary Thompson as the author of this year's Waksberg Award paper.

In her paper entitled "International surveys: Motives and methodologies" Thompson discusses the challenges of organization and data collection that arise in conducting surveys in several countries simultaneously, and also issues in analyses that compare and contrast different countries included in the surveys. She describes several examples of international surveys, and illustrates the challenges by discussing issues arising in the International Tobacco Control survey in greater detail. She also considers several different methods for calibration of measurements and cross-cultural comparisons.

Pascale and McGee present a study on the use of dependent interviewing which is used in many longitudinal surveys to 'feed forward' data from one wave to the next. Using recordings of field interviews, the authors use behavior coding to evaluate the effectiveness of dependent interviewing. The paper gives some interesting insight into how the type of data fed forward influence the way that interviewers ask the question.

The paper by Xu, Shao, Palta and Wang deals with imputation of missing values in longitudinal surveys when the nonresponse pattern is not monotone. The authors assume that the nonresponse mechanism at the current period depends only on the last value of the variable to be imputed and that this variable follows a Markov chain. Imputation is performed through a series of nonparametric regression models. A bootstrap method is employed for variance estimation. The method is illustrated through the use of both simulated as well as real data.

In the context of the prediction framework, Clark and Chambers propose an adaptive calibration approach for selecting an appropriate set of auxiliary information. They apply their method to a wide range of models. Results of a simulation study are presented confirming the good performance of the proposed methods.

The variance estimation of estimators of change between two successive periods of a repeated survey is studied by Qualité and Tillé. In this article, we take into account, among others, the sampling design which uses a panel, total non-response, and the reduction of the sampling weights due to outliers and calibration. The proposed methodology is applied to the Swiss survey of value added.

In his paper, Park considers the problem faced by analysts when using public-release data that have been modified for confidentiality purposes. In particular he looks at the effect of swapping primary sampling units (PSUs), commonly done to protect the identity of survey respondents, on the calculation of variances. He proposes a new PSU swapping algorithm and compares its effect on variance estimation both theoretically and empirically with some existing methods.

Benedetti, Espa and Lafratta propose a sequential process to stratify a finite population. This process is for obtaining a multivariate stratification and uses an approach based on the development of a tree. With this process they produce successively finer and finer partitions of the population until the difference between the optimal sample sizes obtained in two consecutive steps is less than a predetermined level. The proposed approach is applied to the Italian Farm Structure Survey.

Khan, Nand and Ahmed consider the problem of finding optimum stratum boundaries as the problem of determining optimum stratum widths. They formulate it as a mathematical programming problem and solve it by extending Bühler and Deutler's (1975) dynamic programming approach. The paper is an extension of this dynamic programming approach to variables of interest following triangular and standard normal distributions. A small simulation study compares the proposed method to the cumulative square root $(f)$ method of Dalenius and Hodges (1959) revealing gains in efficiency.

In their paper, Díaz-García and Cortez study the problem of allocation in multivariate stratified sampling. They express this problem as a non-linear problem of matrix optimization of integers constrained by a cost function or by a given sample size. They apply their method using data from a forest survey.

Finally, the 2007 ISI (International Statistical Institute) Satellite Conference on Small Area Estimation, SAE 2007, produced a myriad of papers illustrating the latest small area techniques. Several of these papers were submitted to Survey Methodology. Two of them round out the current issue; we expect that more will be published in future issues.

Falorsi and Righi develop a sampling strategy to obtain planned sample sizes for domains subject to predetermined sampling errors, particularly when the cross-classification of variables defining the different partitions would yield a number of strata larger than the overall sample size. The proposed method has the advantage of computational feasibility and the implementation of a small area strategy that comprises the sampling design and the estimation jointly and improves the efficiency of the direct domain estimators.

In the last paper of this issue Pfeffermann, Terryn and Moura consider situations where the target response value is either zero or an observation from a continuous distribution, for example when assessing literacy skills with the possible outcome being either zero, indicating illiteracy, or a positive score measuring the level of literacy. Available methods, however, are not suitable for this kind of data because of the mixed distribution of the responses, having a large peak at zero juxtaposed to a continuous distribution for the rest of the responses. The authors develop a suitable two-part random effects model and show how to estimate the model and assess its goodness of fit, and how to compute small area estimators of interest and measure their precision.

And finally, we are please to inform readers and authors that Survey Methodology is now cited in the ISI Web of knowledge, which included Current Contents/Social and Behaviorial Sciences, Social Sciences Citation Index, and the Science Citation Index Expanded, starting with the June 2007 issue.

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)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)

Alastair Scott (2006)
Carl-Erik Särndal (2007)
Mary Thompson (2008)

## Nominations:

The author of the 2010 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, Leyla Mojadjer, by email to MOHADJL1@WEStat.com. Nominations and suggestions for topics must be received by February 27, 2009.

## 2008 Waksberg Invited Paper

## Author: Mary Thompson

Mary Thompson is University Professor of Statistics at the University of Waterloo in Waterloo, Ontario. She is Co-Director of the University of Waterloo Survey Research Centre, and Director of the Data Management Centre for the International Tobacco Control Policy Evaluation Project. Her research interests over the years have focused on theory of estimation, particularly in complex surveys, and she is the author of a monograph entitled Theory of Sample Surveys, published in 1997. Recently, she has become particularly interested in the analysis of survey data relevant to latent variable models in the social and health sciences.

# Members of the Waskberg Paper Selection Committee (2008-2009) 

Leyla Mojadjer (Chair), Westat
Wayne A. Fuller, Iowa State University
Daniel Kasprzyk, Mathematica Policy Research
Betsy Martin

## Past Chairs:

Graham Kalton (1999-2001)
Chris Skinner (2001-2002)
David A. Binder (2002-2003)
J. Michael Brick (2003-2004)

David R. Bellhouse (2004-2005)
Gordon Brackstone (2005-2006)
Sharon Lohr (2006-2007)
Bob Groves (2007-2008)

# International surveys: Motives and methodologies 

Mary E. Thompson ${ }^{1}$


#### Abstract

The context of the discussion is the increasing incidence of international surveys, of which one is the International Tobacco Control (ITC) Policy Evaluation Project, which began in 2002. The ITC country surveys are longitudinal, and their aim is to evaluate the effects of policy measures being introduced in various countries under the WHO Framework Convention on Tobacco Control. The challenges of organization, data collection and analysis in international surveys are reviewed and illustrated. Analysis is an increasingly important part of the motivation for large scale cross-cultural surveys. The fundamental challenge for analysis is to discern the real response (or lack of response) to policy change, separating it from the effects of data collection mode, differential non-response, external events, time-in-sample, culture, and language. Two problems relevant to statistical analysis are discussed. The first problem is the question of when and how to analyze pooled data from several countries, in order to strengthen conclusions which might be generally valid. While in some cases this seems to be straightforward, there are differing opinions on the extent to which pooling is possible and reasonable. It is suggested that for formal comparisons, random effects models are of conceptual use. The second problem is to find models of measurement across cultures and data collection modes which will enable calibration of continuous, binary and ordinal responses, and produce comparisons from which extraneous effects have been removed. It is noted that hierarchical models provide a natural way of relaxing requirements of model invariance across groups.


Key Words: International surveys; Longitudinal surveys; Analysis of survey data; Random effects; Data collection mode effects; Hierarchical models; Measurement models.

## 1. Introduction

I have chosen the topic of international surveys since one such survey, the International Tobacco Control survey, has been a major part of my activity in the last few years, and there are some interesting intersections with the areas to which Joseph Waksberg gave his attention - particularly frames for telephone surveys, and the effects of stratification with widely varying sampling rates.

The paper will begin with some discussion of the motifs and motives of international surveys and some examples. It will touch on the challenges of organization, data collection and analysis. Finally, it will consider two problems to be addressed in analysis: (i) survey sampling theory and the pooling of data from several countries, and (ii) measurement across data collection modes and cultures.

The first large international survey was the World Fertility Survey, carried out in the 1970's through the International Statistical Institute, and funded by the U.S. Agency for International Development and other sponsors. It was a very ambitious one-time survey. The WFS eventually surveyed over 330,000 women in 61 countries, at a cost of about $\$ 50$ million. It gave countries important comparison data on family sizes, and led to policy measures on population planning in several participating countries. It also produced analytic projects in the hundreds, including path-breaking methodological studies, and laid the foundation for international survey methodology, particularly in
developing countries (Verma, Scott and O'Muircheartaigh 1980; Cleland and Verma 1989).

Another well known example is the Programme for International Student Assessment, a project of the Organization for Economic Co-operation and Development, beginning in 2000. PISA is a continuing survey, carried out every 3 years, with 15 year old youths in developed countries. It is growing in scope, with 67 countries expected to participate in 2009. The results allow countries to monitor the success of their education programs in providing verbal and quantitative literacy.

The Global Youth Tobacco Survey is a one-time survey which began in 2002, sponsored by the World Health Organization and the Centers for Disease Control and Prevention. The GYTS has focused on surveying youth aged 13 to 15 years in developing countries, and had carried out data collection in 129 countries by 2004. The objective is to measure tobacco use uptake among youth, and awareness of the associated health risks.

The European Social Survey (ESS 2008) is an "acad-emically-driven social survey" in over 30 nations, funded by European and national agencies, and designed to "chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations".

Even as the use of local and national surveys is growing everywhere, so too is the incidence of international surveys, carried out by international agencies, non-governmental

[^0]organizations and private sector firms. This burgeoning appears to be part of a trend toward global governance and concern for population health and well-being.

I have seen the purposes of international surveys classified as epidemiology, surveillance, monitoring and evaluation of the effects of policy. Evidently these classifications overlap. It can be argued that PISA, the GYTS and the ESS constitute surveillance and monitoring, because their data are related only indirectly to interventions. The WFS had a direct evaluation aspect, in countries that had introduced family planning programs. The International Tobacco Control (ITC) survey, to be discussed later in this section, is one of the few for which evaluation is the primary purpose.

Apart from scientific concerns, another important role for an international survey is to engage the governments of the countries; it provides a way for them to participate in global policy development even in the face of political and economic obstacles.

For the researcher, international surveys allow the comparison of the populations of countries, the possibility of interpretation of the differences, and sometimes even the possibility of shedding light on causes and effects typically with the underlying aim of improving conditions and informing policy.

The International Tobacco Control Policy Evaluation Project (ITC Project) was initiated by Dr. Geoffrey T. Fong of Psychology at the University of Waterloo, with collaborators around the world (Fong, Cummings, Borland, Hastings, Hyland, Giovino, Hammond and Thompson 2006; Thompson, Fong, Hammond, Boudreau, Dreizen, Hyland, Borland, Cummings, Hastings, Siahpush, Mackintosh and Laux 2006). The impetus was the WHO Framework Convention on Tobacco Control (FCTC), which was passed in May 2003, and has been ratified by over 150 countries. By ratifying the treaty the participating countries pledge to introduce policy measures for tobacco control, such as strong health warning package labels, banning of cigarette advertising, and banning of smoking in public places. The necessity for national legislation has as a consequence that these measures are being introduced at various times and in various ways. For example, Canada in December 2000 introduced graphic warning labels, setting international precedents for the size of label (more than 50 $\%$ of the package) and vivid colour images. Since then a few other countries have adopted this same practice, while others have legislated prominent text warnings. For the current status of health warning regulations around the world, see ITC (2008). The MPower Report (WHO 2008) describes the global tobacco policy environment and six policies of focus for the FCTC.

The purpose of the ITC Project is to try to find out which measures are effective in reducing uptake of tobacco use, and in helping people already using tobacco to quit. Furthermore, it has the ambitious aim of trying to explain how those measures which are effective actually work. The investigating team includes social psychologists and specialists in social marketing, as well as epidemiologists and economists.

By September 2008 the ITC Project was carrying out surveys in 17 countries, with more likely to be added. The surveys began in 2002, in Canada, the US, the UK and Australia. That year, in each of the four countries, approximately 2000 adult smokerswere recruited by telephone using a geographically stratified random digit dial (RDD) frame, of which the science has origins in the famous Mitofsky-Waksberg method (Waksberg 1978). The recruited smokers were interviewed a week or two later, and have been followed up each year since then, regardless of whether they continued to smoke. Wave 6 for the ITC Four Country Survey was completed in February 2008.

Because sufficient sample size is needed to evaluate the effects of measures introduced between the waves, dropouts at each wave have been replaced with a cohort of new recruits. In the ITC Four Country Survey, new recruits in each wave have been selected using the same design as in Wave 1 , without any attempt to match the characteristics of the dropouts. Weights construction at each wave is effectively carried out separately for each cohort, adjusting for differential attrition by region and by age-sex group. This design has helped us to discern "time-in-sample" effects, and time-in-sample is entertained as an explanatory variable in analytic models (Thompson, Boudreau and Driezen 2005).

The first national policy measures following 2002 were an advertising ban and enhanced warning labels in the UK, between Waves 1 and 2; graphic warning labels were introduced in Australia between Waves 4 and 5. In the ITC Four Country Survey we have what is sometimes called a natural experiment or quasi-experiment (Cook and Campbell 1979), where the non-policy countries serve as external controls; moreover, the longitudinal feature of the design provides internal control. The design has been replicated a number of times, with other sets of countries.

For example, it became clear early on that Ireland would be the first country to adopt national smoke-free legislation, coming into effect in March 2004. The ITC collaborators were able to put together parallel surveys in Ireland and the UK before the law came into effect, and to visit the same people a year later. The samples were again recruited nationally using a random digit dial (RDD) frame. There were 755 smokers in Ireland and 411 smokers in the UK who were interviewed at both waves. One interesting
finding concerned support for a ban on smoking in pubs (Fong, Hyland, Borland, Hammond, Hastings, McNeill, Anderson, Cummings, Allwright, Mulcahy, Howell, Clancy, Thompson, Connolly and Driezen 2006). Figure 1 shows the proportions supporting or strongly supporting the ban in bars and pubs, in the two countries, by wave.


Figure 1 Support for smoke-free legislation in two waves of ITC Ireland/UK Survey

In the ITC sample of smokers the support increased between the two waves a little in the UK, and a great deal in Ireland. Moreover, the same survey showed no evidence that the reduction of smoking in public venues was associated with increased smoking in private venues. Showing broad acceptance of the smoke-free law by smokers, the ITC findings and others like them have helped bring about similar laws in Scotland, France, Germany, the rest of the UK, and the Netherlands. An ITC survey was carried out before and after the April 2006 implementation of the ban in Scotland, using the rest of the UK as the control, and the findings were replicated, except that by this time support in the rest of the UK had grown substantially (Hyland, Hassan, Higbee, Fong, Borland, Cummings, Thompson, Boudreau and Hastings 2008).

The model used for testing was simple: a GEE model, where $Y$ is a binary measure of support for the ban, $w$ is country, $t$ is time, the wt term represents an interaction, and $x$ is a vector of fixed individual level covariates:

$$
\begin{aligned}
& \operatorname{logit}\left[P\left(Y_{t}=1 \mid w, x\right)=\alpha_{0}+\alpha_{1} w+\gamma t+\delta w t+x \beta,\right. \\
& \operatorname{Corr}\left(Y_{1}, Y_{2}\right)=\rho .
\end{aligned}
$$

The coefficient $\delta$ represents the difference in increase in support in the two countries, and we tested the hypothesis $H_{0}: \delta=0$. There are other possible parametrizations, but this one has the advantage of matching the plot in Figure 1, which displays marginal proportions; the methodology is widely accepted, and supported by complex survey software.

## 2. Challenges

There are numerous challenges in carrying out an international survey. The WFS papers by Verma et al. (1980) and Cleland and Verma (1989) can be recommended for thoughtful discussions which are very little out of date. In this section we illustrate by describing some of the issues encountered by the ITC survey in organization and data collection.

Unlike the WFS, the ITC survey has been funded in the first instance by national granting programs, primarily the National Institutes of Health in the United States, and the Canadian Institutes for Health Research. The central infrastructure, led by Dr. Fong at the University of Waterloo and by Dr. K. Michael Cummings at Roswell Park Cancer Institute in Buffalo, works directly with investigating teams and agencies in the various countries. We have had to learn how to work with widely varying societies, political systems and cultures. Survey costs and budgets alone differ surprisingly from country to country. When governments contribute funding, they have their own requirements, and data ownership agreements must be negotiated. Since the amount of infrastructure and expertise can be quite different from place to place, the close coordination of the ITC Four Country Survey is difficult to replicate more widely.

For example, in the first half of 2008 the fieldwork was carried out for Wave 3 of a parallel survey (the ITC SouthEast Asia Survey) in Thailand and Malaysia, which are geographically close, and similar in some ways, but different in many dimensions. Thailand is ethnically quite homogeneous, while Malaysia has three major ethnic groups and many minor ones. More than half the population of Thailand lives in rural areas, but most of the Malaysian population is urban, and residential mobility is high. Thailand has extensive experience with surveys, including cohort studies, but when Wave 1 began in 2005, Malaysia was attempting this kind of cohort study for the first time. We tried to prescribe parallel sampling designs in the two countries, but had to make compromises. For example, it was found at the time of Wave 1 that the official sampling frames had different sized building blocks at the lowest level, consisting of clusters of households. This difference made the sample of households rather more dispersed in Malaysia, which had the smaller blocks. The greater dispersion meant greater work and costs. (Design effects are still larger for Malaysia than for Thailand, because of more heterogeneity at the level of the first stage units.)

An important aspect of the project is to try to build capacity for longitudinal health surveys in countries which are relatively new to this kind of work. We provide detailed protocols, training manuals, and data entry templates. We have learned to be more insistent on the identification of
local expertise, particularly statistical expertise. Day-to-day communication is carried out by email and teleconferences. Final data cleaning and construction of survey weights normally occur at the University of Waterloo, but some country teams have been eager to participate in these parts of the operation.

We use telephone surveys, with recruitment by modified RDD, in the four original countries, as well as Ireland, South Korea, France and Germany; recruitment from the National Health Survey sample in New Zealand; and face to face surveys in Thailand, Malaysia Wave 1, China, Bangladesh, Mexico, and Uruguay.

In Malaysia Wave 2 we intended face-to-face data collection, but since both recontact and new recruitment proved to be very difficult because of a combination of factors, we moved to telephone interviewing where feasible in some areas. There was limited scope for comparison of modes, but in the large and mainly urban state of Selangor, 137 Wave 1 smokers (non-quitters) were reinterviewed face-to-face, and 63 were reinterviewed by telephone, making some tentative inferences possible. In Wave 3, an attempt was made to carry out both telephone and face-toface interviewing in some of the same census districts, to enable a better assessment of data collection mode effects, and this study is in progress. At the same time, the proportion of the ITC Malaysia smoker sample interviewed face-to-face has decreased steadily, from $100 \%$ in Wave 1, to $63.5 \%$ in Wave $2,44.4 \%$ in Wave 3. In Wave 4, we expect that telephone will be used exclusively in the mainland states.

The Wave 1 survey in the Netherlands has used parallel internet panel and RDD telephone interviewing, with sample sizes of about 400 and 1,800 respectively. This exercise will provide our best chance yet at being able to account for mode effects in modeling. These effects have been the subject of much research recently. For example, some studies have found that telephone respondents choose the extreme options of a Likert scale more of the time than web respondents do (Wichers and Zenderink 2006; Bronner and Kuijlen 2007).

The internet sample in the Netherlands consists of smokers randomly sampled from a large pre-recruited multipurpose panel of about 200,000 people assembled by the firm TNS NIPO. The telephone sample, representing smokers accessible by land-line telephone, might well represent a different population of smokers. The low telephone response rates make clear that the public in the Netherlands is not as receptive to telephone surveys as the public in most of the other ITC countries. We requested that each group be asked about their accessibility by the other mode, so as to be able to use dual frame methods (Lohr and Rao 2000) to compute appropriate survey weights. We will
also model propensity (Rosenbaum and Rubin 1984) for responding by telephone (say), given demographic variables and the accessibility variables, and control for propensity score in comparison of response patterns by mode.

Response rates vary a great deal, even within the ITC Four Country Survey, the response rates and retention rates being highest in Australia, and lowest in the United States. Certainly this jeopardizes the ability to compare across countries, in the sense that we can only compare the populations represented by the respondents - those in each country who would respond if approached under our protocol. The situation looks slightly better if we break response rates down into components. For example, we have seen from call attempt outcomes, and from our knowledge of the increased use of call filtering devices, that US adults are much harder to contact and recontact than adult residents of the other three countries. However, once contact is established, the US agreement or non-refusal rate is very similar, upwards of $80 \%$, to those of the other three countries.

We have measurement issues even for matters of fact, such as habits of tobacco purchase and use. In some countries like India, Bangladesh and Sudan, all under discussion for inclusion, there are many forms of tobacco in common use. For the developed countries, just keeping the list of cigarette brands current is a full time job. Compounding the difficulty is that whenever we ask about purchasing patterns or noticing advertisements we are asking people to remember what they have done over the previous two weeks, or some longer period. For the most part, we rely on self-report, and for a number of reasons self-report may not be accurate.

For attitudes and beliefs we have known all along that the questions must be suited to the language and the literacy level of the participants, but we were still surprised and sobered to find a high incidence of item non-response in outlying areas of one country, suggesting great difficulty with attitude and belief questions. In our pilot survey in India, the survey took an average of 1.5 hours per participant, despite having been shortened and simplified.

Psycho-social measurements need to be validated in each culture and language. For example, we have started to include a very short depression scale. Here is the version for the ITC Four Country Survey.

- During the last month, have you often been bothered by little interest or pleasure in doing things?
- During the last month, have you often been bothered by feeling down, depressed or hopeless?
- In the last year, have you been told by a doctor or other health care provider that you have depression?

And here is the version that we finally came to for ITC China Wave 2, on the advice of other researchers who reported having been able to validate a version of it.

Below is a list of ways that you might have felt or behaved. Please tell me how often you have felt this way during the past week.

1. I did not feel like eating; my appetite was poor.
2. I felt hopeful about the future.
3. I felt sad.
4. I felt that people dislike me.

Ryder, Yang, Zhu, Yao, Yi, Heine and Bagby (2008) have released results of a very interesting comparative study of the expression of depression.

The catalogue of measurement issues goes on. Even if the question is supposed to be the same in two languages, it may be hard to find equivalences. We try to ensure a good quality translation using committee translation or comparison of independent translations, but must often fall short of perfection. For example, literal translations of English into French or German are a fair amount longer, and it takes considerable skill to make a translation that runs smoothly over the telephone. Thrasher, Quah, Borland, Awang, Sirirassamee, Boado, Miller, Watts and Dorantes (2008) describe a study in cognitive testing of some of the most important questions, across several countries.

There are more subtle cultural differences, particularly the degree to which respondents will give a socially desirable response. We have noticed what may be a higher tendency toward this among Mexicans and among anglophone Canadians. Johnson and Van de Vijver (2003) among others have discussed the possibility that crossnational differences in socially desirable responses may be related to "cultural value systems such as in the individualism and collectivism dimension" of Hofstede (1980).

In a longitudinal survey, we need to be concerned as well about the validity and reliability of repeated measures. As we have already indicated, it is common to observe what are called "time-in-sample effects", where the response proportions tend to drift upward or downward as the cohort proceeds, just because of the fact of being measured.

All these issues feed into the analytic challenges faced by researchers. Fundamentally, the aim of analysis must be to discern the real response (or lack of response) to policy change, separating it from the effects of data collection mode, differential non-response, external events, time-insample, culture, and language. This is a daunting task.

## 3. Pooling of data across countries

In the traditional survey analysis paradigm (Binder 1983; Godambe and Thompson 1986; Skinner 1989), there is a
model for the responses $y$ with parameter $\theta$, and we imagine how we would estimate $\theta$ if we had responses from the whole population, in a census. We would use an efficient unbiased estimating equation like this:

$$
\sum_{i=1}^{N} \phi_{i}\left(y_{i}, \theta\right)=0
$$

to define a census estimate. To obtain the sample estimate, we use a weighted sum of the sample estimating function terms:

$$
\sum_{i \in s} w_{i} \phi_{i}\left(y_{i}, \theta\right)=0
$$

to give an approximately unbiased estimator of the census estimating function. The survey weights are constructed to take into account the sampling design, and underrepresentation of some groups due to non-response and noncoverage. The usual interpretation of $w_{i}$ is the number of population members represented by $i$. The use of this sample estimating function is appealing because of the likely reduction of bias due to informative sampling and non-response; but if the weights are highly variable and the model for the terms is correct, the second equation gives an inefficient way of estimating $\theta$.

Now when we are combining data from two countries with very different sampling fractions, as in the Ireland/UK survey, the weights for one country (the UK) will be much greater than the weights for the other country (Ireland). A literal application of the paradigm would have the data from the UK dominating the analysis. If the model is correct, the most efficient census estimate is the mean of $y$ over the two countries combined. But then the corresponding sample estimate is an inefficient use of the sample. This problem is similar to that arising in case-control studies, as discussed by Scott (2006).

One way of producing better estimates while remaining in the traditional paradigm is to consider that the parameter value for the UK is $\theta-\Delta$, that the parameter value for Ireland is $\theta+\Delta$, and that we are trying to estimate $\theta$, the arithmetic mean of the two. An efficient census estimating function system for $\theta$ and $\Delta$ is equivalent to one which separates into a part for each country. Since rescaling of weights within a country has then no effect on the point estimators and their properties, the survey-weighted sample version of that system yields efficient estimation.

Moreover, the ensuing analysis is the approximately the same as we would obtain from the original paradigm if we had equal sample sizes in the two countries, and rescaled the weights to sum to sample size within each country. As noted by Scott (2006), rescaling the weights in this manner is a very common practice among epidemiologists. It is in a sense a partial application of the $q$-weight method of

Pfeffermann and Sverchkov (1999), where the inverse inclusion probability weight is divided by a kind of expectation of the weight, conditional on an explanatory variable (country).

In estimating a mean parameter $\theta$, a somewhat more appealing suggestion is to consider a random effects model, where $Y=\theta+u+e$, and $u$ is a country random effect, then to develop a census estimating function system which is efficient for estimating the parameter $\theta$. For example, if

$$
Y_{1 i}=\theta+u_{1}+e_{1 i} \text { and } Y_{2 i}=\theta+u_{2}+e_{2 i}
$$

and if the variance components corresponding to $u$ and $e$ are known, then the best combination of the two country means for estimating $\theta$ is

$$
a \bar{Y}_{1}+(1-a) \bar{Y}_{2},
$$

where

$$
a=\frac{1}{2}\left\{\frac{\sigma_{u}^{2}+\sigma_{e}^{2} / N_{2}}{\sigma_{u}^{2}+\frac{\sigma_{e}^{2}}{2 N_{1}}+\frac{\sigma_{e}^{2}}{2 N_{2}}}\right\} .
$$

Notice that if $\sigma_{u}^{2}=0$, the census estimator becomes the mean of $y$ over the two countries combined. However, if $\sigma_{u}^{2}$ is dominant, the best estimator is the arithmetic mean of the country means. From a pooled sample, the usual paradigm gives the same convex combination of withincountry sample-based mean estimators.

More generally, we can replace $\theta$ in each country census estimating function by $\theta+u$, where again $u$ is a country random effect. Then the best combination of two country census estimating functions for $\theta$ is

$$
c_{1} \sum_{i=1}^{N_{1}} \phi_{1 i}\left(Y_{1 i}, \theta, u_{1}\right)+c_{2} \sum_{i=1}^{N_{2}} \phi_{2 i}\left(Y_{2 i}, \theta, u_{2}\right)
$$

where $\quad c_{1}=\left[\operatorname{Var}\left(\sum_{i=1}^{N_{2}} E\left(\phi_{2 i} \mid u_{2}\right)\right)+E\left(\sum_{i=1}^{N_{2}} \operatorname{Var}\left(\phi_{2 i} \mid u_{2}\right)\right)\right]\left[\sum_{i=1}^{N_{1}}\right.$ $\left.E\left(\partial \phi_{1 i} / \partial \theta\right)\right]$, and $c_{2}$ is defined symmetrically. If the first term in square brackets in $c_{1}$ dominates, the corresponding sample-based estimating function weights the terms comparably in the two samples.

Even in the simple case of a mean, the parameters of the random effects model will not be known, and will be difficult to estimate when there are only two countries, but conceptually the model seems to be a useful one. When there are several reasonably similar countries or regions (for example the seven cities of the ITC China survey), linear models with random effects are estimable in the usual paradigm, as described for example in a more general setting by Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998).

As an aside, the GEE analysis of the Ireland data described earlier was a pooled analysis, and all its "effects" were regarded as fixed. The model is nearly "saturated", with two time points and two countries accounting for the four main parameters. It is possible to see that with ordinary survey weights, the estimation of $\beta$ and $\rho$ would be dominated by the UK data. However, if $\beta$ and $\rho$ are known, then as in the case of the parameters $\theta$ and $\Delta$ in the example of the mean, the equations for the main parameters separate into two pairs, one pair for $\alpha_{0}$ and $\gamma$, and one pair for $\alpha_{0}, \alpha_{1}, \gamma$ and $\delta$, each involving weights from only one of the two countries. Thus the estimation of the main parameters is less affected by the scaling of the weights. If the estimation of $\beta$ were also important to us, we might consider it to be the mean of a country level random variable, leading naturally to each of the two samples having appropriate influence. (In fact, in our analysis, we did not do this; the weights were rescaled to sum to sample size within country.)

The foregoing discussion of pooled analysis assumes that there exists a parameter $\theta$ that has the same interpretation and relevance across countries. Most multi-country analyses start from this assumption. Indeed, de Leeuw and Hox (2003) state as a requirement for a meta-analysis that "all studies must estimate the same fixed parameter, and all variance is assumed to be sampling variance". But in fact a central issue for debate is the question of when it is appropriate to make a model that is to apply to the data from several countries simultaneously. Sometimes it may be most appropriate simply to consider the country models to be separate but parallel. For example, in countries at different stages of development, introducing the same relative increase in real price of cigarettes can be expected to lead to decreases in cigarette consumption; but since the linear model is at best a useful local approximation to the complex relationship between price and consumption, there is no reason to suppose that the decreases will be of the same magnitude, or that the two regression estimates are measuring the same quantity.

For another example, one of the models of interest to the ITC project is the mediational model of the figure 2, postulating how "noticing" health warning labels might affect intention to quit.

The distribution at baseline of the intention to quit in the various countries is quite variable. The same is true for the other variables in the model. Is it reasonable to hope that the relationships among these variables might be less variable across countries? In fact it appears that for the original four countries, they are. Even though UK smokers were much the most likely to say they had no plan to quit, it was still the case for them that "health concern" (label-triggered consciousness of health effects) predicts quit intention, and
that "health concern" was elevated with increased noticing of labels (Hammond, Fong, Borland, Cummings, McNeill and Driezen 2007). Thus it is not unreasonable to explore a model like the one in Figure 2 for the data from the four countries, pooled. Regardless of weighting issues, in the regression of the mediator "health concern" on "noticing" health warning labels, and in the regression of intention to quit on both of these, we have found it convenient to take the country means to be fixed effects. On the other hand, since the estimated regression coefficients for the countries modeled separately vary moderately, it is natural in the pooled analysis to conceptualize the regression coefficients as having random country components.

This discussion can be summarized and elaborated in the following points:

- An analysis which pools data across countries should be adopted with caution. For such an analysis to be appropriate, the model structure (the regression equation and its variables) should be correct for all countries, and the assumption of common parameters should be supported by theory and observation. Robust variance estimation which respects the country sampling designs will be necessary when the sampling designs are complex.
- If the set of parameters of a pooled model can (through transformation) be separated into disjoint subsets corresponding to the countries, the estimation of those parameters is not affected by large differences in sampling fractions among countries, and is not affected by rescalings of the weights within countries.
- If a fixed mean or regression parameter is deemed to be common to the countries, estimation using inverse inclusion probability weights will be inefficient if the sampling fractions are widely variable.
- Alternatives to simple weight rescaling are to make the mean or regression parameter a fixed effect varying by country (which leads to separation into disjoint subsets, but increases the number of parameters and removes the "common-ness"); or to make the mean or regression parameter a random effect, varying by country (which leads to approximate separation and retains the "common-ness").
- It is conceptually appealing to make the intercept fixed and the slope random, since baseline levels tend to vary much more by country than slopes do. In implementation, this approach requires enough countries to make estimation of variance components feasible, and a small number of random effects to be integrated.
- When a pooled analysis is problematic, less formal comparison of the results of parallel country analyses may accomplish most of what is desired.


## 4. Calibration of measurements and cross-cultural comparisons

The other statistical problem which I would like to highlight is the use of measurement models to try to calibrate measurements across modes and compare measurements across cultures. A common approach is to consider that with each questionnaire item we are measuring a construct, like "social denormalization" (perception of societal disapproval), and to think of the construct as a continuous variable $\eta$. The distribution of $\eta$, conditional on explanatory variables, determines a distribution of responses to the questionnaire item.

If we have several items of the same kind measuring a construct, a conceptual model for continuous measurements $y$ might be $Y_{i k}=b_{i} \eta_{k}+a_{i}+e_{i k}$ for item $i$ and participant $k$. Here $b_{i}$ represents a positive scaling for item $i, a_{i}$ a location shift, and $e_{i k}$ a normal mean zero measurement error with variance $\sigma_{e i}^{2}$ not dependent on $k$. Assume all $e_{i k}$ to be independent, and independent of the $\eta_{k}$. (This is effectively an assumption that $\eta$ is the only latent determinant of $Y$.) If we take the distribution of $\eta$ to be $N(0,1)$, as we may if $\eta$ is normal with no explanatory variables, then the distribution of $Y_{i k}$ is $N\left(a_{i}, b_{i}^{2}+\sigma_{e i}^{2}\right)$, and for a single item $i$, the parameters $a_{i}$ and $b_{i}^{2}+\sigma_{e i}^{2}$ are estimable from the marginal data on many participants. If there are at least two items with the same variances, then since the item responses for a participant have covariances of form $b_{i_{1}} b_{i_{2}}$, all parameters are estimable from the marginal data on many participants. Given values for the item parameters, the value of $\eta$ for a participant can be "predicted" from the posterior distribution of $\eta$, given the participant's item responses.


Figure 2 Mediational model of policy effects: Warning labels

If the measurement $y$ is binary, it is common to take an Item Response Theory (IRT) model $\operatorname{Prob}\left(Y_{i k}=1 \mid \eta_{k}\right)=$ $H\left(b_{i} \eta_{k}-\gamma_{i}\right)$ where $H$ is the standard normal or the logistic cumulative distribution function (c.d.f.). The parameter $b_{i}$ is the "discrimination parameter" for the item, and $\gamma_{i}$ is a threshold, such that the probability of response 1 exceeds $1 / 2$ when the construct scaled by $b_{i}$ exceeds $\gamma_{i}$. The unconditional probability that $Y_{i k}=1$ is obtained by integrating with respect to the distribution of $\eta_{k}$, given fixed explanatory variables for participant $k$. In this simplest case it appears that at least 3 items are needed (with many participants) for all parameters to be estimable, since they would yield 7 joint probabilities for the estimation of 6 parameters. Again, given values for the item parameters, the value of $\eta$ for a participant can be predicted, given his or her set of item responses. See for example Lu, Thomas and Zumbo (2005). Standard latent variable estimation software can be used to produce these inferences, and their analogues in the case of ordinal measurements.

First let us consider the calibration problem. Suppose there are two data collection modes, and for item $i$ in mode $j$ with participant $k$ we have the continuous measurement

$$
Y_{i j k}=\beta_{j}\left(b_{i} \eta_{k}+a_{i}+e_{i j k}\right)+\alpha_{j}+\varepsilon_{i j k} .
$$

This model, in which $\alpha_{j}$ and $\beta_{j}$ do not depend on the item $i$, might be appropriate for a set of items all of the same general type. Plausible examples are not abundant, but one such might be a series of questions of form: "What percentage of the time would you say you feel ...", where the respondent is asked to give a percentage over the telephone, or asked to mark a position on a line on paper.

If we take the $a_{i}$ and $b_{i}$ to be the parameters of the items using the first data collection mode, we may set $\alpha_{1}=0$ and $\beta_{1}=1$. If $\beta_{2}$ is greater than 1 , there is a tendency for a wider variation, or more extreme responses, under the second collection mode. If $\alpha_{2}$ is greater than 0 , respondents tend to give a higher response under the second collection mode than under the first. Note that the samples for the two modes involve different participants. If we can assume that the distribution of $\eta$ is the same for the two mode samples (an assumption which effectively requires randomization to mode), we have the distribution of $Y_{i l k}$ as before, $N\left(a_{i}, b_{i}^{2}+\sigma_{e i}^{2}+\sigma_{\varepsilon}^{2}\right)$, while the distribution of $Y_{i 2 k}$ is

$$
N\left(\beta_{2} a_{i}+\alpha_{2}, \beta_{2}^{2} b_{i}^{2}+\beta_{2}^{2} \sigma_{e i}^{2}+\sigma_{\varepsilon}^{2}\right)
$$

If $\sigma_{\varepsilon}^{2}=0$, then given data on one item $i$ in the two modes, we can estimate $\alpha_{2}$ and $\beta_{2}$, assuming $\beta_{2}$ is positive. If $\sigma_{\varepsilon}^{2}>0$, the parameters $\alpha_{2}, \beta_{2}$ and $\sigma_{\varepsilon}^{2}$ are estimable provided that there are at least two items available - of the same type, but with differing values of $a$ and $b$.

These considerations can be extended to the more usual case of items with ordinal responses, by imagining an ordinal response probability to be determined by an underlying continuous response. For binary data, we would most simply set

$$
P\left(Y_{i j k}=1 \mid \eta_{k}\right)=H\left(\beta_{j}\left(b_{i} \eta_{k}+a_{i}\right)+\alpha_{j}\right),
$$

with $\alpha_{1}=0$ and $\beta_{1}=1$. If the distribution of $\eta_{k}$ is the same for both modes, then from data on many participants and three items we can identify all parameters. Adding an explanatory variable would decrease the number of items required.

The assumption that the distribution of $\eta$ is the same for the two mode samples is crucial for this kind of calibration, and is difficult to guarantee. It is satisfied if we have interpenetrating probability samples for the two modes in a single survey; then in principle we can imagine a mapping of responses from one mode to the other, through estimated values of $\alpha_{2}$ and $\beta_{2}$. We do not have to estimate the constructs themselves to do this. More rigorously, we can include $\alpha_{2}$ and $\beta_{2}$ as parameters in a model for all responses to a set of similar items.

In some developed countries, sampling frames for households and individuals appear to be moving in the direction of address registries and lists of persons. However, even when there is a common frame for (say) telephone and internet surveys, it is difficult to randomize respondents to data collection modes. The dependence of non-response on demographic variables may well be different for the modes. Moreover, the need to maximize response rates often dictates allowing respondents to choose. In principle, we might imagine that the distribution of $\eta$ might be shifted or tilted according to the "propensity" to choose one mode or the other. Having modeled this propensity in terms of explanatory variables, and having introduced one or two parameters for the dependence of the distribution of $\eta$ on the propensity, we could estimate the item parameters $a_{i}$ and $b_{i}$ from the respondents for the first data collection mode. The estimation of the mapping parameters $\alpha$ and $\beta$ would follow in the same manner as before.

In another kind of circumstance, we might use the two data collection modes in different groups of the population. In that case, the mode effect becomes part of the group effect; it cannot be disentangled from an underlying difference in the distribution of the construct.

Trying to compare measurements across cultures or other groups is different from the calibration problem, since randomizing participants to groups, to keep the distribution of the construct constant, is out of the question. The common wisdom is that to compare the mean of a construct from one group to another, the measuring items must have the same relationship with the construct in the two groups.

When there are several constructs, to compare the relationship among constructs from one group to another requires a kind of "measurement invariance" or equivalence for all items involved. There is a vast literature on crosscultural comparisons and measurement. For example, Johnson (1998) lists fifty-two terms for cross-cultural equivalence that have been introduced by authors in various disciplines.

The multi-group confirmatory factor analysis model is useful for continuous measurement items, and takes the form:

$$
Y_{k}^{g}=\tau^{g}+\Lambda^{g} \eta_{k}^{g}+e_{k}^{g}
$$

where $Y_{k}^{g}$ is the vector of observed responses to the items for respondent $k$ in group $g ; \Lambda^{g}$ is a matrix of slopes or "factor loadings"; the intercept vector $\tau^{g}$ indicates the expected value of $Y_{k}^{g}$ when $\eta_{k}^{g}=0$; and $e_{k}^{g}$ is a measurement error with 0 mean. Then $E\left(Y_{k}^{g}\right)=\tau^{g}+$ $\Lambda^{g} \kappa^{g}$, where $\kappa^{g}$ is the mean of the construct $\eta$ in group $g$. The variance-covariance matrix among the observed values $y_{k}^{g}$ can be expressed as $V\left(Y_{k}^{g}\right)=\Lambda^{g} \Phi^{g} \Lambda^{g^{\prime}}+\Theta^{g}$, where $\Phi^{g}$ is the covariance matrix of the latent constructs and $\Theta^{g}$ is the diagonal matrix of measurement error variances. See de Jong, Steenkamp and Fox (2007), Davidov (2008) and references therein.

The IRT version of the model can be defined in straightforward manner. Using the same parameter notation, in the case of binary items, we have

$$
P\left(Y_{k}^{g}=1 \mid \eta_{k}^{g}\right)=H\left(\tau^{g}+\Lambda^{g} \eta_{k}^{g}\right),
$$

and there is a natural extension to the ordinal case.
The model parameters are not identifiable unless some restrictions are made. In the multi-group confirmatory factor analysis model, many authors postulate a "marker" item for each construct, with a factor loading of 1 and an intercept of 0 for all groups, so that the mean of the construct is identified in each group. This is a very strong assumption. Alternatively, we might imagine choosing the units for the constructs so that they are marginally $\mathrm{N}(0,1)$ within group 1 . The parameters of the items (with sufficiently many items) are thus identified for group 1. If the variances and relationships of the path diagram are assumed to remain true in group 2 , then we can test whether the item parameters also remain the same, and if not, try to redesign the set of items to produce something closer to measurement invariance. On the other hand, if the item parameters are constrained to remain the same, we can test whether the underlying joint distribution of the constructs is also the same. However, formal rejection of the null hypothesis is difficult to interpret. Following Rensvold and Cheung (1998), Barrera Ceballos (2007) has carried out this kind of multi-group analysis for the data of the ITC Mexico survey
and the ITC Uruguay survey, replacing "health concern" in the model of Figure 2 by "social denormalization", or the extent to which the respondent perceives society to disapprove of tobacco use. (The other two constructs are warning label salience and intention to quit.) The relationships appear unexpectedly different in the two countries under the constraints of measurement item invariance, a finding which could be due either to real societal differences or to an imperfect correspondence between the items themselves (i.e., failure of the constraints). Admittedly, with very few constructs having multiple items, the ITC survey instrument was not designed for this kind of analysis,

Ultimately the relationships among the constructs are of paramount importance, along with the question of whether the relationships of the constructs can be said to be alike, though not necessarily identical, from group to group. This is so regardless of whether the marginal distributions of the constructs are the same, or whether measurement items have the same parameters from place to place, or mode to mode. Intuitively, the two kinds of restrictions of the previous paragraph seem too strong. A hierarchical approach of De Jong et al. (2007) offers a way forward.

If item $i$ has $C$ ordered response options, we can write

$$
\begin{aligned}
P\left(Y_{i k}^{g}\right. & \left.=c \mid \eta_{k}^{g}, b_{i}^{g}, \gamma_{i, c}^{g}, \gamma_{i, c-1}^{g}\right)=H\left(b_{i}^{g} \eta_{k}^{g}-\gamma_{i, c-1}^{g}\right) \\
& -H\left(b_{i}^{g} \eta_{k}^{g}-\gamma_{i, c}^{g}\right),
\end{aligned}
$$

$c=1, \ldots, C$. Here the factor loadings are replaced by the discrimination parameters $b$, and the intercepts are replaced by the thresholds $\gamma$. Instead of insisting that these parameters are independent of group label before proceeding, the approach is to model them with groupspecific random effects:

$$
\begin{aligned}
\gamma_{i, c}^{g} & =\gamma_{i, c}+e_{i, c}^{g}, e_{i, c}^{g} \sim N\left(0, \sigma_{\gamma_{i}}^{2}\right), \\
b_{i}^{g} & =b_{i}+r_{i}^{g}, r_{i}^{g} \sim N\left(0, \sigma_{b_{i}}^{2}\right) .
\end{aligned}
$$

The heterogeneity in the latent variable is modeled by a hierarchical structure:

$$
\begin{aligned}
& \eta_{k}^{g}=\kappa^{g}+v_{k}^{g}, v_{k}^{g} \sim N\left(0, \sigma_{g}^{2}\right), \\
& \kappa^{g} \sim N\left(\kappa, \xi^{2}\right) .
\end{aligned}
$$

With sufficiently many items, such a model is estimable, and can be fitted using Markov Chain Monte Carlo methods. The invariance tests of multi-group analysis can still be performed within this framework.

## 5. Discussion and conclusions

Again, the aim of analysis in the ITC context must be to discern the real response (or lack of response) to policy
change, separating it from the effects of data collection mode, differential non-response, external events, time-insample, culture, and language. It may not be necessary to distinguish among all of the confounders, but it is important to allow them to contribute to the model. In this paper we have not addressed external events, which can be modelled in an obvious way if recognized. We have not discussed modelling attrition or time-in-sample effects in detail, but in principle, each one of them can be regarded as part of the mix. Those who are retained from wave to wave of a survey might be regarded as a kind of cultural group. On the other hand, time-in-sample effects are a particular kind of failure of measurement invariance, over time rather than from one group to another. A comprehensive analysis would take account of these, and of other effects of culture, language and data collection mode.

It is by no means the case that the effects of policy would always be identifiable in a full model. But the chances increase if the design involves between country comparisons of longitudinal data, and the replication which comes from observing cohorts with different starting points.

A unifying thread of the two previous sections is the introduction of random effects as a device. The device of introducing random effects for countries and groups in key parameters is natural, and (for large group samples) conceptually compatible with traditional survey analysis, based on weighted estimating functions. There are some obstacles to practical implementation, arising from identifiability and estimability limitations, and the calculation of likelihood functions if more than a few random effects are entertained. At the same time, with increasing availability of numerical methods to handle such models, further research to adapt them to complex international surveys should be very fruitful.

## Acknowledgments

This work is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The ITC project is supported by grants from the National Cancer Institute of the United States (P50 CA11236) Roswell Park Transdisciplinary Tobacco Use Research Center and the Canadian Institutes of Health Research (57897). Thanks are extended to an anonymous reviewer for very helpful comments.

## References

Barrera Ceballos, J.A. (2007). Cross-national comparison of the impact of cigarette warning labels and social denormalization on intention to quit from the International Tobacco Control Survey. Research paper. Statistics and Actuarial Science, University of Waterloo.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

Bronner, K., and Kuijlen, T. (2007). The live or digital interviewer: A comparison between CASI, CAPI and CATI with respect to differences in response behaviour. International Journal of Market Research, 49, 167-190.

Cleland, J., and Verma, V. (1989). The World Fertility Survey: An appraisal of methodology. Journal of the American Statistical Association, 84, 756-767.

Cook, T., and Campbell, D. (1979). Quasi-Experimentation. Chicago: Rand McNally.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. Survey Research Methods, 2, 33-46.

De Jong, M.G., Steenkamp, J.-B.E.M. and Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical irt model. Journal of Consumer Research, 34, 260-278.
de Leeuw, E.D., and Hox, J. (2003). The use of meta-analysis in crossnational studies. In Cross-Cultural Survey Methods (Eds., J.A. Harkness, F.J.R. Van de Vijver and P. Ph. Mohler). Hoboken, NJ: Wiley, 329-346.

ESS (2008). European Social Survey. http://www.europeansocialsurvey. org/.

Fong, G.T., Cummings, K.M., Borland, R., Hastings, G., Hyland, A., Giovino, G.A., Hammond, D. and Thompson, M.E. (2006). The conceptual framework of the International Tobacco Control Policy Evaluation Project. Tobacco Control, 15(Supp 3): iii3-iii11.

Fong, G.T., Hyland, A., Borland, R., Hammond, D., Hastings, G., McNeill, A., Anderson, S., Cummings, K.M., Allwright, S., Mulcahy, M., Howell, F., Clancy, L., Thompson, M.E., Connolly, G. and Driezen, P. (2006). Reductions in tobacco smoke pollution and increases in support for smoke-free public places following the implementation of comprehensive smokefree workplace legislation in the Republic of Ireland: Findings from the ITC Ireland/UK Survey. Tobacco Control, 15(Supp. 3): iii51-iii58.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey populations: Their relationships and estimation. International Statistical Review, 54, 127-138.

Hammond, D., Fong, G.T., Borland, R., Cummings, K.M., McNeill, A. and Driezen, P. (2007). Text and graphic warnings on cigarette packages: Findings from the International Tobacco Control Four Country Study. American Journal of Preventive Medicine, 32, 210-217.

Hofstede, G. (1980). Culture's Consequences. International Differences in Work-Related Values. Beverly Hills, CA: Sage.

Hyland, A., Hassan, L., Higbee, C., Fong, G.T., Borland, R., Cummings, K.M., Thompson, M., Boudreau, C. and Hastings, G. (2008). The impact of smokefree legislation in Scotland: Results from the Scottish International Tobacco Control Policy Evaluation Project. In progress.

ITC (2008). Tobacco Labelling Resource Centre. http://www.igloo. org/tobacco_labelling. Accessed April 24, 2008.

Johnson, T.P., and Van de Vijver, F.J.R. (2003). Social desirability in cross-cultural research. In Cross-Cultural Survey Methods (Eds., J.A. Harkness, F.J.R. Van de Vijver and P. Ph. Mohler). Hoboken, NJ: Wiley, 195-204.

Johnson, T.P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In Cross-Cultural Survey Equivalence (Ed., J.A. Harkness). ZUMA-Nachrichten Spezial 3. Mannheim: ZUMA, 1-40.

Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. Journal of the American Statistical Association, 95, 271280.

Lu, I.R.R., Thomas, D.R. and Zumbo, B.D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. Structural Equation Modeling, 12, 263-277.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society, Series B, 60, 23-40.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. Sankhyā, series B, 61, 166-186.

Rensvold, R.B., and Cheung, G.W. (1998). Testing measurement models for factorial invariance: A systematic approach. Educational and Psychological Measurement, 58, 1017-1034.

Scott, A. (2006). Population-based case control studies. Survey Methodology, 32, 123-132.

Rosenbaum, P.R., and Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association, 79, 516542.

Ryder, A.G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S.J. and Bagby, R.M. (2008). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? Journal of Abnormal Psychology, 117, 300-313.

Skinner, C. (1989). Introduction to Part A. In Analysis of Complex Surveys (Eds., C. Skinner, D. Holt and T.M.F. Smith), Chichester: Wiley. 2.

Thompson, M.E., Boudreau, C. and Driezen, P. (2005). Incorporating time-in-sample in longitudinal survey models. Proceedings: Symposium 2005, Methodological Challenges for Future Information Needs. Session 12: Challenges in Using Data from Longitudinal Surveys. Statistics Canada.

Thompson, M.E., Fong, G.T., Hammond, D., Boudreau, C., Dreizen, P., Hyland, A., Borland, R., Cummings, K.M., Hastings, G.B., Siahpush, M., Mackintosh, A.M. and Laux, F.L. (2006). Methods of the International Tobacco Control (ITC) Four Country Survey. Tobacco Control, 15(Supp 3): iii12-iii18.

Thrasher, J., Quah, A., Borland, R., Awang, R., Sirirassamee, B., Boado, M., Miller, K., Watts, A. and Dorantes, A. (2008). Ensuring valid cross-cultural comparisons in survey research on tobacco: Development, implementation, and results from a transnational cognitive interviewing study. In progress.

Verma, V., Scott, C. and O’Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. Journal of the Royal Statistical Society, Series A, 143, 431-473.
Waksberg, J. (1978). Sampling methods for random digit dialing. Journal of the American Statistical Association, 73, 40-46.

WHO (2008). WHO Report on the Global Tobacco Epidemic, 2008. http://www.who.int/tobacco/mpower/mpower_report_full_2008.pdf Accessed April 14, 2008.
Wichers, B., and Zengerink, E. (2006). It's the culture, stupid! A cross-cultural comparison of data collection methods. Panel Research 2006, Part 4/The respondent - Cross cultural insights, ESOMAR.

# Using behavior coding to evaluate the effectiveness of dependent interviewing 

Joanne Pascale and Alice McGee ${ }^{1}$


#### Abstract

Dependent interviewing (DI) is used in many longitudinal surveys to "feed forward" data from one wave to the next. Though it is a promising technique which has been demonstrated to enhance data quality in certain respects, relatively little is known about how it is actually administered in the field. This research seeks to address this issue through behavior coding. Various styles of DI were employed in the English Longitudinal Study of Ageing (ELSA) in January, 2006, and recordings were made of pilot field interviews. These recordings were analysed to determine whether the questions (particularly the DI aspects) were administered appropriately and to explore the respondent's reaction to the fed-forward data. Of particular interest was whether respondents confirmed or challenged the previously-reported information, whether the prior wave data came into play when respondents were providing their current-wave answers, and how any discrepancies were negotiated by the interviewer and respondent. Also of interest was to examine the effectiveness of various styles of DI. For example, in some cases the prior wave data was brought forward and respondents were asked to explicitly confirm it; in other cases the previous data was read and respondents were asked if the situation was still the same. Results indicate varying levels of compliance in terms of initial question-reading, and suggest that some styles of DI may be more effective than others.


Key Words: Dependent interviewing; Longitudinal surveys; Panel surveys; Behavior coding.

## 1. Introduction

In recent years there has been increased interest in and use of "dependent interviewing" (or DI) in longitudinal surveys. DI (also known as "previously reported data" or PRD) is a technique whereby data collected from one wave are carried forward into the next wave in order to tailor question wording and skip patterns. For example, if at Wave 1 a respondent reported working for Employer X, the Wave 2 DI question would read: "Last time you said you worked for Employer X. Are you still working for Employer X?" This is in contrast to an "independent" (that is, non-DI) method whereby at Wave 2 the respondent would simply be asked "from scratch" if he/she was working, and the name of the employer. A related implementation of DI is to route respondents around detailed questions if the circumstances from one wave to another have not changed. For example, a detailed set of questions about Employer X may be asked in Wave 1 (such as the industry, number of employees, etc.), and if at Wave 2 the respondent reports they are still working for the same employer, those details need not be collected a second time.

The proliferation of automated surveys has contributed to the increased use of DI, since the technique can be difficult and cumbersome to implement in a paper/pencil questionnaire. Another factor contributing to the interest in DI is its potential to enhance data quality in a number of ways. Generally DI can make for a smoother, smarter, more efficient interview by reminding respondents of their
previous answers and allowing them to simply report whether anything has changed since then. Rigorous research evidence demonstrating this potential is beginning to emerge. For example, there is consistent evidence that DI reduces spurious change, particularly in employment characteristics (Polivka and Rothgeb 1993; Jäckle and Lynn 2004). Another source of measurement error that has consistently plagued panel surveys is "seam bias."

The "seam" of a panel survey is the point where one wave is joined with the next wave. For example, in a panel survey with annual waves the seam is between December of one year and January of the following year. Seam bias occurs when more transitions (e.g.: from employment to unemployment) are observed from December to January than for any of the non-seam month pairs (e.g.: February to March, April to May). There is strong evidence that DI significantly reduces (though does not eliminate) this seam bias (Moore, Bates, Pascale and Okon 2006). In terms of respondents' receptivity to DI, there is qualitative evidence that respondents want and expect it (Pascale and Mayer 2004). In the summer of 2006 a major conference was organized to assess the "state of the art" of research on longitudinal studies and several papers demonstrated specific benefits of DI. An edited monograph book of selected papers is to be published by John Wiley and Sons in 2008 (http://www.iser.essex.ac.uk/ulsc/mols2006/).

What the literature seems to lack up to now, however, is evidence of how DI is actually implemented in the field. The current research set out to address this gap. In particular

[^1]we use behavior coding to examine whether interviewers read questions as worded, focusing especially on the dependent words and phrases within the questions, and we examine respondents' reactions to the dependent phrases that is, whether they affirm or dispute the previouslycollected data, and whether providing this information seems to help or hinder the reporting task. Finally we examine whether these behaviors seem to vary at all by "style" of DI - that is, the particular way that the previously-collected information is fed back to the respondent. The vehicle we use for this research is the English Longitudinal Study of Ageing (ELSA), carried out by the National Centre for Social Research (or "NatCen") in collaboration with University College London and the Institute of Fiscal Studies.

## 2. Methods

### 2.1 ELSA: The survey vehicle

ELSA is a study of people aged 50 and over and their younger partners. The study explores the dynamics of health and disability, family structure, public program participation, economic circumstances, and retirement. The ELSA sample was drawn from households that had previously responded to five years of the Health Survey for England (HSE) between 1998 and 2003. The first ELSA wave was administered in 2002 with 12,100 respondents, and followup interviews have been conducted every two years to measure changes in health and social and economic circumstances. Dependent interviewing was embedded in the Wave 2 instrument but due to budget and schedule constraints, little evaluation was done prior to its implementation. Analysis of Wave 2 data, however, raised some concerns about the effect of DI. For example, roughly $20 \%$ of respondents who reported high blood pressure at Wave 1 reported that they no longer had the condition at Wave 2. Due in part to this finding, the current research project was undertaken to generally assess the implementation of DI techniques in the field. Behavior coding was chosen as the evaluation method in order to carefully assess interviewerrespondent interactions, and to measure the extent to which the questions were being administered as written.

### 2.2 Field interviewing and recording

The pilot phase of Wave 3 ELSA was conducted over a 4 week period in January, 2006. Altogether 17 NatCen field interviewers from different areas around the United Kingdom conducted 123 individual interviews. The vast majority of the interviews (106) were conducted with individuals who had been interviewed in the prior ELSA wave, while 17 interviews were conducted with members of
a refreshment sample who were new to ELSA but who had been interviewed in the HSE. Most of the analysis in this paper pertains to those individuals interviewed in the prior ELSA wave. However, for two items in the demographics section - LIVE (whether a household member still lives at the residence) and DOB (household member's date of birth) - both the prior-ELSA-wave and the refresher sample were included in the analysis since those items included data fed forward from the HSE interviews. That is, these two items included DI even for those refresher cases not interviewed in ELSA in the prior wave. All interviews were conducted face-to-face using a computer-assisted personal instrument (CAPI). The questionnaire included questions on a number of topics: household and individual demographics, health status, income and assets. Interviews were recorded using Computer Audio Recorded Interviewing (CARI), a software application that allows field interviews to be recorded directly onto computer laptops as digital sound files. A consent question asking respondents for permission to record the interview was embedded into the beginning of the questionnaire, and if respondents did not consent the recorder was not switched on. Furthermore, in some cases the sound files were corrupted and therefore could not be coded. Among the 123 individual interviews, 104 were coded, and among the 106 prior-ELSA-wave interviews, a total of 87 individual recordings were coded. In both cases the majority of the interview losses stemmed from nonconsent (versus corrupt sound files).

### 2.3 Dependent interviewing question wording

Dependent interviewing was embedded in the instrument across three different topic areas: demographics, health conditions and vehicle ownership (see Figure 1). In the health condition section there were three broad categories of illnesses: eye, cardiovascular disease (CVD), and chronic conditions. Within each of these broad categories there were multiple specific illnesses asked about. For example under eye conditions there were four illnesses (such as glaucoma and cataracts). Items 4 and 5 were repeated for each illness or condition the respondent had reported in the prior wave.

Five different styles of DI were used across these three topic areas, but as was mentioned earlier, no particular research guided those design decisions. As Figure 1 indicates, each of the six items employed a slightly different style of DI. The first two items in the demographics section (LIVE, DOB) do provide previously-reported data but do not explicitly mention having gathered this data in the previous interview. Rather, the past data is simply presented and the respondent is asked to verify it. The third item (CHILD) explicitly states that the data was collected last time and the respondent is asked if the information is correct. Unlike the demographics questions, the health
questions were separated into two distinct items, which appeared on two different screens. The first (LAST-EYE) was simply a statement, informing the respondent of a particular illness they reported during the previous interview. It was meant to be read as a statement and the respondent was not asked or expected to provide a response to this statement; rather the interviewer was meant to press the "enter" key in order for the second of the two-part series to appear. The second item (STILL-EYE) then asked whether the respondent still had the illness or condition. And finally for vehicle ownership the routine was somewhat similar to the health conditions questions; first a statement was read that informed the respondent of what they reported last time, and then a question was asked to determine if that condition still existed (that is: do you still own the vehicle?). The difference was that for the vehicle item the statement on the past condition and the question ("still have it") were wrapped into one single item, while in the health section there were two distinct items - the statement and then the "still?" question.

[^2]Figure 1 Question wording of items using dependent interviewing

### 2.4 Behavior coding

In order to develop the code frame for behavior coding, we first listened to several recordings to get a general feel for the flow of the interview, the frequency and nature of non-standard interviewer behavior, and respondent's reactions to the questions. We determined that the "first exchange" - the interviewer's initial reading of the question and the respondent's first utterance in response to that - was sufficiently rich for analysis and thus developed a code frame to capture only these behaviors, as well as a final outcome. Within these three behaviors (interviewer's
initial question-reading, respondent's initial response, and outcome), we started out with a fairly standard code frame and adapted it based on the content of the recordings and our particular interest in learning about the functioning of the feed-forward phrases embedded within the questions (see Figure 2). For interviewer behavior we used three main code categories: (1) question was read as worded or with only a minor change that did not change the meaning of the question (2) question was read with a "major change" that changed or could change the meaning of the question and (3) the question was omitted. Within the major change code we developed two DI-specific codes. On the recordings it was rather common to hear interviewers changing a statement into a question. For example, in the health section the statement: "Our records show that last time you reported X condition" became a question because interviewers often added "Is that right?" Or, in some cases, interviewers used an intonation and a pause to turn the statement into a question - for example "Our records show that last time you reported X condition?" followed by a pause, waiting for an answer from the respondent. We should note that because coders were working directly from recordings, versus transcripts, they were able to make a judgment regarding the use of intonation to convey either a question or a statement. In other cases a question became a statement. For example in the demographics section the question "Does NAME still live here?" was modified to "And NAME still lives here." (with no intonation indicating a question mark). Since these were the most frequentlyobserved problems we created dedicated codes for them.

In total there were 7 coders, drawn from both the survey methods and the operations units. Researchers conducted a half-day training which lasted 4 hours. The training covered the basic concepts of behavior coding, along with the studyspecific codes and how to apply them. The majority of training time was devoted to coding hypothetical examples of respondent-interviewer interactions and then discussing and comparing individual judgments and the rationale for those judgments in an attempt to apply codes consistently across coders. However, no formal reliability measures were implemented.

Respondent codes were fairly standard, again with the exception of DI-specific codes. An "adequate" code meant that the respondent's initial utterance fit into one of the response categories. We adapted this code to capture whether respondents affirmed or disputed the fed-forward data. There were also codes for a request for clarification and a rereading of the question, and a general "inadequate" code, meaning the respondent's answer did not fit into any of the given response categories. Outcome codes were simply "adequate" and "inadequate."

```
A. Interviewer Codes
S: Standard; read as worded or with a minor change that did not change the meaning
MC1: Fed-forward statement was read as a question (e.g.: "Last time you told us you had high blood pressure. Is that correct?")
MC2: Fed-forward question was read as a statement (e.g.: "And your date of birth was \(25^{\text {th }}\) May 1933.")
MC3: Any other change that did or could change the meaning of the question
O: Omission
I/O: Recording was inaudible or the behavior does not fit into one of the above codes
B. Respondent Codes
AA: Adequate; acknowledged or did not dispute the fed-forward data
AD : Adequate; disputed or challenged the fedforward data
CL: Request for clarification
IA: Inadequate answer or elaboration
DK: Don't know
R: Refused
I/O: Recording was inaudible or the behavior does not fit into one of the above codes
C. Outcome Codes
AA: Adequate; final response fit one of the given response categories
IA: Inadequate; final response did not fit any of the response categories
DK: Don't know
R: Refused
I/O: Recording was inaudible or the behavior does not fit into one of the above codes
```

Figure 2 Behavior codes

Table 1A
Interviewer behavior for demographic items

| Item | Base <br> $(\mathbf{n})$ | Interviewer Behavior Code (in percent) <br> Read as <br> worded | Qus statement <br> ather major <br> change | Omitted | Other |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| LIVE: Does NAME still live here? | 120 | 40 | 33 | 4 | 18 | 5 |
| DOB: Can I just check, is NAME's <br> date of birth [DOB]? | 107 | 57 | 37 | 1 | 1 | 4 |
| CHILD: Our records show that when | 84 | 79 | 8 | 11 | 0 | 2 |

As was predicted from our earlier (unsystematic) listening of the recordings, for the most part when interviewers diverged from the script they turned the question into a statement (e.g.: "Is NAME's date of birth January 1?" would become "And NAME's date of birth is January 1."). This behavior occurred 33-37\% of the time for the first two items (LIVE and DOB) and only $8 \%$ of the time for the last item (CHILD). This may not be too surprising considering the nature of the items. Answers to the first two items may seem obvious - particularly at Wave 3 - and interviewers may have been somewhat reluctant to ask a question with an obvious answer. Indeed LIVE was omitted altogether $18 \%$ of the time, and this could be because the interviewer was talking to the person referenced in the question. The third item, on the other hand, asks about someone else in the household (a child), the information is rather specific (name and date of birth) and the actual question ("are these details correct?") may not seem to have an obvious answer. That is, it may seem like a more "legitimate" question to ask than asking a person, in what appears to be their home, "Do you still live here?" This could explain why this last item was read as worded so frequently - $79 \%$ of the time.

Turning to respondent behavior, on the whole respondents provided a codeable answer straightaway more than $80 \%$ of the time (see Table 1B). (Note that in some cases the base for respondent behavior on any given item is smaller than the base for interviewer behavior for that same item. This is due to a combination of interviewers omitting the item (in which case there was no respondent behavior to code) and missing data.) They rarely disputed the fedforward data (up to only $5 \%$ of the time), and most of the disputes stemmed from keying errors in the name or date of the birth previously recorded.

### 3.2 Health items

As noted above, the health questions were asked in two parts. First a statement about the condition reported during the prior wave was read, and then a question was asked to determine whether the condition still existed. Overall levels of "exact reading" of these items were moderate - ranging from $41-76 \%$ but generally in the low 60s (see Table 2A). When interviewers diverged from the script they tended to turn the statement into a question ( $20-38 \%$ of the time) by adding something along the lines of "Is that correct?" to the end of the statement. Interviewers would then often omit the actual question "Do you still have it?" altogether - 13-18\% of the time. The implications are important here, because it means the respondent is getting a fundamentally different question, specifically "Is it correct that you reported this condition last time?" versus "Do you still have this condition now?"

Another problem was when the actual question "Do you still [have condition X]?" was read, interviewers often read it as a statement rather than a question: "And you still have it." - 3-16\% of the time. This has serious implications for data quality as well, because the respondent is not being given the opportunity to think about whether they really do still have the condition; they are just being told they do.

Regarding respondent behavior (Table 2B), there were fairly high levels of adequate behaviour - over $90 \%$ for both cardiovascular and chronic conditions - and $72 \%$ for eye conditions (however the base here was only 21 cases). Respondents disputed prior wave data for a variety of reasons. Some said they used to have the condition but no longer do, and this is essentially how the questionnaire was expected to operate. But in other cases the fed-forward data were problematic; respondents either denied that they'd reported the condition at the prior wave, or they disagreed with the characterization of the illness. For example, in one case an illness was recorded as cancer in the prior wave and when asked about it in the next wave the respondent said it wasn't cancer. He wasn't sure what the diagnosis was but said it was not cancer. In another case a respondent reported memory impairment at the prior wave but this particular condition was grouped in with other related illnesses in the instrument ("dementia, senility or memory impairment"). When the DI question appeared on the screen the interviewer only read "dementia" and the respondent refuted it. Only when the interviewer went back and read the full question, with all three conditions, did the respondent affirm that he had a memory impairment. And finally, in one case the presence of other household members seemed to be an issue. For example, when the respondent was told he'd reported a certain chronic condition at the prior wave he asked "Did I?" and his wife said "yes."

### 3.3 Vehicle item

The vehicle item was similar to the health items - first providing a statement about what was recorded in the prior wave and then asking a question about whether the situation is still the same. A key difference, however, was that rather than presenting the statement and question as two distinct items on two different screens, they were rolled into one item. Across all items in the questionnaire the vehicle item had the highest level of interviewers reading the question as worded at $82 \%$ (see Table 3A). The problems identified in the health section - interviewers turning the statement into a question, or the question into a statement, or omitting the question - did not turn up very often here, perhaps because the style of DI was different. Specifically, interviewers did not have to have read a statement about the prior report but could move directly into the question: "Do you still have this vehicle?" By not displaying the statement on the prior
wave data as a distinct item, interviewers may have been less tempted to turn that statement into a question by asking, for example, "Is it correct that you reported this vehicle last time?" The result was that the intended question - whether the vehicle was still owned - was being asked, rather than an unintended question ("Did you report owning this vehicle last time?"). However, among the non-standard behaviors there were still several instances of interviewers ( $8 \%$ of the time) turning the question into a statement: "And you still own xx vehicle." This could be a result of interviewers
having seen the vehicle in question on their way to the doorstep.

Respondent behavior here was similar to the health section. Respondents provided a codeable answer straightaway $80 \%$ of the time (see Table 3B). They rarely disputed the fed-forward data ( $6 \%$ of the time), and reasons were mixed. One stemmed from keying errors in the fed-forward registration information, and two were based on real change (in one case the respondent had a different car; in the other case the respondent had given up driving).

Table 1B
Respondent behavior for demographic items

| Item | Base <br> (n) |  | Respondent behavior code (in percent) <br> Adequate; <br> affirmed fed- <br> forward data |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Adequate; <br> disputed fed- <br> forward data | Clarification |  |  |
| Inadequate |  |  |  |  | Other

Table 2A
Interviewer behavior for health items

| Item | Base <br> (n) | Interviewer behavior code (in percent) |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Read as worded | Question read as statement | Statement read as question | Other major change | Omitted | Other |
| LAST-EYE: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition] | 21 | 62 | na | 38 | 0 | 0 | 0 |
| STILL-EYE: Do you still have [condition] | 19 | 63 | 16 | na | 5 | 16 | 0 |
| LAST-CVD: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition] | 100 | 63 | na | 20 | 17 | 0 | 0 |
| STILL-CVD: Do you still have [condition] | 79 | 76 | 3 | na | 8 | 13 | 1 |
| LAST-CHRON: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition] | 59 | 41 | na | 34 | 17 | 5 | 3 |
| STILL-CHRON: Do you still have [condition] | 51 | 61 | 14 | na | 2 | 18 | 6 |

Table 2B
Respondent behavior for health items

| Item | Base (n) | Respondent behavior code (in percent) |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Adequate; affirmed fedforward data | Adequate; disputed fedforward data | Adequate* | Inadequate | Clarification | Other |
| LAST-EYE: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition] | 21 | 62 | 10 | [72] | 10 | 5 | 15 |
| STILL-EYE: Do you still have [condition] | 16 | na | na | 94 | 0 | 0 | 6 |
| LAST-CVD: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition] | 100 | 87 | 5 | [93] | 4 | 0 | 4 |
| STILL-CVD: Do you still have [condition] | 67 | na | na | 69 | 24 | 0 | 8 |
| LAST-CHRON: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition] | 53 | 85 | 4 | [89] | 2 | 0 | 9 |
| STILL-CHRON: Do you still have [condition] | 35 | na | na | 89 | 6 | 0 | 6 |

* For "LAST-XX" items this column shows the sum of "Adequate; affirmed fed-forward data" and "Adequate; disputed fed-forward data"

Table 3A
Interviewer behavior for vehicle item

| Item | Base <br> (n) | Read as <br> worded | Interviewer behavior code (in percent) <br> Question read <br> as statement | Other <br> major <br> change | Omitted |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |$\quad$ Other

Table 3B
Respondent behavior for vehicle item

| Item | Base <br> (n) | Respondent behavior code (in percent) <br> Adequate; <br> affirmed fed- <br> forward data | Adequate; <br> disputed fed- <br> forward data | Inadequate | Clarification | Other |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| VEHICLE: Last time we saw you, you told us <br> that you were the main user of a [MAKE OF | 49 | 74 | 6 | 10 | 4 | 6 |
| VEHICLE], with a [LETTER] registration. Do <br> you still have that vehicle? |  |  |  |  |  |  |

### 3.4 Interviewer-respondent interaction

In addition to item-specific analysis we examined the relationship between interviewer and respondent behavior across all items. We found that whether interviewers read questions as worded, read questions as statements, or read statements as questions, respondents provided an adequate (and affirmative) answer $87.5 \%$ of the time. For other behaviors, cell sizes were too small to conduct meaningful analysis.

## 4. Summary and recommendations

The extent to which interviewers adhered to the standardized script varied quite a bit - questions were read as worded $40-82 \%$ of the time, depending on the particular item. When interviewers diverged from the script, the way they changed the wording varied by topic area and style of DI which, unfortunately, were confounded because each item had a unique style of DI. In the demographics and vehicle items, for the most part interviewers changed the question into a statement ("Does NAME still live here?" became "And NAME still lives here.") In the health section interviewers read statements about what was reported last time as questions. Rather than simply reading the statement "Last time you reported X condition" interviewers would add "Is that correct?" (which is an ambiguous question) and often omitted the question "Do you still have condition $x$ ?" The result was that often the intended question - to determine whether the condition still exists - was obscured or omitted.

For the most part respondents provided codeable answers on the first exchange $72-94 \%$ of the time. It was fairly uncommon for respondents to dispute the fed-forward data ( $0-10 \%$ of the time) but when they did it was for a variety of reasons. Some confirmed the prior wave report but said they no longer have the condition. Some denied the prior wave report, and some disagreed with the details of the fedforward data. Note that this first scenario is what we expect to happen in the instrument so it is actually a misnomer to say the respondent "disputed" the earlier report. Respondents here are not disputing what they said earlier, but rather they are confirming their earlier report and then reporting change. However, when the code frame was developed we heard very few instances of respondents disputing the prior data at all; the majority of cases were respondents simply agreeing with the fed-forward data. We therefore failed to recognize that it would have been valuable to create separate codes for agreeing to the fed-forward data and reporting real change versus actually disputing the prior report. Even with the full dataset, however, the frequency with which respondents did not simply agree to the prior wave data was too
low for a rich analysis, and a larger dataset would be needed to address this issue.

In terms of recommendations, these findings strongly suggest that questionnaire designers should avoid providing statements of prior wave data without an actual question, because interviewers are too tempted to turn these statements into questions, which obscures the question on whether the prior wave situation still exists. If it is important to confirm or verify information reported in a prior wave, this should be done explicitly by adding a discrete question to the statement, such as: "Last time I recorded that you had condition X. Is that correct?" Subsequent questions could then be asked to determine whether the condition still exists. Separating the two concepts in this way would convey to the respondent that there are two distinct issues: one is whether the prior report was recorded accurately, and the other is whether the condition still exists. If researchers do not have a rationale for needing to confirm the accuracy of previously-recorded data, a more efficient approach would be to ask: "Last time I recorded that you had condition X. Do you still have condition X?"

Our findings from the health conditions section suggest that for certain topic areas it is important to feed back prior wave data in the respondent's own words as much as possible. When respondents' descriptions of their illnesses was obscured by either the instrument or the interviewer grouping the illness with other conditions, respondents no longer recognized the illness they originally reported.

Finally our findings suggest a more general recommenddation that the style of DI should be carefully tailored depending on the particular item. For example, for topics unlikely to change from one wave to the next (such as date of birth), avoid re-asking questions because interviewers often read them as statements or omit them altogether. For these topics it may be more effective to either explicitly verify the accuracy of the earlier report (as suggested above), or to avoid bringing back the information at all. A hybrid-type approach for a study with several waves would be to verify the accuracy of previously-recorded data in wave 2 and then accept the data as correct and avoid reaffirming it in all later waves.

## Acknowledgements

The authors thank Carli Lessof for her ongoing support of survey methods research, and Jeff Moore for reviews of earlier drafts of this paper. We also thank Hayley Cheshire, Michelle Gray, Fiona Andrews, Lizzie Hacker, Colin Miceli, Marian Bolden, Audrey Hale and Laura Street for their assistance with data collection and processing. This research was funded by the National Centre for Social Research, and
funding for the ELSA study is provided by the National Institute of Ageing (in the United States), and a consortium of UK government departments coordinated by the Office for National Statistics.

This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on methodological issues are those of the authors and not necessarily those of the US Census Bureau or the National Centre for Social Research.

## References

Jäckle, A., and Lynn, P. (2004). Dependent Interviewing and Seam Effects in Work History Data. Working Paper \#2005-24 of the Institute for Social and Economic Research, Colchester, UK: University of Essex.

Moore, J.C., Bates, N., Pascale, J. and Okon, A. (2006). Tackling Seam Bias Through Questionnaire Design. Invited paper presented at the conference on Methodology on Longitudinal Surveys, July 12-14, Essex, United Kingdom.

Pascale, J., and Mayer, T. (2004). Alternative methods for exploring confidentiality issues related to dependent interviewing. Journal of Official Statistics, 20, 2, 357-377.

Polivka, A., and Rothgeb, J. (1993). Redesigning the CPS Questionnaire. Monthly Labor Review, September, 10-28.

# Imputation for nonmonotone last-value-dependent nonrespondents in longitudinal surveys 

Jing Xu, Jun Shao, Mari Palta and Lin Wang ${ }^{1}$


#### Abstract

In longitudinal surveys nonresponse often occurs in a pattern that is not monotone. We consider estimation of timedependent means under the assumption that the nonresponse mechanism is last-value-dependent. Since the last value itself may be missing when nonresponse is nonmonotone, the nonresponse mechanism under consideration is nonignorable. We propose an imputation method by first deriving some regression imputation models according to the nonresponse mechanism and then applying nonparametric regression imputation. We assume that the longitudinal data follow a Markov chain with finite second-order moments. No other assumption is imposed on the joint distribution of longitudinal data and their nonresponse indicators. A bootstrap method is applied for variance estimation. Some simulation results and an example concerning the Current Employment Survey are presented.


Key Words: Bootstrap; Nonmonotone missingness; Last-value-dependent; Nonignorable nonresponse; Nonparametric regression.

## 1. Introduction

A survey is longitudinal if data are collected from every sampled unit at multiple time points. For example, in the Current Employment Survey (CES), commonly known as the payroll survey conducted by the U.S. Bureau of Labor Statistics, data are obtained from establishments on a monthly basis by mail, telephone, FAX, and electronic data entry (Butani, Harter and Wolter 1997). Other examples include the Survey of Income and Program Participation (SIPP) and many economic surveys conducted by the U.S. Census Bureau. Nonresponse occurs in longitudinal studies. We assume that every sampled unit responds at baseline (the first time point). Nonresponse is monotone if a unit not responding at some time does not return to the survey. Nonmonotone nonresponse, however, often occurs in surveys such as the CES and SIPP and entails a wider variety of nonresponse patterns.

Let $y_{1}, \ldots, y_{T}$ be the values of a variable from a sample unit, where $T$ is the total number of time points, and $\delta_{1}, \ldots, \delta_{T}$ be the response indicators ( $\delta_{t}=1$ if $y_{t}$ is a respondent and $\delta_{t}=0$ if $y_{t}$ is a nonrespondent). Nonresponse is completely at random if $\left(\delta_{1}, \ldots, \delta_{T}\right)$ is statistically independent of $\left(y_{1}, \ldots, y_{T}\right)$, which rarely occurs in surveys. A more realistic assumption is that nonresponse at time point $t$ depends on observed or unobserved past values $y_{1}, \ldots, y_{t-1}$. In this paper, we focus on a stronger assumption, the last-value-dependent nonresponse mechanism, i.e., nonresponse of $y_{t}$ depends on the last value $y_{t-1}$ (observed or unobserved). The last-value-dependent nonresponse mechanism is assumed in many economic surveys (e.g., the CES; see Butani, Harter
and Wolter 1997). If nonresponse is also monotone, then either $y_{t}$ is a nonrespondent with certainty or $y_{t-1}$ is observed. This is a special case of what is referred to as ignorable missingness (Little and Rubin 1987). For nonmonotone nonresponse, however, last-value-dependent nonresponse is nonignorable, as whether $y_{t}$ is a respondent depends on $y_{t-1}$ that may be a nonrespondent.

Existing methods for handling nonmonotone nonresponse can be briefly described as follows. Under parametric modeling, methods such as the maximum likelihood, multiple imputation, or Bayesian analysis can be applied (e.g., Troxel, Harrington and Lipsitz 1998; Troxel, Lipsitz and Harrington 1998; Schafer 1997), if a suitable parametric model for the joint distribution of $\left(y_{1}, \ldots, y_{T}\right)$ and $\left(\delta_{1}, \ldots, \delta_{T}\right)$ can be found. The validity of these methods, however, depends on whether parametric models are correctly specified. A simple linear regression imputation method (see, e.g., Butani etal. 1997) imputes a nonrespondent $y_{t}$ by the predicted value under a fitted linear regression model between $y_{t}$ and $y_{t-1}$, where the regression model is fitted using data from sampled units with both $y_{t}$ and $y_{t-1}$ observed and the prediction is made using the predictor being either the observed $y_{t-1}$ or a previously imputed value of a nonrespondent $y_{t-1}$. Under the nonmonotone nonresponse mechanism (1), however, it can be shown that simple linear regression imputation is biased even if the linear regression model between $y_{t}$ and $y_{t-1}$ is correct. The bias is mainly caused by the erroneous way of using imputed $y_{t-1}$ values to impute missing $y_{t}$ values. A censoring approach creates a dataset with monotone nonresponse by discarding all respondents from a sampled unit after its first missing $y$-value. Methods

[^3]appropriate for monotone nonresponse (Paik 1997; Robins, Rotnitzky and Zhao 1995; Troxel, Lipsitz and Brennam 1997) can then be applied to the reduced dataset. Although this approach produces correct estimators, it is not efficient when $T$ is not small, since many respondents are discarded.

The purpose of this article is to propose an imputation method for longitudinal surveys with nonmonotone nonresponse and the last-value-dependent nonresponse mechanism (1). Imputation is commonly used to compensate for nonresponse in survey problems (Kalton and Kasprzyk 1986; Rubin 1987). Once all nonrespondents are imputed, estimates of parameters (such as the mean of $y_{t}$ ) are computed using standard methods by treating imputed values as observations. Our proposed imputation method produces approximately unbiased and consistent estimators for the means of $y_{1}, \ldots, y_{T}$.

The rest of this paper is organized as follows. In Section 2, we describe our assumptions. Section 3 describes the imputation process. Some properties of the resulting estimators of population means are discussed in Section 4, together with the proposal of a bootstrap procedure for variance estimation. Section 5 contains some simulation results. An example related to the CES is presented in Section 6. The last section contains a summary.

## 2. Assumption and imputation model

Let $P$ be a finite population indexed by $i=1, \ldots, N$, and let $S$ be a sample of size $n$ taken from $P$ according to some sampling design. According to the sampling design, survey weights $w_{i}, i \in S$, are constructed so that for any set of values $\left\{z_{i}: i \in P\right\}$,

$$
E_{s}\left(\sum_{i \in S} w_{i} z_{i}\right)=\sum_{i=1}^{N} z_{i}
$$

where $E_{s}$ is the expectation with respect to $S$. For each unit $i \in P,\left(y_{1, i}, \ldots, y_{T, i}\right)$ is a vector of items of interest obtained at time points $t=1, \ldots, T$. When nonresponse is present, each unit also has the vector $\left(\delta_{1, i}, \ldots, \delta_{T, i}\right)$ of response indicators. For simplicity, we may omit the index $i$ in our discussion.

We adopt a model-assisted approach by assuming that the vector $\left(y_{1, i}, \ldots, y_{T, i}, \delta_{1, i}, \ldots, \delta_{T, i}\right)$ 's are independent and identically distributed (i.i.d.) from a superpopulation. The i.i.d. assumption can be relaxed by dividing $P$ into several sub-populations (called imputation classes) so that the i.i.d. assumption approximately holds within each imputation class. Imputation classes are usually constructed using a categorical variable whose values are observed for all sampled units; for example, under stratified sampling, strata or unions of strata are often used as imputation classes. Each imputation class should contain a large number of sampled
units. When there are many strata of small sizes, imputation classes are often obtained through poststratification (Valliant 1993) and/or combining small strata.

Once imputation classes are constructed, imputation is done within each imputation class. Thus, for simplicity, from now on we assume that there is only one imputation class.

Under the last-value-dependent nonresponse mechanism,

$$
\begin{align*}
& P\left(\delta_{t}=1 \mid y_{1}, \ldots, y_{T}, \delta_{1}, \ldots, \delta_{t-1}, \delta_{t+1}, \ldots, \delta_{T}\right) \\
&=P\left(\delta_{t}=1 \mid y_{t-1}\right), \quad t=2, \ldots, T \tag{1}
\end{align*}
$$

We do not make any other assumption on $P\left(\delta_{t}=1 \mid y_{t-1}\right)$. When there is no nonresponse, we assume that $\left(y_{1}, \ldots, y_{T}\right)$ is a Markov chain, i.e.,

$$
\begin{equation*}
L\left(y_{t} \mid y_{1}, \ldots, y_{T}\right)=L\left(y_{t} \mid y_{t-1}\right), \quad t=2, \ldots, T \tag{2}
\end{equation*}
$$

where $L(\xi \mid \zeta)$ denotes the conditional distribution of $\xi$ given $\zeta$. We do not make any other assumption on $L\left(y_{t} \mid y_{t-1}\right)$ except that $y_{t}$ has a finite second-order moment. In many economic surveys, the following assumption stronger than (2) is assumed:

$$
\begin{equation*}
y_{t}=\beta y_{t-1}+\sqrt{\left|y_{t-1}\right|} \varepsilon_{t}, \quad t=2, \ldots, T \tag{3}
\end{equation*}
$$

where $\beta$ is an unknown parameter, $\varepsilon_{t}$ 's are independent of $y_{t}$ 's, $\varepsilon_{1}=0$, and $\varepsilon_{2}, \ldots, \varepsilon_{T}$ have mean 0 and a common variance (e.g., the CES data; see Butani et al. 1997). Under (3), the best linear unbiased estimator of $\beta$ is the well known ratio estimator.

To consider asymptotics, we adopt the frame work in Krewski and Rao (1981) and Bickel and Freedman (1984). We assume that the finite population $P$ is a member of a sequence of finite populations indexed by $v$. All limiting processes are understood to be as $v \rightarrow \infty$. As $v \rightarrow \infty$, the population size $N$ and the sample size $n$ increase to infinity. In sample surveys, the following regularity conditions on $w_{i}$ 's are typically imposed:

$$
\begin{equation*}
n \max _{i \in P} w_{i} \leq b_{0} N \quad \text { and } \quad n \operatorname{Var}_{s}\left(\sum_{i \in S} w_{i}\right) \leq b_{1} N^{2} \tag{4}
\end{equation*}
$$

where $b_{0}$ and $b_{1}$ are some positive constants and $\operatorname{Var}_{s}$ is the variance with respect to sampling. The first condition in (4) ensures that none of the weights $w_{i}$ is disproportionately large (see Krewski and Rao 1981). The second condition in (4) means that $\operatorname{Var}_{s}\left(\sum_{i \in S} w_{i} / N\right)$ is at most of the order $n^{-1}$. Conditions in (4) are satisfied for stratified simple random sampling designs.

## 3. Imputation process

Our proposed imputation is a type of regression imputation. Thus, one of the key issues to our method is to find the right "predictors" for nonrespondents. For a nonrespondent $y_{t}, t \geq 2$, let $r$ be the time point at which the unit has the last
observed value, i.e., $y_{r}$ is observed but $y_{r+1}, \ldots, y_{t-1}, y_{t}$ are nonrespondents. Under assumptions (1)-(2), we can use $y_{r}$ as a predictor in imputing $y_{t}$.

### 3.1 The case of $\boldsymbol{r}=\boldsymbol{t} \mathbf{- 1}$

We first consider the case of $r=t-1$. Let

$$
\phi_{t, t-1}\left(y_{t-1}\right)=E\left(y_{t} \mid y_{t-1}, \delta_{t}=0, \delta_{t-1}=1\right)
$$

be the conditional expectation (regression function) for a nonrespondent $y_{t}$ with observed $y_{t-1}$. If $\phi_{t, t-1}$ is known, we can simply impute $y_{t}$ by $\phi_{t, t-1}\left(y_{t-1}\right)$. But $\phi_{t, t-1}$ is usually unknown. It is shown in the Appendix that assumption (1) implies that

$$
\begin{equation*}
\phi_{t, t-1}\left(y_{t-1}\right)=E\left(y_{t} \mid y_{t-1}, \delta_{t}=1, \delta_{t-1}=1\right), \quad t=2, \ldots, T . \tag{5}
\end{equation*}
$$

Thus, $\phi_{t, t-1}$ can be estimated by regressing $y_{t}$ on $y_{t-1}$ using data from all sampled units having observed $y_{t}$ and $y_{t-1}$.

The idea of using (5) for imputation is the same as that in the monotone nonresponse case treated by Paik (1997). Unlike the monotone nonresponse case, however, $\phi_{t, t-1}(x)$ may not be linear in $x$ for nonmonotone nonresponse. Hence, we consider the nonparametric method in Cheng (1994) for regression. The kernel estimator of $\phi_{t, t-1}(x)$ is

$$
\begin{aligned}
& \hat{\phi}_{t, t-1}(x)= \\
& \quad \sum_{i \in S} \kappa\left(\frac{x-y_{t-1, i}}{h}\right) w_{i} I_{t, t-1, i} y_{t, i} / \sum_{i \in S} \kappa\left(\frac{x-y_{t-1, i}}{h}\right) w_{i} I_{t, t-1, i}
\end{aligned}
$$

where $\kappa(x)$ is a probability density function, $h>0$ is a bandwidth, and

$$
I_{t, t-1, i}= \begin{cases}1 & \delta_{t, i}=1, \delta_{t-1, i}=1 \quad t=2, \ldots, T \\ 0 & \text { otherwise }\end{cases}
$$

A nonrespondent $y_{t, j}$ with respondent $y_{t-1, j}$ is imputed by

$$
\tilde{y}_{t, j}=\hat{\phi}_{t, t-1}\left(y_{t-1, j}\right)
$$

Cheng (1994) suggested a bandwidth $h=\mathrm{Cn}^{-2 / 5}$, where $C$ is a constant. In general, $C$ may be different from application to application, and should be chosen using techniques developed in the kernel estimation literature (e.g., Cheng 1994 and Chapter 5 of Härdle 1990) and/or empirical studies.

### 3.2 The case of $r<t-1$

When $r<t-1$, the situation is more complicated. Let

$$
\phi_{t, r}\left(y_{r}\right)=E\left(y_{t} \mid y_{r}, \delta_{t}=\cdots=\delta_{r+1}=0, \delta_{r}=1\right)
$$

As nonresponse mechanism (1) is nonignorable, the expected value of $y_{t}$ conditional on $y_{r}$ with $r<t-1$ is not equal for observed and missing $y_{t}$, which precludes the use
of observed $y_{t}$ values as outcomes in estimating $\phi_{t, r}$. It is explicitly shown by Xu (2007) that

$$
\begin{align*}
& \phi_{t, r}\left(y_{r}\right) \\
& \quad \neq E\left(y_{t} \mid y_{r}, \delta_{t}=1, \delta_{t-1}=a_{t-1}, \ldots, \delta_{r+1}=a_{r+1}, \delta_{r}=1\right) \tag{6}
\end{align*}
$$

where $a_{j}=0$ or $1, j=r+1, \ldots, t-1$. On the contrary, in the case of monotone nonresponse the two sides of (6) are the same (Paik 1997) so that the right hand side of (6) can be used as the regression imputation model and observed $y_{t}$ values can be used to estimate $\phi_{t, r}$.

We have to find a conditional expectation of $y_{t}$ (given $y_{r}$ and some response status) that is equal to $\phi_{t, r}\left(y_{r}\right)$ and enables us to carry out imputation. It is shown in the Appendix that

$$
\begin{array}{r}
\phi_{t, r}\left(y_{r}\right)=E\left(y_{t} \mid y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=\delta_{r}=1\right), \\
r=1, \ldots, t-2, t=2, \ldots, T \tag{7}
\end{array}
$$

To estimate $\phi_{t, r}$ by fitting regression according to (7), we do not have observed $y_{t}$ values as outcomes in regression, because units defined by the right hand side of the equation in (7) have $\delta_{t}=0$. If we carry out imputation sequentially as $r=t-1, t-2, \ldots, 1$, then the $y_{t}$ nonrespondents for units with $\delta_{r+1}=1$ have already been imputed. Thus, we can use these previously imputed $y_{t}$ values as outcomes in regression. Although at each fixed time point $t$, imputation is carried out sequentially as $r=t-1, t-2, \ldots, 1$, imputation for different time points can be carried out at any order, because at any time point $t$, imputed values are used as outcomes at time point $t$ only.

Since $\phi_{t, r}$ is usually not linear, we use the kernel regression. For $t=2, \ldots, T$ and $r=t-2, t-3, \ldots, 1$, the conditional expectation $\phi_{t, r}(x)$ for any $x$ is estimated by $\hat{\phi}_{t, r}(x)$

$$
\begin{equation*}
=\sum_{i \in S} \kappa\left(\frac{x-y_{r, i}}{h}\right) w_{i} I_{t, r, i} \tilde{y}_{t, i} / \sum_{i \in S} \kappa\left(\frac{x-y_{r, i}}{h}\right) w_{i} I_{t, r, i} \tag{8}
\end{equation*}
$$

where $\tilde{y}_{t, i}$ is a previously imputed value and

$$
I_{t, r, i}= \begin{cases}1 & \delta_{t, i}=\cdots=\delta_{r+2, i}=0, \delta_{r+1, i}=\delta_{r, i}=1 \\ & r=1, \ldots, t-2, \quad t=2, \ldots, T \\ 0 & \text { otherwise }\end{cases}
$$

A nonrespondent $y_{t, j}$ with last respondent $y_{r, j}$ is imputed by

$$
\tilde{y}_{t, j}=\hat{\phi}_{t, r}\left(y_{r, j}\right) .
$$

The Markov chain assumption (2) ensures that using previously imputed values $\tilde{y}_{t, i}$ as outcomes in (8) produces an asymptotically valid estimator of $\phi_{t, r}$ (see result (11) in the Appendix).

### 3.3 An illustration

To illustrate the previous described imputation process and how nonresponse patterns are grouped into imputation cells, we consider imputation at time point $t=4$ (Table 1). The horizontal direction in Table 1 corresponds to 4 time points and the vertical direction corresponds to different nonresponse patterns, where each pattern is represented by a 4-dimensional vector of 0 's and 1 's with 0 indicating a nonrespondent and 1 indicating a respondent. There are a total of $2^{T-1}=2^{3}=8$ nonresponse patterns. According to the previously described imputation process, at step 1 , we consider nonrespondents at time 4 with last respondents at time 3 , which are patterns 3 and 4 . According to imputation model (5), we fit a regression using data in patterns 7 and 8 indicated by + (used as predictors) and $\times$ (used as outcomes) in the block in Table 1 under title step 1. Then, imputed values (indicated by $\bigcirc$ ) are obtained from the fitted regression using data indicated by ${ }^{*}$ as predictors in the block under title step 1. Next, we focus on the block in Table 1 under title step 2 . The nonrespondents at $t=4$ with last respondents at time 2 are those in pattern 2. According to imputation model (7), we fit a regression using data in pattern 3 indicated by + (used as predictors) and $\otimes$ (previously imputed values used as outcomes). Then, imputed values (indicated by $\bigcirc$ ) are obtained from the fitted regression using data indicated by $*$ as predictors. Finally, we focus on the block in Table 1 under title step 3. The nonrespondents at $t=4$ with last respondents at time 1 are those in pattern 1. According to imputation model (7), we fit a regression using data in pattern 2 indicated by + (used as predictors) and $\otimes$ (previously imputed values used as outcomes). Then, imputed values (indicated by $\bigcirc$ ) are obtained from the fitted regression using data indicated by * as predictors.

Table 1
Illustration of imputation process at $t=4$

|  | Step 1: $r=3$ Time |  |  |  | Step 2: $r=2$ Time |  |  |  | Step 3: $r=1$ Time |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Pattern | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| (1,0,0,0) |  |  |  |  |  |  |  |  | * |  |  | $\bigcirc$ |
| $(1,1,0,0)$ |  |  |  |  |  | * |  | $\bigcirc$ | $+$ |  |  | $\otimes$ |
| $(1,1,1,0)$ |  |  | * | $\bigcirc$ |  | + |  | $\otimes$ |  |  |  |  |
| $(1,0,1,0)$ |  |  | * | $\bigcirc$ |  |  |  |  |  |  |  |  |
| $(1,0,0,1)$ |  |  |  |  |  |  |  |  |  |  |  |  |
| $(1,1,0,1)$ |  |  |  |  |  |  |  |  |  |  |  |  |
| (1,0,1,1) |  |  | + | $\times$ |  |  |  |  |  |  |  |  |
| $(1,1,1,1)$ |  |  | $+$ | $\times$ |  |  |  |  |  |  |  |  |

+ : observed data used in regression fitting as predictors
$\times$ : observed data used in regression fitting as responses
$\otimes$ : imputed data used in regression fitting as responses
*: observed data used as predictors in imputation
O: imputed values


## 4. Estimation of population means using imputed data

Let $\bar{Y}_{t}$ be the finite population mean at time point $t$. The sample mean based on observed and imputed data is

$$
\begin{equation*}
\hat{\bar{Y}}_{t}=\sum_{i \in S} w_{i} \tilde{y}_{t, i} \tag{9}
\end{equation*}
$$

where $\tilde{y}_{t, i}$ is equal to the observed value if $\delta_{t, i}=1$ and is an imputed value if $\delta_{t, i}=0$. We now establish that, as an estimator for the population mean at time point $t, \hat{\bar{Y}}_{t}$ in (9) is consistent and asymptotically normal under the asymptotic frame work described in Section 2.

Theorem 1. Assume (1)-(2) and (4), and the asymptotic frame work described in Section 2. Assume further the following regularity conditions:
(C1) $E\left(y_{t}^{2}\right)<\infty, t=1, \ldots, T$.
(C2) $0<P\left(I_{t, r}=1\right)<1$ and $E\left[\sigma_{t, r}^{2}\left(y_{r}\right) / p_{t, r}\left(y_{r}\right)\right]<\infty$, where $p_{t, t-1}(x)=P\left(\delta_{t}=1 \mid y_{t-1}=x, \delta_{t-1}=1\right), p_{t, r}(x)=$ $P\left(\delta_{t}=0 \mid y_{r}=x, \delta_{r}=\delta_{r+1}=1, \delta_{t-1}=\cdots=\delta_{r+2}=0\right)$, $r=1, \ldots, t-2, \sigma_{t, r}^{2}(x)=\operatorname{Var}\left(y_{t} \mid y_{r}=x, I_{t, r}=1\right), I_{t, r}$ is the same as $I_{t, r, i}$ with $\delta_{t, i}$ 's replaced by $\delta_{t}$ 's, $r=1, \ldots, t-1, t=2, \ldots, T$.
(C3) $\phi_{t, r}(x)$ and $g_{t, r}(x)=p_{t, r}(x) f_{r}(x)$ have bounded second derivatives such that

$$
\begin{aligned}
E\left[\left\{\sigma_{t, r}^{2}\left(y_{r}\right)+\phi_{t, r}\right.\right. & \left.\left(y_{r}\right)\right\} g_{t, r}^{\prime \prime}\left(y_{r}\right) \\
& \times\left\{1-p_{t, r}\left(y_{r}\right)\right\} / \sqrt{g_{t, r}\left(y_{r}\right)}<\infty
\end{aligned}
$$

and

$$
\begin{aligned}
E\left[\left\{\phi_{t, r}\left(y_{r}\right) g_{t, r}^{\prime \prime}\left(y_{r}\right)^{2}\right.\right. & \left.+\phi_{t, r}^{\prime}\left(y_{r}\right) g_{t, r}^{\prime}\left(y_{r}\right)\right\} \\
& \left.\times\left\{1-p_{t, r}\left(y_{r}\right)\right\} / g_{t, r}\left(y_{r}\right)\right]<\infty
\end{aligned}
$$

where $f_{r}(x)$ is the probability density function of $y_{r}, r=1, \ldots, t-1, t=2, \ldots, T$.
(K) The kernel function $\kappa$ is a bounded and symmetric probability density function on the real line with finite second moment.
(B) The bandwidth $h$ satisfies $n h^{2} /(\log n)^{2} \rightarrow \infty$ and $n h^{4} \rightarrow 0$ as $n \rightarrow \infty$.

Then, for $t=1, \ldots, T$,

$$
\begin{equation*}
\sqrt{n}\left(\hat{\bar{Y}}_{t}-\mu_{t}\right) \rightarrow_{d} N\left(0, \sigma_{t}^{2}\right) \tag{10}
\end{equation*}
$$

where $\mu_{t}=E\left(y_{t}\right), \sigma_{t}^{2}$ is an unknown parameter, and $\rightarrow_{d}$ denotes convergence in distribution with respect to the joint distribution of $\left(y_{1, i}, \ldots, y_{T, i}, \delta_{2, i}, \ldots, \delta_{T, i}\right)$ and sampling (model and design).

The proof is given in the Appendix. Conditions (K) and (B) are exactly the same as those in Cheng (1994) and
conditions (C1)-(C3) are the same as those in Cheng (1994) applied to units defined by the right hand sides of (5)-(7).

Because of the complexity of the imputation, it is difficult to obtain an explicit form of $\sigma_{t}^{2}$ in (9). We consider the bootstrap method. A correct bootstrap can be obtained by applying the imputation process in each of the bootstrap samples, i.e., by imputing each bootstrap data set exactly the way the original data set is imputed (Shao and Sitter 1996). More specifically, we proceed as follows.

1. Within each imputation class, draw a bootstrap sample as a simple random sample with replacement from the sample, where the bootstrap sample size is the same as the number of sampled units in the imputation class. Combine the bootstrap samples to form $S^{*}$. The bootstrap data set contains all observed data, weights, and response indicators of units in $S^{*}$.
2. Apply the proposed imputation procedure to the bootstrap data set. Calculate the bootstrap analogue $\hat{\bar{Y}}_{t}{ }^{*}$.
3. Independently repeat the previous steps $B$ times to obtain $\hat{\bar{Y}}_{t}^{* 1}, \ldots, \hat{\bar{Y}}_{t}^{* B}$. The sample variance of $\hat{\bar{Y}}_{t}^{* 1}, \ldots$, $\hat{\bar{Y}}_{t}^{* B}$ is our bootstrap variance estimator for $\hat{\bar{Y}}_{t}$.

Note that the bootstrap method requires a large amount of repeated computation, which is the price paid for replacing a theoretical derivation of asymptotic variances. One may also use other valid bootstrap methods for survey data described in Shao and Tu (1995, Chapter 6).

Performance of the proposed bootstrap variance estimator is evaluated by simulation in the next section.

## 5. Simulation

A simulation study was conducted to evaluate the performance of the proposed imputation method in terms of the estimation of the mean of $y_{t}$. We considered a sample
of size 1,000 . Each sample unit has longitudinal data of size $T=4$. The population mean values for the 4 time points are $1.33,1.94,2.73$, and 3.67 , respectively. Longitudinal data were generated according to two models. In the first model, $\left(y_{1}, \ldots, y_{T}\right)$ is multivariate normal and follows the $\operatorname{AR}(1)$ model with correlation coefficient 0.9 and standard error 1 . In the second model, $\left(\log y_{1}, \ldots, \log y_{T}\right)$ is multivariate normal and follows the $\operatorname{AR}(1)$ model with correlation coefficient 0.9 and standard error 1 . All data at $t=1$ are observed. For $t=2, \ldots, T$, nonrespondents were generated using

$$
P\left(\delta_{t}=0 \mid y_{t-1}\right)=\frac{\exp \left\{1-1.2 y_{t-1}\right\}}{1+\exp \left\{1-1.2 y_{t-1}\right\}}
$$

for the case of normal data and

$$
P\left(\delta_{t}=0 \mid y_{t-1}\right)=\frac{\exp \left\{2-0.7 y_{t-1}\right\}}{1+\exp \left\{2-0.7 y_{t-1}\right\}}
$$

for the case of log-normal data. These nonresponse models were chosen so that the unconditional probabilities of nonresponse are about the same in the two cases (see Table 2).

For comparison, we included five estimators in the simulation: the sample mean of complete data, which is used as the gold standard; the sample mean of respondents that ignores nonrespondents; the sample mean based on simple linear regression imputation described in Section 1; the sample mean based on censoring (as described in Section 1) and linear regression imputation for monotone missing data (Paik 1997); and the sample mean based on our proposed imputation procedure. In the nonparametric regression described in Sections 3.1 and 3.2, the standard normal density was used as the kernel $\kappa(x)$ and the bandwidth $h$ was chosen to be around $4 n^{-2 / 5}$, which was used in the simulation in Cheng (1994).

Table 2
Unconditional probabilities of nonresponse in the simulation study

|  |  |  | Nonresponse probability |  |  |
| :--- | :--- | ---: | :--- | :--- | :--- |
|  | Nonresponse pattern |  | The normal case | The log-normal case |  |
| $t=3$ | Monotone | $(1,0,0)$ | 0.14 |  | 0.16 |
|  |  |  |  |  |  |
|  | Intermittent | $(1,1,0)$ | 0.12 | 0.26 | 0.09 |
| $t=4$ | $(1,0,1)$ | 0.25 | 0.25 | 0.20 | 0.20 |
|  | $(1,1,1)$ | 0.49 | 0.49 | 0.55 | 0.55 |
|  | $(1,0,0,0)$ | 0.04 |  | 0.06 |  |
|  | $(1,1,0,0)$ | 0.02 |  | 0.02 |  |
|  | $(1,1,1,0)$ | 0.04 | 0.10 | 0.02 | 0.10 |
|  | Intermittent | $(1,0,0,1)$ | 0.10 |  | 0.10 |
|  | $(1,0,1,0)$ | 0.04 |  | 0.03 |  |
|  | $(1,0,1,1)$ | 0.21 |  | 0.17 |  |
|  | $(1,1,0,1)$ | 0.10 | 0.45 | 0.06 | 0.36 |
|  | $(1,1,1,1)$ | 0.45 | 0.45 | 0.54 | 0.54 |

Tables 3-4 report (based on 1,000 simulation runs) the relative bias and variance of mean estimators, the bootstrap variance estimator (based on 200 bootstrap replications), the coverage probability of approximate $95 \%$ confidence intervals (CI) obtained using point estimator $\pm 1.96 \times \sqrt{\text { bootstrap variance }}$, and the length of CI. The results in Tables 3-4 can be summarized as follows.

1. Bias. The proposed imputation method produces estimators with negligible bias in all cases under consideration. The sample mean of respondents only is clearly biased unless $t=1$. Although in some cases the values of the bias are small, the bias leads to very low coverage probability of the CI, because the variance of the sample mean is much smaller than its squared bias. The simple linear regression imputation method is correct only when $t=2$ and data are normally distributed. Its relative bias at $t=3$ in the normal case is very small, but at $t=4$, it has a relative bias of $1.1 \%$ that leads to a coverage probability $76.3 \%$ only for its CI. The method of censoring and linear regression imputation is correct in the normal case and has little bias. In the log-normal case, however, both the simple linear regression imputation and the method of censoring with linear regression imputation are biased, due to the fact that the regression functions are not linear.
2. The bootstrap and CI. The bootstrap variance as an estimator of the variance of the mean estimator performs well in all cases, even when the mean estimator is biased. The related CI has a coverage probability close to the nominal level $95 \%$ when the mean estimator has no bias.
3. Proposed imputation versus censoring. When censoring and linear regression imputation is used, the mean estimator is biased in the log-normal case and, thus, the proposed imputation method is clearly better. In the normal case, both methods are correct. However, the results in Table 3 show the effect of discarding observed data. When $t=2$, censoring is better than the proposed imputation method, because no unit is actually censored and the censoring method uses the correct linear regression whereas the proposed imputation method fits a nonparametric regression. When $t=3$, censoring is about the same as the proposed imputation method but when $t=4$, censoring is a lot worse than the proposed imputation method. From Table 2, on average $25 \%$ sample units are censored when $t=3$ and $45 \%$ sample units are censored when $t=4$. The gain in using a correct linear regression is not enough to compensate the effect of discarding observed data, especially when $T=4$.

Table 3
Simulation results for estimation of means (Normal case)

| Method | Quantity | $\boldsymbol{t}=\mathbf{1}$ | $\boldsymbol{t}=\mathbf{2}$ | $\boldsymbol{t}=\mathbf{3}$ | $\boldsymbol{t}=\mathbf{4}$ |
| :--- | :--- | ---: | ---: | ---: | ---: |
| complete data | relative bias | $0.0 \%$ | $0.0 \%$ | $0.0 \%$ | $0.0 \%$ |
|  | variance $\times 10^{3}$ | 0.962 | 0.981 | 1.052 | 1.033 |
|  | bootstrap variance | 1.002 | 1.002 | 1.002 | 1.006 |
|  | coverage prob of CI | $95.4 \%$ | $94.9 \%$ | $94.5 \%$ | $94.5 \%$ |
|  | length of CI | 0.124 | 0.124 | 0.124 | 0.124 |
|  | relative bias |  | $16.8 \%$ | $8.3 \%$ | $3.5 \%$ |
|  | variance $\times 10^{3}$ |  | 1.319 | 1.240 | 1.051 |
|  | bootstrap variance |  | 1.364 | 1.178 | 1.062 |
|  | coverage prob of CI |  | $0.0 \%$ | $0.0 \%$ | $2.4 \%$ |
|  | length of CI |  | 0.145 | 0.134 | 0.128 |
|  | relative bias |  | $0.0 \%$ | $0.0 \%$ | $1.1 \%$ |
| simple linear | variance $\times 10^{3}$ |  | 1.121 | 1.434 | 1.185 |
| regression imputation |  | 1.172 | 1.466 | 1.192 |  |
|  | bootstrap variance |  | $94.9 \%$ | $94.7 \%$ | $76.3 \%$ |
|  | coverage prob of CI |  | 0.134 | 0.150 | 0.135 |
|  | length of CI |  | $0.0 \%$ | $0.0 \%$ | $0.0 \%$ |
| censoring and linear | relative bias |  | 1.121 | 1.437 | 1.642 |
| regression imputation | variance $\times 10^{3}$ |  | 1.172 | 1.476 | 1.819 |
|  | bootstrap variance |  | $94.9 \%$ | $94.7 \%$ | $96.1 \%$ |
|  | coverage prob of CI |  | 0.134 | 0.150 | 0.167 |
|  | length of CI |  | $0.2 \%$ | $0.3 \%$ | $0.2 \%$ |
| proposed imputation | relative bias |  | 1.196 | 1.438 | 1.264 |
|  | variance $\times 10^{3}$ |  | 1.231 | 1.401 | 1.224 |
|  | bootstrap variance |  | $95.0 \%$ | $93.7 \%$ | $94.1 \%$ |
|  | coverage prob of CI |  | 0.137 | 0.146 | 0.137 |

Table 4
Simulation results for estimation of means (Log-normal case)

| Method | Quantity | $\boldsymbol{t}=\mathbf{1}$ | $\boldsymbol{t}=\mathbf{2}$ | $\boldsymbol{t}=\mathbf{3}$ | $\boldsymbol{t}=\mathbf{4}$ |
| :--- | :--- | ---: | ---: | ---: | ---: |
| complete data | relative bias | $0.0 \%$ | $0.0 \%$ | $0.0 \%$ | $0.0 \%$ |
|  | variance | 0.069 | 0.172 | 0.383 | 1.074 |
|  | bootstrap variance | 0.067 | 0.161 | 0.418 | 1.138 |
|  | coverage prob of CI | $94.4 \%$ | $93.8 \%$ | $94.9 \%$ | $94.6 \%$ |
|  | length of CI | 1.008 | 1.557 | 2.511 | 4.142 |
|  | relative bias |  | $28.1 \%$ | $18.8 \%$ | $10.8 \%$ |
|  | variance |  | 0.366 | 0.614 | 1.378 |
|  | bootstrap variance |  | 0.344 | 0.668 | 1.461 |
|  | coverage prob of CI |  | $0.1 \%$ | $2.1 \%$ | $31.6 \%$ |
|  | length of CI |  | 2.267 | 3.171 | 4.690 |
|  | relative bias |  | $7.0 \%$ | $12.6 \%$ | $12.5 \%$ |
| simple linear | variance |  | 0.266 | 0.877 | 1.589 |
| regression imputation | bootstrap variance |  | 0.240 | 0.807 | 1.611 |
|  | coverage prob of CI |  | $71.6 \%$ | $39.3 \%$ | $23.2 \%$ |
|  | length of CI |  | 1.894 | 3.481 | 4.938 |
| relative bias |  | $7.0 \%$ | $12.1 \%$ | $13.8 \%$ |  |
| censoring and linear | regression imputation | variance |  | 0.266 | 0.874 |
|  | bootstrap variance |  | 0.240 | 0.836 | 2.735 |
|  | coverage prob of CI |  | $71.6 \%$ | $43.9 \%$ | $36.4 \%$ |
|  | length of CI |  | 1.894 | 3.540 | 6.277 |
| proposed imputation | relative bias |  | $0.1 \%$ | $0.1 \%$ | $0.1 \%$ |
|  | variance |  | 0.189 | 0.447 | 1.119 |
|  | bootstrap variance |  | 0.179 | 0.482 | 1.236 |
|  | coverage prob of CI |  | $94.5 \%$ | $95.7 \%$ | $95.6 \%$ |
|  | length of CI |  | 1.644 | 2.697 | 4.317 |

## 6. An example

In the CES introduced in Section 1, data for employment are collected from nonagricultural establishments on a monthly basis. In any particular month after the baseline, the nonresponse rate is about $20-40 \%$ and nonresponse is nonmonotone. In CES, it is typically assumed that (1)-(2) hold. In fact, assumption (3) that is stronger than assumption (2) is often assumed (Butani, Harter and Wolter 1997). We consider a stratified simple random sample from a CES dataset (a subset of a sample from the 1980's). Stratum sizes, sample size by stratum, and nonresponse rate by stratum are listed in Table 5. For each imputation method, imputation is carried out within a group of strata (group $1=$ strata $1-4$; group $2=$ strata $5-7$; group $3=$ strata $8-11$; group $4=$ stratum 12 ; group $5=$ strata $13-15$; group $6=$ stratum 16).

To estimate the employment counts from month 1 (baseline) to month 8, we applied the five methods in the simulation study in Section 5. The kernel and bandwidth in nonparametric regression were the same as those in the simulation (Section 5). Since population employment counts are obtained once a year from Unemployment Insurance
administrative records, nonrespondents in any month actually become available later so that the sample mean of complete data is available as a standard. The sample means based on different methods are reported in Table 6 together with their bootstrap variance estimates (based on bootstrap sample size 200). For the sample means based on respondents and three imputation methods, we also computed the estimated relative bias defined as (sample mean/sample mean of complete data) -1 .

The result in Table 6 shows that the sample mean based on the proposed imputation method is very comparable to the sample mean from the complete data, whereas the sample mean of respondents is clearly biased. Due to the fact that nonresponse is nonignorable, the simple linear regression imputation shows some bias starting from month 4, although the estimated relative bias is at most $5.5 \%$ in absolute value. The method of censoring with linear regression imputation has some bias after month 4, probably due to the fact that data are not normally distributed so that fitting linear regression is not correct. Furthermore, it has larger estimated variances compared with the proposed method, indicating the effect of discarding observed data.

Table 5
Stratum size, sample size, and nonresponse rate in the CES example

|  | Stratum |  |  |  |  |  |  |  | Sample |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Stratum | size | size | $\boldsymbol{t}=\mathbf{1}$ | $\boldsymbol{t}=\mathbf{2}$ | $\boldsymbol{t}=\mathbf{3}$ | $\boldsymbol{t}=\mathbf{4}$ | $\boldsymbol{t}=\mathbf{5}$ | $\boldsymbol{t}=\mathbf{6}$ | $\boldsymbol{t}=\mathbf{7}$ | $\boldsymbol{t}=\mathbf{8}$ |
| 1 | 223 | 102 | 0 | 32.4 | 39.2 | 34.3 | 27.5 | 30.4 | 28.4 | 33.3 |
| 2 | 1,649 | 110 | 0 | 28.2 | 31.8 | 30.0 | 34.5 | 33.6 | 26.4 | 29.1 |
| 3 | 1,900 | 120 | 0 | 37.5 | 39.2 | 34.2 | 44.2 | 41.7 | 40.0 | 43.8 |
| 4 | 419 | 98 | 0 | 41.8 | 48.0 | 35.7 | 38.8 | 44.9 | 38.8 | 38.8 |
| 5 | 1,947 | 132 | 0 | 37.1 | 33.3 | 25.8 | 25.0 | 27.3 | 23.5 | 28.8 |
| 6 | 2,391 | 180 | 0 | 41.1 | 36.1 | 42.8 | 37.8 | 39.4 | 39.4 | 38.3 |
| 7 | 5,365 | 256 | 0 | 35.2 | 34.0 | 33.6 | 36.7 | 35.2 | 40.6 | 39.1 |
| 8 | 2,330 | 201 | 0 | 30.3 | 36.8 | 40.3 | 34.8 | 37.3 | 37.3 | 37.8 |
| 9 | 1,164 | 113 | 0 | 35.4 | 29.2 | 33.6 | 30.1 | 29.2 | 32.7 | 33.6 |
| 10 | 593 | 106 | 0 | 37.7 | 44.3 | 40.6 | 47.2 | 41.5 | 37.7 | 32.1 |
| 11 | 2,222 | 182 | 0 | 24.2 | 26.4 | 27.5 | 27.5 | 28.0 | 20.3 | 27.5 |
| 12 | 6,880 | 512 | 0 | 40.0 | 39.6 | 40.6 | 41.0 | 41.4 | 39.8 | 38.9 |
| 13 | 2,373 | 160 | 0 | 36.9 | 40.6 | 36.2 | 33.8 | 39.4 | 30.0 | 36.9 |
| 14 | 50 | 42 | 0 | 40.5 | 38.1 | 28.6 | 45.2 | 33.3 | 33.3 | 31.0 |
| 15 | 4,100 | 241 | 0 | 36.5 | 38.6 | 34.4 | 34.9 | 42.7 | 33.2 | 32.8 |
| 16 | 3,951 | 412 | 0 | 37.9 | 36.9 | 36.9 | 38.1 | 40.3 | 40.5 | 39.3 |

Table 6
Estimates in the CES example

| Method | Quantity | $\boldsymbol{t}=\mathbf{1}$ | $\boldsymbol{t}=\mathbf{2}$ | $\boldsymbol{t}=\mathbf{3}$ | $\boldsymbol{t}=\mathbf{4}$ | $\boldsymbol{t}=\mathbf{5}$ | $\boldsymbol{t}=\mathbf{6}$ | $\boldsymbol{t}=\mathbf{7}$ | $\boldsymbol{t}=\mathbf{8}$ |
| :--- | :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| complete data | mean | 38.05 | 38.41 | 38.70 | 38.95 | 39.16 | 38.91 | 38.79 | 38.81 |
|  | $\widehat{\text { var }}$ | 0.805 | 0.814 | 0.828 | 0.830 | 0.990 | 0.832 | 0.822 | 0.852 |
| respondents | $\widehat{(\text { bian }}$ |  | 54.03 | 54.31 | 54.15 | 55.08 | 55.15 | 54.20 | 54.50 |
|  | $\widehat{\text { var }}$ |  | $40.7 \%$ | $40.3 \%$ | $39.0 \%$ | $40.7 \%$ | $41.7 \%$ | $39.7 \%$ | $40.4 \%$ |
|  | mean |  | 1.647 | 1.488 | 1.506 | 1.990 | 1.708 | 1.413 | 1.491 |
| simple linear |  | 38.45 | 38.72 | 39.33 | 38.81 | 41.04 | 40.54 | 39.79 |  |
| regression imputation | $\widehat{\text { bias }}$ |  | $0.1 \%$ | $0.1 \%$ | $1.0 \%$ | $-0.9 \%$ | $5.5 \%$ | $4.5 \%$ | $2.5 \%$ |
|  | $\widehat{\text { var }}$ |  | 0.821 | 0.834 | 0.866 | 0.963 | 1.979 | 1.465 | 1.008 |
| censoring and linear | mean |  | 38.45 | 38.71 | 39.17 | 38.25 | 40.46 | 40.34 | 40.30 |
| regression imputation | $\widehat{\text { bias }}$ |  | $0.1 \%$ | $0.0 \%$ | $0.6 \%$ | $-2.3 \%$ | $4.0 \%$ | $4.0 \%$ | $3.8 \%$ |
|  | $\widehat{\text { var }}$ |  | 0.821 | 0.833 | 0.881 | 1.289 | 1.497 | 1.630 | 1.660 |
| proposed imputation | mean |  | 38.37 | 38.68 | 38.97 | 39.10 | 39.05 | 38.72 | 38.88 |
|  | $\widehat{\text { bias }}$ |  | $-0.1 \%$ | $-0.1 \%$ | $0.0 \%$ | $-0.1 \%$ | $0.4 \%$ | $-0.2 \%$ | $0.2 \%$ |
|  | $\widehat{\text { var }}$ |  | 0.813 | 0.834 | 0.837 | 1.019 | 0.962 | 0.924 | 0.910 |

bias: (sample mean / sample mean of complete data) - 1
$\widehat{\text { var: }}$ bootstrap variance estimate

## 7. Concluding remarks

For longitudinal data with nonmonotone nonresponse, we propose an imputation method under the assumptions that the nonresponse mechanism is last-value-dependent and the longitudinal data follow a Markov chain. Our method is nonparametric and produces consistent and asymptotically normally distributed estimators of population means. Because the asymptotic variances of the estimators of population means are very complicated, we propose a simple bootstrap method for variance estimation. Some simulation results show that the proposed method works well. The CES is our motiving example and is used for illustration.

In general, nonresponse of data at time point $t$ may depend not only on the last value at time $t-1$, but also on other past values at time points prior to $t-1$. Furthermore, the longitudinal data may not be a Markov chain, i.e., there may be long time dependence among data from different
time points. In either case, our proposed method is not applicable. A general method is still under development.

## Appendix

## Proof of (5)

Let $L(\xi)$ denote the distribution of $\xi$ and $L(\xi \mid \zeta)$ denote the conditional distribution of $\xi$ given $\zeta$. Then,

$$
\begin{aligned}
L\left(y_{t} \mid y_{t-1}\right. & \left., \delta_{t}=0, \delta_{t-1}=1\right) \\
& =\frac{L\left(y_{t}, y_{t-1}, \delta_{t}=0, \delta_{t-1}=1\right)}{L\left(y_{t-1}, \delta_{t}=0, \delta_{t-1}=1\right)} \\
& =\frac{L\left(\delta_{t}=0 \mid y_{t}, y_{t-1}, \delta_{t-1}=1\right) L\left(y_{t}, y_{t-1}, \delta_{t-1}=1\right)}{L\left(\delta_{t}=0 \mid y_{t-1}, \delta_{t-1}=1\right) L\left(y_{t-1}, \delta_{t-1}=1\right)} \\
& =L\left(y_{t} \mid y_{t-1}, \delta_{t-1}=1\right),
\end{aligned}
$$

where the third equality follows from (1). Similarly, we can show that

$$
L\left(y_{t} \mid y_{t-1}, \delta_{t}=1, \delta_{t-1}=1\right)=L\left(y_{t} \mid y_{t-1}, \delta_{t-1}=1\right)
$$

Hence, $L\left(y_{t} \mid y_{t-1}, \delta_{t}=0, \delta_{t-1}=1\right)=L\left(y_{t} \mid y_{t-1}, \delta_{t}=1, \delta_{t-1}=1\right)$ and result (5) follows.

## Proof of (7)

Using the same notation as in the previous proof, we have

$$
\begin{aligned}
L\left(y_{t} \mid y_{r}, \delta_{t}=\right. & \left.\cdots=\delta_{r+1}=0, \delta_{r}=1\right) \\
= & \frac{L\left(y_{t}, y_{r}, \delta_{t}=\cdots=\delta_{r+1}=0, \delta_{r}=1\right)}{L\left(y_{r}, \delta_{t}=\cdots=\delta_{r+1}=0, \delta_{r}=1\right)} \\
= & \frac{L\left(\delta_{r+1}=0 \mid y_{t}, y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r}=1\right)}{L\left(\delta_{r+1}=0 \mid y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r}=1\right)} \\
& \times \frac{L\left(y_{t}, y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r}=1\right)}{L\left(y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r}=1\right)} \\
= & L\left(y_{t} \mid y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r}=1\right),
\end{aligned}
$$

where the last equality follows from (1). Similarly, we can show that

$$
\begin{aligned}
& L\left(y_{t} \mid y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=\delta_{r}=1\right) \\
& \quad=L\left(y_{t} \mid y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r}=1\right)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
L\left(y_{t} \mid y_{r}, \delta_{t}=\cdots\right. & \left.=\delta_{r+1}=0, \delta_{r}=1\right) \\
& =L\left(y_{t} \mid y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=\delta_{r}=1\right)
\end{aligned}
$$

and result (7) follows.

## Proof of theorem 1

Let $t(=2, \ldots, T)$ be fixed and $n_{t, r}=$ the number of units with $I_{t, r, i}=1, r=1, \ldots, t-1$. We first show that, for $r<t-1$,

$$
\begin{equation*}
E\left(\phi_{t, r+1}\left(y_{r+1}\right) \mid y_{r}, I_{t, r}=1\right)=E\left(y_{t} \mid y_{r}, I_{t, r}=1\right) \tag{11}
\end{equation*}
$$

Under assumption (1),

$$
\begin{aligned}
& L\left(y_{t} \mid y_{t-1}, \ldots, y_{1}, \delta_{t-1}, \ldots, \delta_{1}\right) \\
& =\frac{L\left(\delta_{t-1} \mid y_{t}, y_{t-1}, \ldots, y_{1}, \delta_{t-2}, \ldots, \delta_{1}\right)}{L\left(\delta_{t-1} \mid y_{t-1}, \ldots, y_{1}, \delta_{t-2}, \ldots, \delta_{1}\right)} L\left(y_{t} \mid y_{t-1}, \ldots, y_{1}, \delta_{t-2}, \ldots, \delta_{1}\right) \\
& =L\left(y_{t} \mid y_{t-1}, \ldots, y_{1}, \delta_{t-2}, \ldots, \delta_{1}\right) \\
& \ldots \\
& =L\left(y_{t} \mid y_{t-1}, \ldots, y_{1}\right) \\
& =L\left(y_{t} \mid y_{t-1}\right)
\end{aligned}
$$

where the last equality follows from assumption (2). Then

$$
\begin{aligned}
L\left(y_{t} \mid y_{t-1},\right. & \left.\delta_{t-1}\right) \\
= & \int L\left(y_{t} \mid y_{t-1}, \ldots, y_{1}, \delta_{t-1}, \ldots, \delta_{1}\right) \\
& d L\left(y_{t-2}, \ldots, y_{1}, \delta_{t-2}, \ldots, \delta_{1} \mid y_{t-1}, \delta_{t-1}\right) \\
= & L\left(y_{t} \mid y_{t-1}\right) \int d L\left(y_{t-2}, \ldots, y_{1}, \delta_{t-2}, \ldots, \delta_{1} \mid y_{t-1}, \delta_{t-1}\right) \\
= & L\left(y_{t} \mid y_{t-1}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& L\left(y_{t}, \delta_{t} \mid y_{t-1}, \ldots, y_{1}, \delta_{t-1}, \ldots, \delta_{1}\right) \\
&= L\left(\delta_{t} \mid y_{t}, y_{t-1}, \ldots, y_{1}, \delta_{t-1}, \ldots, \delta_{1}\right) \\
& \times L\left(y_{t} \mid y_{t-1}, \ldots, y_{1}, \delta_{t-1}, \ldots, \delta_{1}\right) \\
&= L\left(\delta_{t} \mid y_{t-1}\right) L\left(y_{t} \mid y_{t-1}\right) \\
&= L\left(\delta_{t} \mid y_{t}, y_{t-1}, \delta_{t-1}\right) L\left(y_{t} \mid y_{t-1}, \delta_{t-1}\right) \\
&= L\left(y_{t}, \delta_{t} \mid y_{t-1}, \delta_{t-1}\right) .
\end{aligned}
$$

Hence, $\quad\left\{\left(y_{t}, \delta_{t}\right), t=1, \ldots, T\right\} \quad$ is a Markov chain. Consequently,

$$
\begin{aligned}
E\left(y_{t} \mid y_{r+1},\right. & \left.\delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=1\right) \\
& =E\left(y_{t} \mid y_{r+1}, y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=\delta_{r}=1\right)
\end{aligned}
$$

Then, the left hand side of $(11)$ is equal to

$$
\begin{aligned}
E\left[E\left(y_{t} \mid y_{r+1}, y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=\delta_{r}=1\right) \mid\right. \\
\left.y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=\delta_{r}=1\right] \\
=E\left(y_{t} \mid y_{r}, \delta_{t}=\cdots=\delta_{r+2}=0, \delta_{r+1}=\delta_{r}=1\right),
\end{aligned}
$$

which is the right hand side of (11).
It follows from the construction of $\tilde{y}_{t, i}$ described in Section 3, result (11), and the proof of Theorem 2.1 in Cheng (1994) that

$$
\sqrt{n_{t, r}}\left(\bar{y}_{t, r}-\mu_{t, r}\right) \rightarrow_{d} N\left(0, \sigma_{t, r}^{2}\right)
$$

for some $\sigma_{t, r}^{2}$, where $\bar{y}_{t, r}=1 / n_{t, r} \sum_{i=1}^{n} I_{t, r, i} \tilde{y}_{t, i} \quad$ and $\mu_{t, r}=E\left[\phi_{t, r}\left(y_{r}\right) \mid I_{t, r}=1\right], r=1, \ldots, t-1$. The result follows from

$$
\begin{aligned}
\mu_{t} & =\sum_{r=1}^{t-1} E\left(y_{t} \mid I_{t, r}=1\right) P\left(I_{t, r}=1\right) \\
& =\sum_{r=1}^{t-1} E\left[E\left(y_{t} \mid y_{r}, I_{t, r}=1\right) \mid I_{t, r}=1\right] P\left(I_{t, r}=1\right) \\
& =\sum_{r=1}^{t-1} E\left[\phi_{t, r}\left(y_{r}\right) \mid I_{t, r}=1\right] P\left(I_{t, r}=1\right) \\
& =\sum_{r=1}^{t-1} \mu_{t, r} P\left(I_{t, r}=1\right)
\end{aligned}
$$

and

$$
\begin{aligned}
& \sqrt{n}\left(\hat{\bar{Y}}_{t}-\mu_{t}\right) \\
& =\sum_{r=1}^{t-1}\left[\frac{\sqrt{n_{t, r}}}{\sqrt{n}} \sqrt{n_{t, r}}\left(\bar{y}_{t, r}-\mu_{t, r}\right)+\mu_{t, r} \sqrt{n}\left(P\left(I_{t, r}=1\right)-\frac{n_{t, r}}{n}\right)\right] .
\end{aligned}
$$

## Acknowledgements

The authors would like to thank two referees and one associate editor for their helpful comments and suggestions. The research was partially supported by the NSF Grants DMS-0404535 and SES-0705033.

## References

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. The Annals of Statistics, 470482.

Butani, S., Harter, R. and Wolter, K. (1997). Estimation procedures for the Bureau of Labor Statistics Current Employment Statistics Program. Proceedings of the Section on Survey Research Methods, American Statistical Association, 523-528.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. Journal of the American Statistical Association, 81-87.
Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press, Cambridge, UK.
Kalton, G., and Kasprzyk, D. (1986). The treatment of missing data. Survey Methodology, 12, 1-16.
Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. The Annals of Statistics, 1010-1019.

Little, R.J., and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley \& Sons, Inc.
Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. Journal of the American Statistical Association, 1320-1329.
Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression method for repeated outcomes in the presence of missing data. Journal of the American Statistical Association, 90, 106-121.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley \& Sons, Inc.
Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. New York: CRC Press.
Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. Journal of the American Statistical Association, 1278-1288.
Shao, J., and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer.
Troxel, A.B., Harrington, D.P. and Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. Applied Statistics, 47, 425-438.

Troxel, A.B., Lipsitz, S.R. and Brennam, T.A. (1997). Weighted estimating equations with nonignorable missing response data. Biometrics, 53, 857-869.

Troxel, A.B., Lipsitz, S.R. and Harrigton, D.P. (1998). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotome missing data. Biometrika, 85, 661-672.
Valliant, R. (1993). Poststratification and conditional variance estimation. Journal of the American Statistical Association, 89-96.
Xu , J. (2007). Methods for intermittent missing responses in longitudinal data. Ph.D. Thesis, Department of Statistics, University of Wisconsin-Madison.

# Adaptive calibration for prediction of finite population totals 

Robert G. Clark and Raymond L. Chambers ${ }^{1}$


#### Abstract

Sample weights can be calibrated to reflect the known population totals of a set of auxiliary variables. Predictors of finite population totals calculated using these weights have low bias if these variables are related to the variable of interest, but can have high variance if too many auxiliary variables are used. This article develops an "adaptive calibration" approach, where the auxiliary variables to be used in weighting are selected using sample data. Adaptively calibrated estimators are shown to have lower mean squared error and better coverage properties than non-adaptive estimators in many cases.


Key Words: Sample surveys; Sample weighting; Prediction approach; Ridge estimation; Model selection; Stepwise procedures.

## 1. Introduction

Predictors of finite population totals are commonly calculated by weighted sums of sample values. Auxiliary variables are often available, whose sample values and population totals are known. Weights can be constructed so that weighted sums of auxiliary variables agree with the known population totals, a process called calibration (Deville and Särndal 1992). Predictors of finite population totals based on calibrated weights generally have much lower prediction bias than predictors calculated without auxiliary information.

Existing literature on finite population prediction essentially assumes that a set of useful auxiliary variables is chosen without reference to sample data. In practice, however, there may be a large set of potential auxiliary variables, not all of which should be used. Using additional auxiliary variables generally reduces the bias of calibrated predictors but increases the variance, so that using too many auxiliary variables can actually increase the mean squared error of calibrated predictors. The choice of which auxiliary variables to use is often not obvious, and sample data may be required to determine which set of auxiliary variables is appropriate for predictors of the totals of particular variables of interest. This paper develops methods for making this determination. Our approach may be called adaptive calibration, because the set of variables is chosen adaptively from sample data, rather than statically without reference to the sample at hand.

The prediction framework to finite population estimation will be used (see for example Brewer 1963; Royall 1970; Valliant, Dorfman and Royall 2000). In this approach, the population values of the variables of interest are treated as random variables. The aim is to predict the population total (which is also a random variable) or other finite population quantities using sample data on the variable of interest, and
population data on some auxiliary variables. The sample may have been selected using probability sampling or some other method, and is conditioned upon in inference. A stochastic model for the variable of interest is a central feature. One feature of the prediction framework is that misspecification of the model, for example due to omitting important auxiliary variables, can lead to substantial bias.

An alternative framework is the model-assisted approach (Särndal, Swensson and Wretman 1992). In this approach, a stochastic model is used but the model plays a less crucial role. The randomized nature of sampling is exploited to ensure that estimators are approximately unbiased even if the model is incorrect. When the model is correct, both approaches give approximately unbiased estimators, but the model-based approach would generally give lower variances of estimators of interest. If the model is mis-specified, then model-based predictors and variance estimators may be more biased, however robust model-based methods have been developed to combat this problem. For example Royall and Herson (1973a, 1973b) discuss robust prediction and Royall and Cumberland (1978, 1981a, 1981b) developed variance estimators that are robust to heteroscedasticity. For comparisons of the prediction and model-assisted frameworks, see for example Smith (1976) and Hansen, Madow and Tepping (1983).

The problem of selecting a set of auxiliary variables in the model-assisted framework was considered by Silva and Skinner (1997) and Skinner and Silva (1997). They found that adding calibration variables reduces the mean squared error (MSE) up to a point, after which adding further variables increases the MSE. Choosing calibration variables adaptively, based on sample data, gave better estimates than either calibrating on all variables or no variables. The applicability of this work to model-based prediction is not clear, because the role of the model is very different in the two frameworks. Mis-specified models can lead to

[^4]substantially biased model-based predictors, whereas model-assisted estimators are approximately unbiased even if important variables are omitted. As a result, different strategies for model selection could be appropriate in the two frameworks. Moreover, the differences between alternative approaches would be expected to be more pronounced in the prediction framework than in the modelassisted framework.

Chambers, Skinner and Wang (1999) proposed an approach for selecting calibration variables in the prediction framework, using forward, backward or stepwise selection. (This paper will henceforth be referred to as CSW.) The decision whether to omit (or add) a variable at each step was based on minimizing the estimated squared error of prediction (MSEP) for the predictor of interest. The approach was not evaluated by simulation study, and the estimators of MSEP used were not robust to heteroscedasticity.

The purpose of this paper is to develop the basic approach of CSW to apply to a wider range of situations, including heteroscedastic populations and multi-stage samples, and to evaluate the approach using realistic simulation studies. Estimators of the MSEP which are robust to heteroscedasticity, and to correlation in the case of multi-stage surveys, will be used. The performance of the estimators will be evaluated by simulation from two populations: financial data on farms generated from a farm survey and labour force data from a population census.

Following CSW, the basic approach will be to build a set of auxiliary variables using stepwise selection of variables, starting with some initial set. This algorithm builds up a set of auxiliary variables by a sequence of many decisions between two nested sets of variables. We compare several alternative criteria for deciding between two nested sets, including statistical significance and a number of alternative estimators of the mean squared error of prediction (MSEP). Three alternative estimators of MSEP are considered: a nonrobust estimator; an estimator of MSEP which is robust to heteroscedasticity; and an estimator which is robust both to heteroscedasticity and correlations within primary sampling units in multi-stage sampling.

Section 2 contains notation and definitions. Section 3 derives the difference in the MSEP of two predictors based on nested models, and develops several alternative estimators of this difference. Section 4 contains simulation results for a farm survey and a multi-stage household survey. Section 5 is a discussion. We conclude that adaptive calibretion generally performs better than static calibration, provided that a non-robust estimator of the MSEP, or statistical significance, is used as the objective in model selection.

## 2. Notation and definitions

A variable of interest $Y_{i}$ is observed for a sample $s$ of $n$ units, which is a subset of a finite population $U$ containing $N$ units. The aim is to estimate the population total $T_{Y}=$ $\sum_{i \in U} Y_{i}$ and other finite population quantities of $Y$. A vector of auxiliary variables $\boldsymbol{x}_{\boldsymbol{i}}$ is available for $i=1, \ldots, n$, with known population total $\boldsymbol{T}_{x}=\sum_{i \in U} \boldsymbol{x}_{i}$.

Weighted estimators of $T_{Y}$ are given by $\hat{T}_{Y}=\sum_{i \in s} w_{i} Y_{i}$, where $w_{i}$ can depend on the auxiliary variables but not on the variable of interest. A set of weights is said to be calibrated on $\boldsymbol{x}_{\boldsymbol{i}}$ if $\sum_{i \in s} w_{i} \boldsymbol{x}_{\boldsymbol{i}}=\boldsymbol{T}_{\boldsymbol{x}}$.

The best linear unbiased predictor (BLUP) based on a linear regression model is one example of a calibrated estimator. The most commonly used BLUP is based on the model

$$
\begin{align*}
E\left[Y_{i}\right] & =\boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}} \\
\operatorname{var}\left[Y_{i}\right] & =\sigma_{i}^{2}=v_{i} \sigma^{2}  \tag{1}\\
\operatorname{cov}\left[Y_{i}, Y_{j}\right] & =0(i \neq j)
\end{align*}
$$

(with $v_{i}$ assumed to be known) and is given by

$$
\begin{equation*}
\hat{T}_{Y}=\sum_{i \in s} Y_{i}+\sum_{i \in r} \hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{\boldsymbol{i}} \tag{2}
\end{equation*}
$$

where $r=U-s$ is the set of non-sample units and

$$
\begin{equation*}
\hat{\boldsymbol{\beta}}=\left\{\sum_{i \in s} v_{i}^{-1} \boldsymbol{x}_{\boldsymbol{i}} \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\right\}^{-1} \sum_{i \in s} v_{i}^{-1} \boldsymbol{x}_{\boldsymbol{i}} Y_{i} \tag{3}
\end{equation*}
$$

is a weighted least squares estimator of $\boldsymbol{\beta}$. The BLUP can also be written in weighted form as

$$
\hat{T}_{Y}=\sum_{i \in s} w_{i} Y_{i}
$$

where the weights $w_{i}$ are given by

$$
\begin{equation*}
w_{i}=1+\boldsymbol{T}_{x r}^{T}\left\{\sum_{j \in s} v_{j}^{-1} \boldsymbol{x}_{\boldsymbol{j}} \boldsymbol{x}_{j}^{T}\right\}^{-1} v_{i}^{-1} \boldsymbol{x}_{\boldsymbol{i}} \tag{4}
\end{equation*}
$$

and $\boldsymbol{T}_{x r}=\sum_{i \in r} \boldsymbol{x}_{\boldsymbol{i}}$. It is straightforward to verify that $\sum_{i \in s} w_{i} \boldsymbol{x}_{\boldsymbol{i}}=\boldsymbol{T}_{\boldsymbol{x}}$.

For heteroscedastic data, it is usually difficult to model $v_{i}$ reliably. In this case, robust estimators of the prediction variance of the BLUP are available, which do not rely on knowledge of $v_{i}$ (Royall and Cumberland 1978, 1981a, 1981b). For multi-stage samples, the assumption of independence may be violated. In this case, the BLUP based on (1) may still be used, and a robust ultimate cluster variance estimator of its prediction variance can be used (e.g., Valliant et al. 2000, Chapter 9). An alternative approach, which will not be considered here, would be to construct a BLUP based on a model that includes the
within-cluster correlations (Royall 1976). Section 3 will discuss robust and non-robust estimation of the mean squared error of prediction of the BLUP in more detail.

A decision needs to be made on what to include in $\boldsymbol{x}_{\boldsymbol{i}}$ in the BLUP. Stepwise selection, forward selection and backward selection are algorithms that can be used to decide which subset of the available auxiliary variables should be used. All three algorithms include many choices between two nested sets of auxiliary variables. Suppose the choice is between (A) using a predictor $\hat{T}_{A}$ based on $\boldsymbol{x}_{\boldsymbol{i}}$ and (B) using a predictor $\hat{T}_{B}$ based on a subvector $\boldsymbol{x}_{\boldsymbol{1} i}$. We can partition $\boldsymbol{x}_{\boldsymbol{i}}$ as $\boldsymbol{x}_{\boldsymbol{i}}=\left(\boldsymbol{x}_{1 i}^{T}, \boldsymbol{x}_{2 i}^{\boldsymbol{T}}\right)^{T}$. The number of elements of $\boldsymbol{x}_{i}, \boldsymbol{x}_{\mathbf{1 i}}$ and $\boldsymbol{x}_{\mathbf{2 i}}$ are denoted by $p, p_{1}$ and $p_{2}$, respectively.

We similarly partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta}=\left(\boldsymbol{\beta}_{1}^{\boldsymbol{T}}, \boldsymbol{\beta}_{\mathbf{2}}^{\boldsymbol{T}}\right)^{T}$. Predictor $\hat{T}_{A}$ is unbiased under model A :

$$
\begin{equation*}
E\left[Y_{i}\right]=\boldsymbol{\beta}^{T} \boldsymbol{x}_{i}=\boldsymbol{\beta}_{1}^{T} \boldsymbol{x}_{1 i}+\boldsymbol{\beta}_{2}^{T} \boldsymbol{x}_{2 i} \tag{5}
\end{equation*}
$$

The predictor $\hat{T}_{B}$ is unbiased for model B,

$$
\begin{equation*}
E\left[Y_{i}\right]=\boldsymbol{\beta}_{1}^{T} \boldsymbol{x}_{\mathbf{1 i}} \tag{6}
\end{equation*}
$$

which is the special case of model A where $\boldsymbol{\beta}_{\mathbf{2}}=\mathbf{0}$.

## 3. Estimation of the difference in the MSEP

### 3.1 Comparing predictors from nested models

Following CSW, our approach is to estimate the difference in the MSEPs of the two estimators:

$$
\Delta=E\left[\left(\hat{T}_{A}-T_{Y}\right)^{2}\right]-E\left[\left(\hat{T}_{B}-T_{Y}\right)^{2}\right]
$$

where the expectations are evaluated with respect to model A, because model B is a special case of this model. Typically, $\hat{T}_{A}$ will be less biased than $\hat{T}_{B}$ but have higher variance. Either predictor can have higher or lower MSEP depending on the particular population and sample.

For single stage sampling, it is usually reasonable to assume $Y_{i}$ and $Y_{j}$ independent for all $i \neq j$. Section 3.2 will derive $\Delta$ and an estimator of it in this case. Section 3.3 will describe the instructive special case where variances are equal and BLUPs are used; this was the case considered by CSW. Section 3.4 extends this by describing a hetero-scedasticity-robust estimator of $\Delta$. Section 3.5 further extends the approach by deriving $\Delta$ and an estimator of it for multi-stage sampling where there may be correlations between values from the same cluster.

### 3.2 Estimating $\Delta$ in single-stage sampling with known variance

In addition to model (5), we assume in this subsection that $Y_{i}$ and $Y_{j}$ are independent for $i \neq j$ and that
$\operatorname{var}\left[Y_{i}\right]=\sigma_{i}^{2}=\sigma^{2} v_{i}$ where $v_{i}$ are known. In this case, the MSEP of any predictor $\hat{T}=\sum_{i \in s} w_{i} Y_{i}$ is given by

$$
\begin{aligned}
\operatorname{MSEP}[\hat{T}] & =E\left[\left(\hat{T}-T_{Y}\right)^{2}\right] \\
& =\left\{E\left[\sum_{i \in s} w_{i} Y_{i}-\sum_{i \in U} Y_{i}\right]\right\}^{2}+\operatorname{var}\left[\sum_{i \in s}\left(w_{i}-1\right) Y_{i}-\sum_{i \in r} Y_{i}\right] \\
& =\left\{\boldsymbol{\beta}^{T}\left(\sum_{i \in s} w_{i} \boldsymbol{x}_{\boldsymbol{i}}-\sum_{i \in U} \boldsymbol{x}_{\boldsymbol{i}}\right)\right\}^{2}+\sum_{i \in s}\left(w_{i}-1\right)^{2} \sigma_{i}^{2}+\sum_{i \in r} \sigma_{i}^{2} .
\end{aligned}
$$

Writing $\boldsymbol{d}=\sum_{i \in s} w_{i} \boldsymbol{x}_{\boldsymbol{i}}-\boldsymbol{T}_{\boldsymbol{x}}$, we can rewrite the MSEP as

$$
\operatorname{MSEP}[\hat{T}]=\boldsymbol{d}^{T}\left(\boldsymbol{\beta} \boldsymbol{\beta}^{\boldsymbol{T}}\right) \boldsymbol{d}+\sum_{i \in s}\left(w_{i}-1\right)^{2} \sigma_{i}^{2}+\sum_{i \in r} \sigma_{i}^{2}
$$

Let $\boldsymbol{d}_{\boldsymbol{A}}=\sum_{i \in s} w_{A i} \boldsymbol{x}_{\boldsymbol{i}}-\boldsymbol{T}_{\boldsymbol{x}}$ and $\boldsymbol{d}_{\boldsymbol{B}}=\sum_{i \in s} w_{B i} \boldsymbol{x}_{\boldsymbol{i}}-\boldsymbol{T}_{\boldsymbol{x}}$. Then $\Delta$ is given by:

$$
\begin{align*}
\Delta & =\operatorname{MSEP}\left[\hat{T}_{A}\right]-\operatorname{MSEP}\left[\hat{T}_{B}\right] \\
& \left.=\boldsymbol{d}_{\boldsymbol{A}}^{\boldsymbol{T}}\left(\boldsymbol{\beta} \boldsymbol{\beta}^{\boldsymbol{T}}\right) \boldsymbol{d}_{\boldsymbol{A}}-\boldsymbol{d}_{\boldsymbol{B}}^{\boldsymbol{T}} \boldsymbol{(} \boldsymbol{\beta}^{\boldsymbol{T}}\right) \boldsymbol{d}_{\boldsymbol{B}} \\
& +\sum_{i \in S}\left(w_{A i}-1\right)^{2} \sigma_{i}^{2}-\sum_{i \in S}\left(w_{B i}-1\right)^{2} \sigma_{i}^{2} . \tag{7}
\end{align*}
$$

To estimate $\Delta$, we first consider how to estimate $\boldsymbol{\beta}$ and the variance of $\hat{\boldsymbol{\beta}}$. The usual weighted least squares estimator is $\hat{\boldsymbol{\beta}}=S_{x}^{-1} S_{x y}$ where $S_{x}=\sum_{i \in s} v_{i}^{-1} \boldsymbol{x}_{\boldsymbol{i}} \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}$ and $S_{x y}=\sum_{i \in s} s_{i}^{-1}$ $\boldsymbol{x}_{i} Y_{i}$. The usual estimator of the variance of $\hat{\boldsymbol{\beta}}$ is $\widehat{\operatorname{var}}[\hat{\boldsymbol{\beta}}]=$ $\hat{\sigma}^{2} S_{x}^{-1}$ where $\hat{\sigma}^{2}=\sum_{i \in S}\left(Y_{i}-\hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{\boldsymbol{i}}\right)^{2} v_{i}^{-1} /(n-p)$.

We can estimate $\left(\boldsymbol{\beta} \boldsymbol{\beta}^{\boldsymbol{T}}\right)$ unbiasedly by ( $\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{T}-\widehat{\operatorname{var}}[\hat{\boldsymbol{\beta}}]$ ). Hence the following is an unbiased estimator of $\Delta$ :

$$
\begin{align*}
\hat{\Delta} & =\boldsymbol{d}_{\boldsymbol{A}}^{\boldsymbol{T}}\left(\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}-\widehat{\operatorname{var}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d}_{\boldsymbol{A}}-\boldsymbol{d}_{\boldsymbol{B}}^{\boldsymbol{T}}\left(\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{T}-\widehat{\operatorname{var}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d}_{\boldsymbol{B}} \\
& +\sum_{i \in s}\left(w_{A i}-1\right)^{2} \hat{\sigma}^{2} v_{i}-\sum_{i \in S}\left(w_{B i}-1\right)^{2} \hat{\sigma}^{2} v_{i} \tag{8}
\end{align*}
$$

Expression (7) applies, and estimator (8) is an unbiased estimator of it, for any weighted predictors $\hat{T}_{A}$ and $\hat{T}_{B}$. We are concerned with the special case where $\hat{T}_{A}$ and $\hat{T}_{B}$ are BLUPs. In this case, $\hat{T}_{A}$ is calibrated to $\boldsymbol{T}_{\boldsymbol{x}}$ so that $\boldsymbol{d}_{\boldsymbol{A}}=\mathbf{0}$, and so (8) simplifies to

$$
\begin{align*}
\hat{\Delta} & =-\boldsymbol{d}_{\boldsymbol{B}}^{\boldsymbol{T}}\left(\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{T}-\widehat{\operatorname{var}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d}_{\boldsymbol{B}} \\
& +\sum_{i \in s}\left(w_{A i}-1\right)^{2} \hat{\sigma}^{2} v_{i}-\sum_{i \in S}\left(w_{B i}-1\right)^{2} \hat{\sigma}^{2} v_{i} \tag{9}
\end{align*}
$$

### 3.3 An important special case

In this Subsection, we make the assumptions stated in Section 3.2, and further assume that $v_{i}=1$ for all $i$. We also assume that the dimension of $\boldsymbol{x}_{2 i}$ is 1 , i.e. that we are considering whether or not one particular auxiliary variable from $\boldsymbol{x}_{\boldsymbol{i}}$ is to be used in prediction. Expressions (7) and (9) simplify in this case.

Let $u_{i}$ be the residual of a regression of $x_{2 i}$ on $\boldsymbol{x}_{1 i}$ :

$$
\begin{aligned}
& u_{i}=x_{2 i}-C^{T} \boldsymbol{x}_{1 i} \\
& C=\left(\sum_{i \in s} \boldsymbol{x}_{1 i} \boldsymbol{x}_{1 i}^{T}\right)^{-1} \sum_{i \in S} \boldsymbol{x}_{1 i} x_{2 i} .
\end{aligned}
$$

Using straightforward linear algebra operations, it can be shown that

$$
\boldsymbol{\beta}^{T} \boldsymbol{d}_{\boldsymbol{B}}=-\beta_{2} \sum_{i \in r} u_{i}
$$

and that

$$
\sum_{i \in s}\left(w_{A i}-1\right)^{2}-\sum_{i \in s}\left(w_{B i}-1\right)^{2}=\left(\sum_{i \in r} u_{i}\right)^{2} S_{u}^{-1}
$$

where $S_{u}=\sum_{i \in s} u_{i}^{2}$.
Hence (7) becomes

$$
\Delta=\sigma^{2}\left(\sum_{i \in r} u_{i}\right)^{2} S_{u}^{-1}-\left(\sum_{i \in r} u_{i}\right)^{2} \beta_{2}^{2}
$$

$\operatorname{CSW}$ show that $d_{B}^{T} \widehat{\operatorname{var}}\left[\hat{\mathrm{\beta}}_{2}\right] d_{B}=\hat{\sigma}^{2}\left(\sum_{i \in r} u_{i}\right)^{2} S_{u}^{-1}$. Hence (9) becomes

$$
\hat{\Delta}=\left(\sum_{i \in r} u_{i}\right)^{2}\left(2 \hat{\sigma}^{2} S_{u}^{-1}-\hat{\beta}_{2}^{2}\right)
$$

It is proposed that $\hat{T}_{A}$ be adopted when $\hat{\Delta}<0$, and $\hat{T}_{B}$ be used otherwise. It follows that we adopt $\hat{T}_{A}$ whenever $\hat{\beta}_{2}^{2}>2 \hat{\sigma}^{2} S_{u}^{-1}$. As noted by CSW, this is equivalent to adopting $\hat{T}_{A}$ whenever $F=\hat{\boldsymbol{\beta}}_{2}^{2} /\left(\hat{\sigma}^{2} S_{u}^{-1}\right)$ is greater than 2 . Notice that $F$ is the usual F -statistic for testing the null hypothesis that $\beta_{2}=0$. For large $n$, the cutoff for the Ftest at the $5 \%$ significance level is 3.96 , whereas we have arrived at a cutoff of 2 for adopting the larger set of variables. Thus, the decision to use A instead of B on the basis of a test of significance requires more evidence against B than a simple comparison of the estimated MSEPs of $\hat{T}_{A}$ and $\hat{T}_{B}$ would suggest. That is, using $\hat{\Delta}$ leads to larger models compared to using significance testing.

### 3.4 Heteroscedasticity-robust estimation of $\Delta$

The estimators of $\Delta$ in Sections 3.2 and 3.3 relied on knowing $\operatorname{var}\left[Y_{i}\right]$ at least up to a constant of proportionality. In practice, variances are at best known approximately, and methods which do not rely on an assumption of known variance may perform better. We will use an estimator of $\sigma_{i}^{2}$ which, assuming model (5), is approximately unbiased for $\sigma_{i}^{2}$ in general, and exactly unbiased if $\sigma_{i}^{2}=\sigma^{2}$ :

$$
\hat{\sigma}_{i}^{2}=\frac{n}{n-p}\left(Y_{i}-\hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{\boldsymbol{i}}\right)^{2} .
$$

(An alternative estimator would be $\hat{\sigma}_{i}^{2}=\left(Y_{i}-\hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{i}\right)^{2}$, as in Royall and Cumberland 1981b.)

The estimator of $\boldsymbol{\beta}$ would still be the weighted least squares estimator given by (3). The variance of $\hat{\boldsymbol{\beta}}$ is

$$
\begin{aligned}
\operatorname{var}[\hat{\boldsymbol{\beta}}] & =\operatorname{var}\left[S_{x}^{-1} S_{x y}\right] \\
& =\operatorname{var}\left[S_{x}^{-1} \sum_{i \in s} \boldsymbol{x}_{\boldsymbol{i}} Y_{i}\right] \\
& =S_{x}^{-1}\left(\sum_{i \in s} \boldsymbol{x}_{\boldsymbol{i}} \boldsymbol{x}_{i}^{T} \sigma_{i}^{2}\right) S_{x}^{-1} .
\end{aligned}
$$

This can be estimated by

$$
\widehat{\operatorname{var}}_{\text {robust }}[\hat{\boldsymbol{\beta}}]=S_{x}^{-1}\left(\sum_{i \in s} \boldsymbol{x}_{\boldsymbol{i}} \boldsymbol{x}_{i}^{T} \hat{\sigma}_{i}^{2}\right) S_{x}^{-1} .
$$

Hence we can estimate $\Delta$ by

$$
\begin{align*}
\hat{\Delta} & =\boldsymbol{d}_{A}^{T}\left(\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{T}-\widehat{\operatorname{var}}_{\text {robust }}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d}_{\boldsymbol{A}} \\
& -\boldsymbol{d}_{\boldsymbol{B}}^{T}\left(\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{T}-\widehat{\operatorname{var}}_{\text {robust }}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d}_{\boldsymbol{B}} \\
& +\sum_{i \in s}\left(w_{A i}-1\right)^{2} \hat{\sigma}_{i}^{2}-\sum_{i \in s}\left(w_{B i}-1\right)^{2} \hat{\sigma}_{i}^{2} . \tag{10}
\end{align*}
$$

### 3.5 Estimation of $\Delta$ in multi-stage sampling

The estimators of $\Delta$ in Sections 3.2, 3.3 and 3.4 all assumed that the values of $Y$ are independent for different units. In multi-stage sampling, a sample of primary sampling units (PSUs) is initially selected. A sample of units within the selected PSUs is then selected. For example, PSUs may be areas and units may be households or people; or PSUs could be schools and units could be students. Typically units from the same PSUs tend to be similar, so that values of $Y_{i}$ and $Y_{j}$ may be correlated if $i$ and $j$ belong to the same PSU. This section develops an estimator of $\Delta$ which is approximately unbiased even when there are correlations between values of $Y$ within the same PSU.

Let $s_{I}$ be the sample of PSUs, selected from the population $U_{I}$. Let $s_{g}$ be the sample of units from PSU $g$, where $g \in s_{I}$. Let $r_{I}=U_{I}-s_{I}$ and $r_{g}=U_{g}-s_{g}$. We assume model (5), and further assume that $Y_{i}$ and $Y_{j}$ are uncorrelated for $i \in g_{1}$ and $j \in g_{2}$ if $g_{1} \neq g_{2}$. The values $Y_{i}$ and $Y_{j_{\hat{N}}}$ may be correlated if $i \neq j$ with $i, j \in U_{g}$.

Let $\hat{T}=\sum_{i \in s} w_{i} Y_{i}$ be any predictor and let $\boldsymbol{d}=$ $\sum_{i \in s} w_{i} \boldsymbol{x}_{\boldsymbol{i}}-\boldsymbol{T}_{\boldsymbol{x}}$. The bias of $\hat{T}$ is $\boldsymbol{\beta}^{T} \boldsymbol{d}$, as in Section 3.2. The variance of $\left(\hat{T}-T_{Y}\right)$ is

$$
\begin{aligned}
\operatorname{var}\left[\hat{T}-T_{Y}\right] & =\operatorname{var}\left[\sum_{i \in s}\left(w_{i}-1\right) Y_{i}-\sum_{i \in r} Y_{i}\right] \\
& =\operatorname{var}\left[\sum_{g \in s_{l}}\left(\sum_{i \in s_{g}}\left(w_{i}-1\right) Y_{i}-\sum_{i \in r_{g}} Y_{i}\right)-\sum_{g \in \epsilon_{l}} \sum_{i \in U_{g}} Y_{i}\right] \\
& =\sum_{g \in s_{l}} \operatorname{var}\left(\sum_{i \in s_{g}}\left(w_{i}-1\right) Y_{i}-\sum_{i \in r_{g}} Y_{i}\right)+\sum_{g \in \epsilon_{I}} \operatorname{var}\left[\sum_{i \in U_{g}} Y_{i}\right] .
\end{aligned}
$$

It is further assumed that the variance of $\sum_{i e_{g}} Y_{i}$ and the covariance between $\sum_{i \in r_{8}} Y_{i}$ and $\sum_{i \in s_{g}}\left(w_{i}-1\right) Y_{i}$ are negligible relative to other terms. This is the case if cluster sampling is used (because in this case $s_{g}=U_{g}$ and $r_{g}$ is empty) or if the sampling fraction within PSUs is small. The variance becomes

$$
\operatorname{var}\left[\hat{T}-T_{Y}\right] \approx \sum_{g \in S_{I}} \operatorname{var}\left[\sum_{i \in s_{g}}\left(w_{i}-1\right) Y_{i}\right]+\sum_{g \in r_{I}} \operatorname{var}\left[\sum_{i \in U_{g}} Y_{i}\right] .
$$

Applying this to $\Delta$, we get:

$$
\begin{align*}
\Delta & =\operatorname{MSEP}\left[\hat{T}_{A}\right]-\operatorname{MSEP}\left[\hat{T}_{B}\right] \\
& =\boldsymbol{d}_{\boldsymbol{A}}^{T}\left(\boldsymbol{\beta} \boldsymbol{\beta}^{T}\right) \boldsymbol{d}_{\boldsymbol{A}}-\boldsymbol{d}_{\boldsymbol{B}}^{T}\left(\boldsymbol{\beta} \boldsymbol{\beta}^{T}\right) \boldsymbol{d}_{\boldsymbol{B}}+\sum_{g \in s_{I}} \operatorname{var}\left[\sum_{i \in s_{g}}\left(w_{A i}-1\right) Y_{i}\right] \\
& -\sum_{g \in s_{I}} \operatorname{var}\left[\sum_{i \in s_{g}}\left(w_{B i}-1\right) Y_{i}\right] . \tag{11}
\end{align*}
$$

To estimate $\Delta$, we need estimators of the variance of $\hat{\boldsymbol{\beta}}$, and of $\left(\sum_{i \in s_{g}}\left(w_{i}-1\right) Y_{i}\right)$.

Firstly, notice that

$$
\begin{aligned}
\operatorname{var}[\hat{\boldsymbol{\beta}}] & =\operatorname{var}\left[S_{x}^{-1} \sum_{g \in s_{l}} \sum_{i \in s_{g}} \boldsymbol{x}_{i} Y_{i}\right] \\
& =S_{x}^{-1} \sum_{g \in s_{I}} \operatorname{var}\left[\sum_{i \in s_{g}} \boldsymbol{x}_{i} Y_{i}\right] S_{x}^{-1} .
\end{aligned}
$$

This can be estimated using the "ultimate cluster variance" method by

$$
\widehat{\operatorname{var}}_{\mathrm{ucv}}[\hat{\boldsymbol{\beta}}]=S_{x}^{-1} \sum_{\boldsymbol{g} \in S_{i}}\left(\sum_{i \in s_{g}} \boldsymbol{x}_{\boldsymbol{i}}\left(Y_{i}-\hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{\boldsymbol{i}}\right)\right)^{2} S_{x}^{-1} .
$$

This is a well known estimator of the variance of a weighted sum from clustered data, and is equivalent to Valliant et al. (2000, 9.5 .5 , page 312 ). The variance has been called a "sandwich-level variance estimator using the cluster-level residuals" (Valliant et al. 2000) and an "ultimate cluster variance" (e.g., Wolter 1985 describes essentially the same idea in a randomization framework).

The variance of $\left(\sum_{i \in s_{g}}\left(w_{i}-1\right) Y_{i}\right)$ can also be estimated by the ultimate cluster variance method:

$$
\widehat{\operatorname{var}}\left[\sum_{i \in s_{g}}\left(w_{i}-1\right) Y_{i}\right]=\left\{\sum_{i \in s_{g}}\left(w_{i}-1\right)\left(Y_{i}-\hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{\boldsymbol{i}}\right)\right\}^{2} .
$$

Hence we can estimate $\Delta$ by

$$
\begin{align*}
\hat{\Delta}_{\mathrm{ucv}} & =\boldsymbol{d}_{\boldsymbol{A}}^{T}\left(\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{T}-\widehat{\operatorname{var}}_{\mathrm{ucv}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d}_{\boldsymbol{A}} \\
& -\boldsymbol{d}_{\boldsymbol{B}}^{T}\left(\hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^{T}-\widehat{\operatorname{var}}_{\mathrm{ucv}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d}_{\boldsymbol{B}} \\
& +\sum_{g \in s_{I}}\left\{\sum_{i \in s_{g}}\left(w_{A i}-1\right)\left(Y_{i}-\hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{i}\right)\right\}^{2} \\
& -\sum_{g \in s_{I}}\left\{\sum_{i \in s_{g}}\left(w_{B i}-1\right)\left(Y_{i}-\hat{\boldsymbol{\beta}}^{T} \boldsymbol{x}_{\boldsymbol{i}}\right)\right\}^{2} . \tag{12}
\end{align*}
$$

## 4. Simulation study

### 4.1 Simulation of farm survey

## Population and sampling scheme

The population distribution of the auxiliary variables, the sample and population size, and heteroscedasticity and other properties of the variable of interest would all be expected to play a part in the performance of the adaptive BLUPs. To make a realistic assessment of the performance of these estimators, a simulation study based on a large, realistic population is needed.

We generated a simulation population of 80,000 units, using sample data on 1652 farms from the 1988 Australian Agricultural and Grazing Industry Survey (AAGIS) as a starting point. Total cash crop was used as the survey variable of interest, and potential auxiliary variables included DSE (a derived size estimate), number of sheep, crops area, number of beef cattle, region (29 regions) and industry ( 5 industries). DSE was a linear combination of the sheep, crops area and beef cattle variables. The dataset also contained a sampling weight which was approximately equal to the inverse of the selection probability. 27 outliers with very large values of DSE were removed, as these would normally be placed in a completely enumerated stratum in a survey. A population of 80,000 was then constructed by probability proportional to size sampling with replacement, with probabilities proportional to the estimation weight on the original sample file.

250 samples were then selected without replacement from the simulation population. The samples were stratified by Region and DSE, with DSE divided into four categories, to give 116 strata. The category boundaries were set such that the category sums of DSE were equal. Total sample sizes of $250,500,1,000$ and 1,500 were simulated. The stratum sample sizes were proportional to the original AAGIS sample sizes by Region and DSE.

## Auxiliary variables and stepwise selection method

Auxiliary variables were included corresponding to the model containing: an intercept; sheep (x1); crops area (x2); beef cattle (x3); Industry; interaction of Industry and x1, x2 and x 3 ; and Region. This gives a total of 52 potential auxiliary variables. Some of these variables are collinear, but are still included in the set of potential variables, to give the model selection process a wider choice of possible models. We also considered the set of 139 auxiliary variables which included this set as well as the interaction of Region and x1, x2 and x3. Models were constructed by forward selection starting with the intercept-only model. Variables were added based on which step most reduced the estimated MSEP, for several alternative estimators of $\Delta$.

Stepwise selection was also trialled but was substantially slower to run for the larger variable set, and did not greatly improve the efficiency of the adaptive BLUPs.

An adaptive BLUP was also calculated based on statistical significance, with $p<0.05$ being the cutoff for inclusion. For each progressive model, the statistical significance of adding each of the variables not in the model was assessed, using a standard $t$-test. The variable with the lowest p -value was included in the model at each step. When there were no further significant variables which could be added, the procedure terminated and this was the model chosen.

A number of modifications were needed for the forward selection algorithm to work reliably: auxiliary variables were not added to the model if they had a pairwise Pearson correlation of 0.95 or higher (or -0.95 or lower); and variables were not added if this would result in the calibration equations not being solvable.

## Estimators used

Several BLUPs were calculated: with all auxiliary variables included; with just Intercept and DSE; and with auxiliary variables chosen by forward selection using the non-robust estimator of $\Delta$ (described in Section 3.2) or the heteroscedasticity-robust estimator of $\Delta$ (described in Section 3.4), from either the set of 52 or the set of 139 potential auxiliary variables. (The larger set of 139 variables was only evaluated for sample sizes of 500 and above.)

Ridge estimators (e.g., Bardsley and Chambers 1984) are an alternative approach to the problem of variable selection, so we included them in the simulation to compare their performance to that of the adaptive BLUPs. The estimators we have so far considered either include or exclude each variable. If a variable is included, then the weights must calibrate on that variable exactly, in the sense that $\sum_{i \in s} w_{i} \boldsymbol{x}_{\boldsymbol{i}}=\boldsymbol{T}_{\boldsymbol{x}}$. Ridge regression introduces a penalty for non-calibration, but does not necessarily require that the weights provide perfect calibration for all variables. In ridge regression, the vector of sample weights $\boldsymbol{w}$ is chosen to minimise

$$
\sum_{i \in s}\left(w_{i}-1\right)^{2} v_{i}^{-1}+\sum_{j=1}^{p} c_{j}^{-1}\left(\sum_{i \in s} w_{i} x_{i j}-T_{x j}\right)^{2}
$$

The $c_{j}$ are non-negative cost coefficients indicating the priority to be placed on meeting calibration constraint $j$. A value of 0 indicates that the constraint must be met precisely and larger cost coefficients result in placing less weight on the constraint. Thus the ridge estimator allows for a smooth reduction in the effective dimension of the model, by effectively interpolating between including a calibration variable $\left(c_{j}=0\right)$ and excluding it $\left(c_{j}=\infty\right)$.

Typically the $c_{j}$ are set to $\lambda c_{j}^{*}$, where $c_{j}^{*}$ reflect a somewhat subjective assessment of the relative importance of each constraint, and $\lambda$ is chosen to ensure that the final weights $w_{i}$ have reasonable properties, for example are all greater than or equal to 0 , or to 1 . We set $c_{j}^{*}$ to 0 for the constant (reflecting an intercept in the model), to 1 for $x 1, x 2$ and $x 3$, to 10 for the region indicators, to 5 for the industry indicators, and to 100 for interactions. The choice of $c_{j}^{*}$ was based on which variables were thought to be likely to be most useful. The value of $\lambda$ was numerically determined for each sample to be the smallest value such that all weights were greater than or equal to 1.

All of the methods were based on the same procedure for modelling $\operatorname{var}_{M}\left[Y_{i}\right]$. Firstly, a simple model with the intercept, $\mathrm{x} 1, \mathrm{x} 2$ and x 3 was fitted to the sample values of $Y$ using ordinary least squares. The $\log$ of the squared residuals from this model were then regressed against the log of DSE. The fitted values of this model were raised to the power of $e$ to give estimates of $\sigma_{i}^{2}$ for each $i \in s$. The estimated values of $\sigma_{i}^{2}$ were then truncated so that no values were more than 4 times, or less than one quarter, of the median value. This adjustment was made to avoid extreme values of $\sigma_{i}^{-2}$ which might lead to instability in calculating weighted least squares estimates of $\hat{\boldsymbol{\beta}}$. Results were somewhat sensitive to the variance modelling procedure, particularly the final adjustment to avoid extreme values: BLUPs based on a crude variance model with $\sigma_{i}^{2} \propto \mathrm{DSE}_{i}$ had variances around $10-20 \%$ higher than the BLUPs shown here.

## Results

Table 1 shows the Relative Root Mean Squared Error (RRMSE) of the various calibrated predictors. The first four rows of the table are for the first set of auxiliary variables ( 52 potential variables) and the last three rows are for the second set (139 potential auxiliary variables). Biases are not shown but were generally a relatively small component of the mean squared error for all of the predictors shown, except for the BLUP based on an intercept and DSE model, which was quite biased. This was somewhat surprising as we expected that a good trade-off between bias and variance would imply that biases were a non-negligible component of the mean squared error. Details on the biases and relative variances of the predictors can be found in Tables A1 and A2 of Clark and Chambers (2008).

Of the adaptive BLUPs, the significance criteria performed the best in all cases, followed by the nonrobust criteria, with the robust criteria performing worst. For the smaller set of 52 potential variables, the adaptive BLUPs based on the nonrobust and significance criteria performed better than the nonadaptive BLUPs for $\mathrm{n}=250$ and $\mathrm{n}=500$; for $\mathrm{n}=1,000$ and $\mathrm{n}=1,500$, they performed
slightly worse than the BLUP with all variables but better than the intercept and size BLUP. For the larger set of 139 potential variables, the adaptive BLUPs based on the nonrobust and significance criteria performed better than the nonadaptive BLUPs for all sample sizes, particularly for smaller values of $n$.

Table 1
RRMSE (\%) of AAGIS predictors of total cash crops

| \# Vars | n |  | BLUP |  | Adaptive BLUP |  |  |
| :---: | ---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | all | int+size | nonrobust $\hat{\boldsymbol{\Delta}}$ | robust $\hat{\boldsymbol{\Delta}}$ | Sig.Test |  |
| 52 | 250 | 3.59 | 3.02 | 2.97 | 3.09 | 2.87 | 3.30 |
|  | 500 | 2.35 | 2.54 | 2.33 | 2.33 | 2.30 | 2.31 |
|  | 1,000 | 1.56 | 2.21 | 1.58 | 1.64 | 1.57 | 1.54 |
|  | 1,500 | 1.36 | 2.22 | 1.39 | 1.41 | 1.37 | 1.37 |
| 139 | 500 | 3.52 | 2.54 | 2.99 | 3.44 | 2.29 | 2.27 |
|  | 1,000 | 1.77 | 2.21 | 1.75 | 1.92 | 1.72 | 1.59 |
|  | 1,500 | 1.56 | 2.22 | 1.51 | 1.64 | 1.42 | 1.42 |

The Ridge estimator generally performed about as well as the best of the adaptive BLUPs when there were 52 auxiliary variables, and slightly better when there were 139 potential variables.

Table 2 shows how many auxiliary variables were selected for the two adaptive BLUPs. The robust $\hat{\Delta}$ led to larger sets of auxiliary variables than the non-robust, with about 10 more auxiliary variables selected. The significance criteria led to even smaller variable sets (6-10 less variables than from the non-robust criteria).

Table 2
Mean (Interquartile range) of number of auxiliary variables selected in AAGIS

| \# Vars | n | nonrobust $\hat{\Delta}$ | robust $\hat{\boldsymbol{\Delta}}$ | Sig.Test |
| :---: | :---: | :---: | :---: | :---: |
| 52 | 250 | $16.0(14.0-18.0)$ | $26.9(24.0-29.0)$ | $9.6(8.0-11.0)$ |
|  | 500 | $18.6(16.0-21.0)$ | $27.4(25.0-30.0)$ | $11.5(10.0-13.0)$ |
|  | 1,000 | $23.6(21.0-26.0)$ | $29.6(26.0-33.0)$ | $14.4(13.0-16.0)$ |
|  | 1,500 | $27.3(25.0-29.0)$ | $32.3(30.0-35.0)$ | $17.2(16.0-18.8)$ |
| 139 | 500 | $42.1(37.0-47.0)$ | $69.4(62.0-75.0)$ | $23.2(21.0-26.0)$ |
|  | 1,000 | $51.5(47.0-56.0)$ | $74.2(69.0-79.8)$ | $29.9(27.0-33.0)$ |
|  | 1,500 | $59.2(55.0-64.0)$ | $75.8(71.0-81.0)$ | $34.9(32.0-38.0)$ |

Table 3 shows the confidence interval (CI) non-coverage of the various predictors. $90 \%$ CIs were defined as the estimator +/- 1.64 standard errors, where the variance was estimated using a heteroscedasticity-robust variance estimator (Royall and Cumberland 1978). Following common practice, CIs were based on estimated variance not estimated mean squared error of prediction. The simulation estimates of the non-coverage rates are fairly rough given that only 250 simulations were used. A larger simulation study could be used to give more precise estimates of coverage, but this was not pursued due to the
computationally intensive nature of the stepwise selection process. Table 3 suggests that: the BLUP using just intercept plus size had high non-coverage as did the adaptive BLUP based on robust $\hat{\Delta}$. The other estimators generally had non-coverage rates close to the nominal $10 \%$.

Table 3
Confidence interval non-coverage in AAGIS

| \# Vars | n |  |  |  | BLUP |  | Adaptive BLUP |  |  | Ridge |  |
| :---: | ---: | ---: | ---: | ---: | :---: | ---: | ---: | :---: | :---: | :---: | :---: |
|  |  | all |  |  |  | int+size | nonrobust $\boldsymbol{\Delta}$ |  |  |  |  |
| robust $\boldsymbol{\Delta}$ | Sig.Test |  |  |  |  |  |  |  |  |  |  |
| 52 | 250 | 10.0 | 6.4 | 10.4 | 16.8 | 11.2 | 10.0 |  |  |  |  |
|  | 500 | 8.0 | 13.2 | 12.0 | 17.2 | 10.8 | 8.0 |  |  |  |  |
|  | 1,000 | 7.6 | 20.4 | 9.2 | 12.0 | 8.4 | 8.4 |  |  |  |  |
|  | 1,500 | 8.8 | 34.8 | 9.2 | 13.2 | 9.6 | 8.8 |  |  |  |  |
| 139 | 500 | 16.8 | 13.2 | 18.0 | 29.2 | 12.8 | 8.8 |  |  |  |  |
|  | 1,000 | 12.4 | 20.4 | 14.0 | 20.4 | 13.2 | 7.2 |  |  |  |  |
|  | 1,500 | 13.6 | 34.8 | 13.6 | 19.6 | 12.4 | 11.2 |  |  |  |  |

Total cash crops is a major variable of interest in the AAGIS survey, but the totals of other variables are also important, including Farm Equity. For practical reasons, a single set of weights is normally used for all variables. Table 4 shows how well the adaptive calibration weights designed for the Total Cash Crops (TCC) variable performed when used to estimate the total of Farm Equity. For the case of 52 potential auxiliary variables, the adaptive BLUP weights chosen based on TCC (using non-robust $\hat{\Delta}$ ) performed reasonably well, as did the ridge estimator. Improvements could be made, however, by choosing auxiliary variables based on Equity.

Table 4
RRMSE (\%) of AAGIS predictors of total equity

| \# Vars | n | BLUP | Adaptive BLUP <br> (nonrobust $\hat{\boldsymbol{\Delta}}$ ) | Ridge |  |  |
| :---: | ---: | :---: | :---: | :---: | :---: | :---: |
|  |  | all | int+size | based on <br> TCC | based on <br> Equity |  |
| 52 | 250 | 6.85 | 6.45 | 6.51 | 6.13 | 6.78 |
|  | 500 | 4.44 | 4.44 | 4.61 | 4.40 | 4.28 |
|  | 1,000 | 3.09 | 3.12 | 3.42 | 3.14 | 3.10 |
|  | 1,500 | 2.54 | 2.58 | 2.90 | 2.58 | 2.54 |
| 139 | 500 | 5.53 | 4.93 | 4.98 | 4.74 | 4.20 |
|  | 1,000 | 3.68 | 4.03 | 3.23 | 3.15 | 3.08 |
|  | 1,500 | 3.04 | 3.63 | 2.66 | 2.60 | 2.57 |

### 4.2 Simulation of Labour Force Survey

## Population and Sampling Scheme

A simulation population was constructed by selecting a simple random sample without replacement of 30,000 people aged $15-64$ from the $1 \%$ sample file of the 1991 Australian Census of Population and Housing. The variable of interest was Employment ( 1 for employed people, 0 for others). The simulation population was divided into
simulated primary sampling units (PSUs) containing 75 people each, in such a way that the intra-cluster correlation was 0.05 . (This is a fairly typical intra-class correlation for the employment variable within primary sampling units in a household survey. See for example Clark and Steel 2002). The algorithm for defining clusters was to sort the data by a randomly generated $N\left(0, \gamma^{2}\right)$ variable plus the employment variable, then to define clusters as sequential sets of 75 people, where $\gamma$ was chosen so as to give the desired intracluster correlation.

The simulation consisted of 250 repeated two-stage samples. The first stage was a simple random sample without replacement of $m$ PSUs and the second stage was a simple random sample without replacement of 20 people from each selected PSU. The total sample size was set to be $n=200,400$ and 1,000 people. Most national household surveys have sample sizes much larger than this, but it is common to construct estimation post-strata within states or provinces, and the sample sizes for these areas would often be in the range 200-1,000.

The potential auxiliary variables were age by sex, where age was recorded in single years for 16-24 year olds, then in five year age groups $25-29,30-34, \ldots, 55-59$ year olds, and $60+$ year olds.

## Non-response

One of the main reasons why age and sex are used as auxiliary variables in household surveys is that nonresponse is known to depend on age and sex. For example, young men are typically the group with the lowest response rates. Non-response was simulated by assuming that the logit of the probability of response was equal to $1.8-$ $((\text { age }-50) / 25)^{2}$ for men, and $2-0.7((\text { age }-50) / 25)^{2}$ for women. This model gave a response rate of $75 \%$. The initial sample size was increased so that the final responding sample size was equal to $n=200,400$ or 1,000 .

## Auxiliary variables and stepwise selection method

The potential auxiliary variables were based on age by sex cells. The definition of the x-variables is shown in Table 5. This parameterization was chosen so that the auxiliary variables corresponding to specific ages or agegroups can be dropped while still giving a sensible model. For example, if all auxiliary variables were included except for $x_{4 i}$, then the model expected value for people aged 17 would be the same as those aged 16, rather than being equal to the intercept parameter. Even better results might be obtained from using more sophisticated parameterizations such as spline models and this will be investigated in a future study.

Table 5
Potential auxiliary variables in labour force survey simulation

| Variable | Definition |
| :--- | :--- |
| $x_{1 i}$ | 1 (corresponding to intercept in model for $Y$ ) |
| $x_{2 i}$ | 1 if person i male -1 if female |
| $x_{3 i}$ | 1 if person i aged 16 or over |
| $x_{4 i}$ | 1 if person i aged 17 or over |
| $\vdots$ | $\vdots$ |
| $x_{12, i}$ | 1 if person i aged 25 or over |
| $x_{13, i}$ | 1 if person i aged 30 or over |
| $\vdots$ | $\vdots$ |
| $x_{19, i}$ | 1 if person i aged 60 or over |
| $x_{20, i}$ | $x_{3 i}$ if person i male $-x_{3 i}$ if female |
| $\vdots$ | $\vdots$ |
| $x_{36, i}$ | $x_{19, i}$ if person i male $-x_{19, i}$ if female |

Stepwise selection was used to select variables, starting with the intercept-only model. At each step, variables could be added or removed, according to which gave the best reduction in the criteria. If the stepwise selection began cycling (for example, adding x 1 , then adding x 2 , then removing x 1 , then removing x 2 , then adding x 1 , etc), then the model building process stopped, and the the current model was used as the final model. The estimators of $\Delta$ used were the non-robust estimator, the robust (to heteroscedasticity) estimator and the ultimate cluster variance (UCV) estimator which is robust to heteroscedasticity and correlations within PSUs. Significance tests were not used as they would need to incorporate correlations within PSUs to be realistic. Results for the ridge estimator are not shown because negative weights rarely occurred in this simulation, so that this estimator performed very similarly to the BLUP using all auxiliary variables.

## Results

Table 6 shows the RRMSE of the various adaptive and non-adaptive BLUPs. There was relatively little difference in RRMSE between the BLUP with intercept only and the BLUP with all auxiliary variables. It is therefore not surprising that at best minor gains were made by using the adaptive BLUPs rather than using the BLUP with all variables. The adaptive BLUP using the non-robust $\hat{\Delta}$ gave the lowest RRMSE in all cases.

Table 6
RRMSE of labour force survey predictors of employment

| n | BLUP |  | Adaptive BLUP |  |  |
| ---: | :---: | :---: | :---: | :---: | :---: |
|  | all | intercept | nonrobust $\hat{\Delta}$ | robust $\hat{\Delta}$ | UCV $\hat{\Delta}$ |
| 200 | 6.54 | 6.77 | 6.44 | 7.06 | 6.96 |
| 400 | 4.72 | 4.76 | 4.61 | 4.72 | 4.65 |
| 1,000 | 2.45 | 2.70 | 2.43 | 2.45 | 2.49 |

Table 7 shows the mean number of variables selected for each of the adaptive BLUPs. Of the 36 potential auxiliary variables, between about 5 and 9 variables were selected based on the non-robust $\hat{\Delta}$. The number of variables selected increased as the sample size increased. The heteroscedasticity-robust criterion resulted in larger sets of auxiliary variables, and the UCV criterion gave even larger sets.

Table 7
Mean (Interquartile range) of number of auxiliary variables selected in labour force simulation

| n | Variable Selection Method |  |  |
| ---: | :---: | :---: | :---: |
|  | nonrobust | robust | UCV |
| 200 | $6.5(5.0-8.0)$ | $13.4(10.0-16.0)$ | $16.1(13.0-19.0)$ |
| 400 | $7.4(6.0-8.0)$ | $12.1(9.0-15.0)$ | $14.5(12.0-17.0)$ |
| 1,000 | $8.6(7.0-10.0)$ | $11.6(10.0-13.0)$ | $14.2(12.0-17.0)$ |

Table 8 shows the confidence interval (CI) non-coverage of the various predictors. $90 \%$ CIs were defined as the estimator +/- 1.64 standard errors, where the variance was estimated using a UCV variance estimator. Table 8 shows that the BLUP using all auxiliary variables had high noncoverage for $n=200$ and 400. The adaptive BLUP using nonrobust $\hat{\Delta}$ had reasonably close to nominal coverage, while the other adaptive BLUPs had high non-coverage.

Table 8
Confidence interval non-coverage (\%) for predictors of employment

| $\mathbf{n}$ | BLUP |  | Adaptive BLUP |  |  |
| ---: | ---: | :---: | :---: | :---: | :---: |
|  | all | intercept | nonrobust $\hat{\Delta}$ | robust $\hat{\Delta}$ | UCV $\hat{\boldsymbol{\Delta}}$ |
| 200 | 17.6 | 12.0 | 12.0 | 20.0 | 24.0 |
| 400 | 17.2 | 12.0 | 14.8 | 16.8 | 17.6 |
| 1,000 | 6.4 | 11.6 | 7.6 | 6.8 | 9.6 |

Table 9 shows how well the various weights performed when used to estimate a different variable, unemployment (equal to 1 for unemployed people and 0 otherwise). Adaptive BLUPs were calculated using the non-robust $\hat{\Delta}$, with the variable of interest given by Employment, and by Unemployment. The adaptive BLUP with variables chosen for Employment had RRMSE between the non-adaptive BLUP with all variables and the non-adaptive BLUP with intercept only. This suggests that this adaptive BLUP gives reasonable results even when applied to variables other than employment. The adaptive BLUP based on Unemploment actually had higher RRMSE. This may be because the auxiliary variables had little or no predictive power for unemployment, so that attempting to tailor the choice of auxiliary variables for this variable of interest did not work well.

Table 9
RRMSE of labour force survey predictors of unemployment

| $n$ | BLUP |  | Adaptive BLUP |  |
| ---: | :---: | :---: | :---: | :---: |
|  | all | intercept | based on emp | based on unemp |
| 200 | 36.3 | 32.6 | 34.5 | 36.0 |
| 400 | 24.1 | 21.7 | 22.8 | 23.7 |
| 1,000 | 14.5 | 14.2 | 14.1 | 14.2 |

## 5. Discussion

The simulation studies described here showed that adaptive BLUPs can give useful gains compared to simple non-adaptive alternatives. In both the farm survey and the labour force survey simulations, the adaptive BLUPs based on a nonrobust estimator of $\Delta$ and based on significance testing both had lower MSEP than non-adaptive estimators in almost all cases. In the case of the farm survey, the gains were sometimes substantial compared to either always using the full model or always using the intercept plus size variable model. In the case of the labour force survey, the gains were minor. The adaptive BLUPs also gave reasonable confidence interval coverage.

The adaptive BLUPs based on the robust and UCV criteria performed much worse than the other adaptive BLUPs. This is surprising, as the AAGIS data is known to be heteroscedastic and the Labour Force data was clustered suggesting that the UCV criteria should have given good results. Further analysis of the farms survey simulation showed that $\hat{\Delta}_{\text {robust }}$ had higher variances than $\hat{\Delta}_{\text {nonrobust }}$ in the great majority of cases, particularly for auxiliary variables with little predictive power - see the Appendix of Clark and Chambers 2008 for details. This suggests that the robust method would tend to select counter-productive auxiliary variables more often and could explain its poor performance.

There was a general tendency for all of the adaptive procedures to choose too many auxiliary variables, but despite this, the adaptive estimators generally performed better than or similar to simple non-adaptive alternatives. We suggest that in practice, an automatic model search (using either a non-robust $\hat{\Delta}$ or a statistical significance criterion) should be used in conjunction with some subjective judgement. For example, models could be selected from several sets of potential auxiliary variables of different sizes. If the larger sets gave only small apparent improvements, then the statistician might decide to restrict to a smaller set, even if apparently slightly suboptimal.

Ridge estimators also performed reasonably well in terms of RRMSE and confidence interval coverage. They generally gave similar results to the adaptive BLUPs for estimating the total of the variable of interest when the choice of auxiliary variables was based on this variable.

However, when the adaptive BLUP weights were applied to different variables, the ridge estimators performed slightly better. An even better approach may be adaptively choose both which auxiliary variables to include and how to apply ridging, based on some criterion calculated from the sample. This will be the topic of future research.

One concern that has been raised with the prediction approach to finite population sampling is its non-robustness to the omission of important auxiliary variables. In our simulations from farm economic data and social data, the adaptive predictors had low bias and lower mean-squared error than the non-adaptive estimators in most of the wide range of cases in our simulation study, and were never substantially worse. Provided that all design variables are considered as potential auxiliary variables, adaptive calibration provides a robust and efficient strategy for finite population prediction.

## Acknowledgements

This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics. An associate editor and two referees made detailed and thoughtful comments which considerably improved the paper.

## References

Bardsley, P., and Chambers, R. (1984). Multipurpose estimation from unbalanced samples. Applied Statistics, 33(3), 290-299.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deductible from the assumption of an underlying stochastic process. Australian Journal of Statistics, 5, 93-105.

Chambers, R., Skinner, C. and Wang, S. (1999). Intelligent calibration. Bulletin of the International Statistical Institute, 58(2), 321-324.

Clark, R.G., and Chambers, R.L. (2008). Adaptive calibration for prediction of finite population totals. University of Wollongong Centre for Statistical and Survey Methodology. Available from http://www.cssm.uow.edu.au.

Clark, R.G., and Steel, D.G. (2002). The effect of using household as a sampling unit. International Statistical Review, 70(2), 289-314.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87(418), 376-382.

Hansen, M., Madow, W. and Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. Journal of the American Statistical Association, 78, 776793.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. Biometrika, 57(2), 377-387.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. Journal of the American Statistical Association, 71(355), 657-664.

Royall, R.M., and Cumberland, W.G. (1978). Variance estimation in finite population sampling. Journal of the American Statistical Association, 73(362), 351-358.

Royall, R.M., and Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. Journal of the American Statistical Association, 76(373), 66-80.

Royall, R.M., and Cumberland, W.G. (1981b). The finite-population linear regression estimator and estimators of its variance - an empirical study. Journal of the American Statistical Association, 76(376), 924-930.

Royall, R.M., and Herson, J. (1973a). Robust estimation in finite populations 1. Journal of the American Statistical Association, 68(344), 880-889.

Royall, R.M., and Herson, J. (1973b). Robust estimation in finite populations 2: Stratification on a size variable. Journal of the American Statistical Association, 68(344), 890-893.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model assisted survey sampling. New York: Springer-Verlag.

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

Skinner, C., and Silva, P.L.N. (1997). Variable selection for regression estimation in the presence of nonresponse. Proceedings of the Section on Survey Research Methods, American Statistical Association, 76-81.

Smith, T.M.F. (1976). The foundations of survey sampling: a review. Journal of the Royal Statistical Society, Series A, 139(2), 183-202.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: John Wiley \& Sons, Inc.

Wolter, K.M. (1985). Introduction to variance estimation. New York: Springer-Verlag.

# Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added 

Lionel Qualité and Yves Tillé ${ }^{1}$


#### Abstract

We propose a method for estimating the variance of estimators of changes over time, a method that takes account of all the components of these estimators: the sampling design, treatment of non-response, treatment of large companies, correlation of non-response from one wave to another, the effect of using a panel, robustification, and calibration using a ratio estimator. This method, which serves to determine the confidence intervals of changes over time, is then applied to the Swiss survey of value added.


Key Words: Covariance; Stratified sampling; Panel.

## 1. Introduction

In longitudinal surveys, the precision of changes over time depends directly on the rate of overlap of the samples. We begin by reviewing known results for disjoint simple designs (on this subject, see Kish 1965; Sen 1973; Wolter 1985; Laniel 1988; Hidiroglou, Särndal and Binder (1995); Holmes and Skinner 2000; Nordberg 2000; Fuller and Rao 2001; Berger 2004). Next, we calculate the variance of such changes for simple designs in which the samples overlap. When the sampling ratios are very low, most of these results are well known and are described, for example, in Caron and Ravalet (2000). Results that take account of finite population corrections can be seen in Tam (1984).

We precisely calculated the variances of estimators for a larger class of sampling designs with a finite population. Finite population corrections can play a major role in business surveys, since large companies are sometimes selected with very high probabilities of inclusion. The calculations become much more complicated with a finite population for the following reason: if the size of the population is finite, two disjoint samples are not independent. If the population is infinite, two independent samples are disjoint. Several estimators are examined: the difference of the cross-sectional estimators; the difference estimated solely on the common portion; and relative changes. The calculations become even more complex when the population is dynamic (with births, deaths, changes of structure). The theory that we develop below is limited to the case in which the population does not change over time.

In the first part, we describe the two-dimensional simple random sampling design (on this subject, see Goga 2003) and we give the corresponding Horvitz-Thompson estimators. We calculate the variance of the estimator of
changes that is based on this sampling design. In a second part, we give the variance of other simple estimators: the relative change or the totals quotient, and the difference estimator based on the overlap of the samples. We then describe how these results adapt to the presence of ignorable non-response and the use of more complex estimators, which introduce weights modified to obtain calibrated estimators, or variables modified by a robustification procedure.

These results for simple designs are easy to generalize to stratified designs, provided that companies do not change strata from one wave to the next. Lastly, we apply this method to the Swiss survey of value added, taking all components of the survey into account: stratification, the panel effect, non-response, correlation between nonresponses from one wave to the next, calibration using a ratio estimator, and robustification.

## 2. Estimation of the difference in simple designs

Let there be a population $U=\{1, \ldots, k, \ldots, N\}$ of size $N$ in which two samples are taken: $s_{1}$ and $s_{2}$ of respective sizes $n_{1}$ and $n_{2}$. These samples may have a common portion (see Figure 1).

Assume that $s_{1}$ and $s_{2}$ are samples taken according to a simple design without replacement, and sizes $n_{1}$ and $n_{2}$ are therefore not random. Samples $s_{1}$ and $s_{2}$ can be broken down into three parts $s_{A}=s_{1} \backslash s_{2}, s_{B}=s_{2} \backslash s_{1}$, and $s_{C}=$ $s_{C}=s_{1} \cap s_{2}$. Let $n_{A}=\left|s_{A}\right|, n_{B}=\left|s_{B}\right|, n_{C}=\left|s_{C}\right|, n_{1}=n_{A}+$ $n_{C}, n_{2}=n_{B}+n_{C}$. The sizes of $s_{A}, s_{B}$, and $s_{C}$, may be random. This design generalizes the following hypothetical cases:

[^5]- If samples $s_{1}$ and $s_{2}$ are selected independently, $n_{C}$ is then a random variable;
- If sample $s_{1}$ is selected first, and sample $s_{2}$ is selected in the complement of $s_{1}$ in $U$, then $s_{C}$ is empty and $n_{C}=0$;
- if sample $s_{1}$ is selected first, and sample $s_{2}$ consists of the union of a subsample of fixed size of $s_{1}$ and a sample of fixed size of the complement of $s_{1}$ in $U$, then $n_{C}$ is not random, and the situation is the same as in case A of Tam (1984).


Figure 1 Overlapping samples
We make the additional hypothesis that conditional on $n_{A}, n_{B}$, and $n_{C}$, samples $s_{A}, s_{B}$, and $s_{C}$, are simple, without replacement and of fixed size. They come from the following sampling design:

Definition 1. Two-dimensional simple fixed-size sampling $\operatorname{design}\left(n_{A}, n_{B}, n_{C}\right)$ :

$$
\begin{aligned}
& p_{\text {simple }}\left(s_{1}, s_{2} \mid n_{A}, n_{B}, n_{C}\right)= \\
& \begin{cases}\frac{n_{A}!n_{B}!n_{C}!\left(N-n_{A}-n_{B}-n_{C}\right)!}{N!} & \text { if } n_{A}=\left|s_{A}\right| \\
0 & n_{B}=\left|s_{B}\right|, n_{C}=\left|s_{C}\right| \\
0 & \text { otherwise },\end{cases}
\end{aligned}
$$

where $s_{A}=s_{1} \backslash s_{2}, s_{B}=s_{2} \backslash s_{1}$ and $s_{C}=s_{1} \cap s_{2} \quad$ (on this subject, see Goga 2003).

The law for drawing the pair $\left(s_{1}, s_{2}\right)$, which we do not know in general, is thus assumed to be of the form

$$
p\left(s_{1}, s_{2}\right)=p_{\text {simple }}\left(s_{1}, s_{2} \mid n_{A}, n_{B}, n_{C}\right) \operatorname{Pr}\left(\left|s_{1} \cap s_{2}\right|=n_{C}\right)
$$

Let there be two variables $x$ and $y$ whose values, taken on the units of $U$, are denoted respectively $x_{k}$ and $y_{k}, k \in U$. Variables $x$ and $y$ may represent the same variable measured at two different times. Also assume that $x$ can be observed only for $s_{1}$ and $y$ for $s_{2}$. The objective is to estimate the totals

$$
X=\sum_{k \in U} x_{k} \text { and } Y=\sum_{k \in U} y_{k}
$$

as well as the difference $Y-X$. The Horvitz-Thompson estimators of $X$ and $Y$ are given by

$$
\hat{X}_{1}=\frac{N}{n_{1}} \sum_{k \in s_{1}} x_{k} \text { and } \hat{Y}_{2}=\frac{N}{n_{2}} \sum_{k \in s_{2}} y_{k} .
$$

### 2.1 Natural estimation of the difference

### 2.1.1 Variance of the estimation of the difference

A first approach for estimating $\Delta=Y-X$ is to use the difference of the cross-sectional estimators $\hat{\Delta}=\hat{Y}_{2}-\hat{X}_{1}$, which is an unbiased estimator conditional on $n_{C}$ according to the following simple design:

$$
E\left(\hat{\Delta} \mid n_{C}\right)=Y-X
$$

and is therefore also unbiased under design $p$ unconditional on $n_{C}$.
Proposition 1: The variance of $\hat{\Delta}$ is:

$$
\begin{align*}
\operatorname{var}(\hat{\Delta})= & N^{2}\left(\frac{1}{n_{1}}-\frac{1}{N}\right) S_{x}^{2}+N^{2}\left(\frac{1}{n_{2}}-\frac{1}{N}\right) S_{y}^{2} \\
& -2 N^{2}\left(\frac{E\left(n_{C}\right)}{n_{1} n_{2}}-\frac{1}{N}\right) S_{x y} \tag{1}
\end{align*}
$$

where

$$
\begin{aligned}
& S_{x}^{2}=\frac{1}{N-1} \sum_{k \in U}\left(x_{k}-\bar{X}\right)^{2}, S_{y}^{2}=\frac{1}{N-1} \sum_{k \in U}\left(y_{k}-\bar{Y}\right)^{2} \\
& S_{x y}=\frac{1}{N-1} \sum_{k \in U}\left(x_{k}-\bar{X}\right)\left(y_{k}-\bar{Y}\right) .
\end{aligned}
$$

The demonstration of (1) is appended.

### 2.1.2 Specific cases and precision gain

Result (1) can be used to deal directly with the following specific cases of co-ordination:

- if the two samples form a panel, $n_{C}=n_{1}=n_{2}$, then

$$
\operatorname{var}(\hat{\Delta})=N^{2}\left(\frac{1}{n_{C}}-\frac{1}{N}\right)\left(S_{x}^{2}+S_{y}^{2}-2 S_{x y}\right)
$$

- if the samples are disjoint (also see Ardilly and Tillé 2003, pages 24-28), $n_{C}=0$, and

$$
\begin{aligned}
\operatorname{var}(\hat{\Delta}) & =N^{2}\left(\frac{1}{n_{1}}-\frac{1}{N}\right) S_{x}^{2}+N^{2}\left(\frac{1}{n_{2}}-\frac{1}{N}\right) S_{y}^{2} \\
& +2 N S_{x y} .
\end{aligned}
$$

Surprisingly, the covariance does not depend on the sizes of the samples. It is negative if $x$ and $y$ are positively correlated, and it becomes negligible in relation to the variance terms when the size of the population is large;

- if $q$ is the set rate of overlap of the two samples and $n_{1}=n_{2}=n$, we are back to case A developed by Tam (1984). We then obtain $n_{C}=q n$, and

$$
\operatorname{var}(\hat{\Delta})=N^{2}\left(\frac{1}{n}-\frac{1}{N}\right)\left(S_{x}^{2}+S_{y}^{2}\right)-2 N^{2}\left(\frac{q}{n}-\frac{1}{N}\right) S_{x y}
$$

- if the two samples are independent, $E\left(n_{C}\right)=$ $n_{1} n_{2} / N$, and we have

$$
\operatorname{var}_{\mathrm{IND}}(\hat{\Delta})=N^{2}\left(\frac{1}{n_{1}}-\frac{1}{N}\right) S_{x}^{2}+N^{2}\left(\frac{1}{n_{2}}-\frac{1}{N}\right) S_{y}^{2}
$$

If the size of the population is large and if the variables $x$ and $y$ have dispersions that are close to one another, the gain (or loss) of precision due to co-ordination in relation to the selection of two samples independently is

$$
\begin{equation*}
G=\frac{\operatorname{var}(\hat{\Delta})}{\operatorname{var}_{\mathrm{IND}}(\hat{\Delta})} \approx 1-\rho q, \tag{2}
\end{equation*}
$$

where $\rho$ is the coefficient of correlation between $x$ and $y, \rho=S_{x y} / S_{x} S_{y}$ and $q$ is the overlap rate, $q=$ $2 E\left(n_{C}\right) /\left(n_{1}+n_{2}\right)$. Expression (2) provides a simple multiplicative coefficient serving to take account of the effect of correlation and overlap.

### 2.1.3 Estimation of the variance of $\hat{\Delta}$

To estimate the variance, two cases must be considered:

- if $E\left(n_{C}\right)$ is known, which may be the case (for example, when the two samples are known to be independent), then

$$
\begin{align*}
\widehat{\operatorname{var}}(\hat{\Delta}) & =N^{2}\left(\frac{1}{n_{1}}-\frac{1}{N}\right) s_{x 1}^{2}+N^{2}\left(\frac{1}{n_{2}}-\frac{1}{N}\right) s_{y 2}^{2} \\
& -2 N^{2}\left(\frac{E\left(n_{C}\right)}{n_{1} n_{2}}-\frac{1}{N}\right) s_{x y C} . \tag{3}
\end{align*}
$$

where

$$
s_{x 1}^{2}=\frac{1}{n_{1}-1} \sum_{s_{1}}\left(x_{k}-\bar{x}_{1}\right)^{2}, s_{y 2}^{2}=\frac{1}{n_{2}-1} \sum_{s_{2}}\left(y_{k}-\bar{y}_{2}\right)^{2},
$$

and

$$
s_{x y C}=\frac{1}{n_{C}-1} \sum_{s_{C}}\left(x_{k}-\bar{x}_{C}\right)\left(y_{k}-\bar{y}_{C}\right) .
$$

This estimator is unbiased, but it can sometimes take on negative values;

- if $E\left(n_{C}\right)$ is not known, the only information concerning co-ordination is $n_{C}$.

$$
\begin{align*}
\widehat{\operatorname{var}}(\hat{\Delta}) & =N^{2}\left(\frac{1}{n_{1}}-\frac{1}{N}\right) s_{x 1}^{2}+N^{2}\left(\frac{1}{n_{2}}-\frac{1}{N}\right) s_{y 2}^{2} \\
& -2 N^{2}\left(\frac{n_{C}}{n_{1} n_{2}}-\frac{1}{N}\right) s_{x y C} . \tag{4}
\end{align*}
$$

This estimator is unbiased conditional on $n_{C}$ and is therefore also unconditionally unbiased. It can also sometimes take on negative values. We will see further on that in some applications involving nonresponse, $E\left(n_{C}\right)$ is not known.

To use estimator (3), it is necessary to have at least two units in the overlap of the samples $\left(n_{C} \geq 2\right)$, unless $E\left(n_{C}\right)=n_{1} n_{2} / N$. If $E\left(n_{C}\right)=n_{1} n_{2} / N$, which is the case where the two samples are independent, the third term of estimator (3) is nil. As to estimator (4), it is not defined when $n_{C}=1$, unless $n_{1} n_{2}=N$.

### 2.2 Estimation using the common portion

The difference can also be estimated using only the common portion of the sample, which yields the estimator

$$
\hat{\Delta}_{C}=N\left(\bar{y}_{C}-\bar{x}_{C}\right),
$$

with $\quad \bar{y}_{C}=1 / n_{C} \sum_{k \in s_{C}} y_{k} \quad$ and $\quad \bar{x}_{C}=1 / n_{C} \sum_{k \in s_{C}} x_{k}$. This estimator is unbiased unconditionally and conditionally on $n_{C}$.

### 2.2.1 Estimation of the variance of $\hat{\boldsymbol{\Delta}}_{C}$

The conditional variance of $\hat{\Delta}_{C}$ is equal to

$$
\operatorname{var}\left(\hat{\Delta}_{C} \mid n_{C}\right)=N^{2}\left(\frac{1}{n_{C}}-\frac{1}{N}\right)\left(S_{y}^{2}+S_{x}^{2}-2 S_{x y}\right)
$$

The unconditional variance is equal to

$$
\operatorname{var}\left(\hat{\Delta}_{C}\right)=N^{2}\left[E\left(\frac{1}{n_{C}}\right)-\frac{1}{N}\right]\left(S_{y}^{2}+S_{x}^{2}-2 S_{x y}\right)
$$

This unconditional variance may be difficult to calculate when $n_{C}$ is random.

### 2.2.2 Comparison of the variances of $\hat{\Delta}$ and $\hat{\Delta}_{C}$

If we want to compare the two estimators of the difference, we can calculate

$$
\begin{aligned}
& \operatorname{var}(\hat{\Delta})-\operatorname{var}\left(\hat{\Delta}_{C}\right)=N^{2}\left[\frac{1}{n_{1}}-E\left(\frac{1}{n_{C}}\right)\right] S_{y}^{2} \\
& \quad+N^{2}\left[\frac{1}{n_{2}}-E\left(\frac{1}{n_{C}}\right)\right] S_{x}^{2}-2 N^{2}\left[\frac{E\left(n_{C}\right)}{n_{1} n_{2}}-E\left(\frac{1}{n_{C}}\right)\right] S_{x y} .
\end{aligned}
$$

If $n_{1}=n_{2}=n, S_{x}^{2}=S_{y}^{2}=S^{2}$, and $E\left(1 / n_{C}\right) \approx 1 / E\left(n_{C}\right)$, then we obtain

$$
\begin{aligned}
& \operatorname{var}(\hat{\Delta})-\operatorname{var}\left(\hat{\Delta}_{C}\right) \\
& \approx \frac{1}{q n}[q-1] 2 N^{2} S^{2}-2 \frac{1}{q n}\left[q^{2}-1\right] \rho N^{2} S^{2} \\
&=\frac{2 N^{2} S^{2}}{q n}(1-q)[\rho(1+q)-1],
\end{aligned}
$$

where $q=2 E\left(n_{C}\right) /\left(n_{1}+n_{2}\right)$ is the overlap rate. The estimator $\hat{\Delta}_{C}$ is therefore more precise than $\hat{\Delta}$ if

$$
\rho \geq \frac{1}{1+q}
$$

For example, if $q=0.7$, it is preferable to use only the common portion once $\rho \geq 1 /(1+0.7) \approx 0.588$ (on this subject, see Caron and Ravalet 2000, page 346). In cases where the overlap is sizable and the correlation is high, the estimator based on the difference of the cross-sectional estimators is therefore not very relevant.

## 3. Taking unit non-response into account

Non-response is considered to be independent of the selection design. According to the model, each unit decides randomly whether or not to respond, and the probabilities of response are equal between units. This is the most elementary model. However, if a unit does not respond in the first wave, it is highly probable that it will also not respond in the second wave. The model takes this dependency into account by considering separately four cases:

- the unit responds in both the first wave and the second;
- the unit responds in the first wave but not in the second;
- the unit does not respond in the first wave but it responds in the second;
- the unit responds in neither the first wave nor the second.

Non-response is commonly modelled by a multivariate Bernoullian design, which means that the probability of responding is the same for all statistical units and also that one unit decides to respond independently of the response of the other units. The non-response design is as follows:

$$
q\left(r_{A}, r_{B}, r_{C}, r_{D}\right)=\phi_{A}^{\operatorname{card}_{A}} \phi_{B}^{\operatorname{card}_{B}} \phi_{C}^{\operatorname{card}_{C}} \phi_{D}^{\operatorname{card}_{D}}
$$

where $r_{A}, r_{B}, r_{C}, r_{D} \subset U$, and $r_{A}, r_{B}, r_{C}, r_{D}$ are mutually exclusive, and where

- $\phi_{A}^{\mathrm{card}_{A}}$ is the probability of responding in wave 1 but not in wave 2 ;
- $\phi_{B}^{\operatorname{cardr}_{B}}$ is the probability of responding in wave 2 but not in wave 1 ;
- $\phi_{C}^{\operatorname{card}_{C}}$ is the probability of responding in both wave 1 and wave 2 ;
- $\phi_{D}^{\mathrm{card} r_{D}}$ is the probability of responding in neither wave 1 nor wave 2 .

The modelled non-response phase thus consists in selecting four disjoint samples according to Bernoullian designs with different intensities. Since it is assumed to be independent of the sampling design, conditional on the sample sizes observed, the design resulting from the selection and the non-response is a simple multivariate design. If inference is conducted conditional on the sample sizes, the estimation of probabilities $\phi_{A}, \phi_{B}, \phi_{C}, \phi_{D}$ is not necessary and an unbiased inference can be conducted, as if dealing with a simple design. The theory of the preceding section therefore applies directly to the respondents, and all the information on the overlap of the two samples is found in $\left|s_{C}\right|$, regardless of whether this overlap is due to the design or to the link that exists between non-responses in the two waves. Note that even if the model is fairly simple, it takes account of the fact that if a unit has not responded in one wave, it will probably be less likely to respond in the following wave. Also, this model will be applied in relatively small, homogeneous strata.

## 4. Other measures of changes over time

The measurement of change over time is not always expressed in terms of differences. Such change is often measured in the form of a quotient or a relative difference. We therefore consider the following three measures:

- the difference $\hat{\Delta}=\hat{Y}_{2}-\hat{X}_{1}$;
- the relative change $\hat{\Delta}_{R}=\left(\hat{Y}_{2}-\hat{X}_{1}\right) / \hat{X}_{1}=\hat{Y}_{2} / \hat{X}_{1}-1$;
- the quotient $\hat{Q}=\hat{Y}_{2} / \hat{X}_{1}$.

The variance of $\hat{\Delta}$ may be expressed simply as a function of the estimators of variance of $\hat{Y}_{2}$ and $\hat{X}_{1}$ and the estimator of their covariance (see expression 4). The variance of $\hat{\Delta}_{R}$ is equal to the variance of $\hat{Q}$. They may be approached and then estimated using a residuals technique (on this subject, see Woodruff 1971; Binder and Patak 1994; Deville and Särndal 1992; Deville 1999),

$$
\begin{aligned}
\widehat{\operatorname{var}}\left(\hat{\Delta}_{R}\right) & =\widehat{\operatorname{var}}(\hat{Q}) \\
& =\frac{1}{\hat{X}_{1}^{2}}\left[\widehat{\operatorname{var}}\left(\hat{Y}_{2}\right)+\hat{Q}^{2} \widehat{\operatorname{var}}\left(\widehat{X}_{1}\right)-2 \hat{Q} \widehat{\operatorname{cov}}\left(\widehat{X}_{1}, \hat{Y}_{2}\right)\right] .
\end{aligned}
$$

This variance can thus be simply estimated once we have estimators of $\operatorname{var}\left(\hat{Y}_{2}\right), \operatorname{var}\left(\hat{X}_{1}\right)$ and $\operatorname{cov}\left(\hat{X}_{1}, \hat{Y}_{2}\right)$.

## 5. Ratio estimation and robustification

Two techniques are commonly used for estimating the results of sample surveys: the use of a ratio estimator to
calibrate on the total of a dummy variable, and robustification of the estimators. These techniques must be taken into account in determining the precision of the final results.

### 5.1 Calibration

If an estimator is calibrated on known totals, the variance may be estimated simply by a residuals technique (see Woodruff 1971; Binder and Patak 1994; Deville and Särndal 1992; Deville 1999). For example, if $\mathbf{z}_{k 1}$ and $\mathbf{z}_{k 2}$ are column vectors of dummy variables on which the estimators $\hat{X}_{1 \text { Cal }}$ and $\hat{Y}_{2 \text { Cal }}$ are calibrated in waves 1 and 2, then the variances can be estimated by a residuals technique: $\operatorname{var}\left(\hat{X}_{1 \mathrm{Cal}}\right) \approx \operatorname{var}\left(\hat{E}_{1}\right)$ and $\operatorname{var}\left(\hat{Y}_{2 \mathrm{Cal}}\right) \approx \operatorname{var}\left(\hat{E}_{2}\right)$, where $\hat{E}_{1}$ et $\hat{E}_{2}$ are Horvitz-Thompson estimators of the totals of the residuals, with the latter being given for a simple design and for the generalized regression estimator by:

$$
\begin{aligned}
& e_{k 1}=x_{k}-\mathbf{z}_{k 1}^{\prime} \hat{\mathbf{B}}_{1}, \\
& e_{k 2}=y_{k}-\mathbf{z}_{k 2}^{\prime} \hat{\mathbf{B}}_{2},
\end{aligned}
$$

with

$$
\begin{aligned}
& \hat{\mathbf{B}}_{1}=\left(\sum_{k \in s_{1}} q_{k 1} \mathbf{z}_{k 1} \mathbf{z}_{k 1}^{\prime}\right)^{-1} \sum_{k \in s_{1}} q_{k 1} \mathbf{z}_{k 1} x_{k 1} \\
& \hat{\mathbf{B}}_{2}=\left(\sum_{k \in s_{2}} q_{k 2} \mathbf{z}_{k 2} \mathbf{z}_{k 2}^{\prime}\right)^{-1} \sum_{k \in s_{2}} q_{k 2} \mathbf{z}_{k 2} y_{k 2}
\end{aligned}
$$

where $q_{k j}, j=1,2$, is a coefficient that serves to take account of possible heteroscedasticity.

In the case of a sampling design with unequal probabilities, e.g., a stratified sampling design such as in the Swiss survey of value added, the residuals are obtained by using a weighted regression. It is sufficient to replace $\hat{\mathbf{B}}_{1}$ and $\hat{\mathbf{B}}_{2}$ respectively by

$$
\begin{align*}
& \hat{\mathbf{B}}_{1}=\left(\sum_{k \in s_{1}} \frac{q_{k 1} \mathbf{z}_{k 1} \mathbf{z}_{k 1}^{\prime}}{\pi_{k 1}}\right)^{-1} \sum_{k \in s_{1}} \frac{q_{k 1} \mathbf{z}_{k 1} x_{k 1}}{\pi_{k 1}}, \quad \text { and }  \tag{5}\\
& \hat{\mathbf{B}}_{2}=\left(\sum_{k \in s_{2}} \frac{q_{k 2} \mathbf{z}_{k 2} \mathbf{z}_{k 2}^{\prime}}{\pi_{k 2}}\right)^{-1} \sum_{k \in s_{2}} \frac{q_{k 2} \mathbf{z}_{k 2} y_{k 2}}{\pi_{k 2}}, \tag{6}
\end{align*}
$$

where $\pi_{k j}$ is the probability of inclusion of unit $k$ in the sample for wave $j, j=1,2$.

### 5.2 Robustification

It is often useful to apply a robustification technique which offers a way to treat outliers. Simply consider that outliers have been detected and the weights of the individuals whose values are considered outliers have been modified by a factor $u_{k j}(s)$ in wave $j$. This factor is included between 0 and 1 and is equal to 1 for units that have values considered normal. The variance of the robustified estimator can be approached by advancing the
classical hypothesis that weights $u_{k j}(s)$ depend only slightly on the sample $s$ that was drawn (see Hulliger 1999). All that is needed, then, is to replace the variables $x_{k}$ and $y_{k}$ observed by $u_{k 1} x_{k}$ and $u_{k 2} y_{k}$ in the variance estimators.

By bringing together all the components of the mean square error of a change over time so as to take account of all components of that variance - namely the design, the panel effect, non-response, calibration and robustification we obtain, for the relative change in a stratum,

$$
\begin{align*}
& \widehat{\operatorname{MSE}}\left(\hat{\Delta}_{R}\right)=\widehat{\operatorname{MSE}}(\hat{Q})= \\
& \frac{1}{\hat{X}_{1}}\left[\widehat{\operatorname{var}}\left(\widehat{E U}_{1}\right)+\hat{Q}^{2} \widehat{\operatorname{var}}\left(\widehat{E U}_{1}\right)-2 \hat{Q} \widehat{\operatorname{cov}}\left(\widehat{E U}_{1}, \widehat{E U} 2\right)\right], \tag{7}
\end{align*}
$$

where

$$
\begin{aligned}
& \hat{X}_{1}=\frac{N}{m_{1}} \sum_{R_{1}} x_{k}, \hat{Y}_{2}=\frac{N}{m_{2}} \sum_{R_{2}} y_{k}, \hat{Q}=\frac{\hat{Y}_{2}}{\hat{X}_{1}}, \\
& e u_{k 1}=u_{k 1} x_{k}-u_{k 1} \mathbf{z}_{k 1}^{\prime} \hat{\mathbf{B}}_{1}, \\
& e u_{k 2}=u_{k 2} y_{k}-u_{k 2} \mathbf{z}_{k 2}^{\prime} \hat{\mathbf{B}}_{2}, \\
& \widehat{E U_{j}}=\frac{N}{m_{j}} \sum_{R_{j}} e u_{k j}, \overline{E U}_{j}=\frac{\widehat{E U_{j}}}{N}, j=1,2, \\
& \hat{\mathbf{B}}_{1}=\left(\sum_{k \in D_{1}} \frac{q_{k 1} u_{k 1}^{2} \mathbf{z}_{k 1} \mathbf{z}_{k 1}^{\prime}}{\pi_{k 1}}\right)^{-1} \sum_{k \in D_{1}} \frac{q_{k 1} u_{k 1}^{2} \mathbf{z}_{k 1} x_{k}}{\pi_{k 1}}, \\
& \hat{\mathbf{B}}_{2}=\left(\sum_{k \in D_{2}} \frac{q_{k 2} u_{k 2}^{2} \mathbf{z}_{k 2} \mathbf{z}_{k 2}^{\prime}}{\pi_{k 2}}\right)^{-1} \sum_{k \in D_{2}} \frac{q_{k 2} u_{k 2}^{2} \mathbf{z}_{k 2} y_{k}}{\pi_{k 2}} . \\
& \widehat{\operatorname{var}}\left(\widehat{E U_{j}}\right)= \\
& N^{2}\left(\frac{1}{m_{j}}-\frac{1}{N}\right) \frac{1}{m_{j}-1} \sum_{R_{j}}\left(e u_{k j}-\overline{E U}_{j}\right)^{2}, j=1,2, \\
& \widehat{\operatorname{cov}}\left(\widehat{E U_{1}}, \widehat{E U_{2}}\right)= \\
& N^{2}\left(\frac{m_{C}}{m_{1} m_{2}}-\frac{1}{N}\right) \frac{1}{m_{C}-1} \sum_{R_{C}}\left(e u_{k 1}-\overline{E U_{1}}\right) \\
& \times\left(e u_{k 2}-\overline{E U_{2}}\right) \text {. }
\end{aligned}
$$

$R_{1}$ and $R_{2}$ designate the set of respondents in the first and the second waves in the stratum, $m_{1}=\left|R_{1}\right|, m_{2}=\left|R_{2}\right|$, $\mid R_{C}=R_{1} \cap R_{2}$ and $m_{C}=\left|R_{1} \cap R_{2}\right| . D_{1}$ and $D_{2}$ are the sets of respondents in the two waves in the domain in which the calibration was carried out.

## 6. The Swiss survey of value added

### 6.1 Description of survey

The Swiss survey of value added is a survey of companies, conducted annually. Its purpose is to provide estimators of the main parameters of output in Switzerland: the value of gross output, the amount of intermediate consumption, the value added created by companies, and the cost of labour. The sampling design used is a stratified sampling of companies. In 1999, a sample of 11,210 companies (employing at least two persons) was selected and surveyed. This sample was run again in 2000 and 2001. Over that period, then, this is a panel survey. In the absence of a business register making it possible to identify births and deaths, the population of companies was considered constant during this period. The only adjustment to the annual data is made using a ratio estimation on the total of full-time equivalents (FTEs) per activity domain, available from an external source.

Stratification is based on the first two digits of the Nomenclature Générale des Activités économiques (general classification of economic activities) (NOGA2) and the size of the company (see Renfer 2000). In each activity stratum, three size strata are created: small companies employing 219 persons in FTE, medium-size companies, from 20 to $M$ FTE, and large companies of more than $M$ FTE. The stratum containing large companies is a take-all stratum, while small and medium-size companies are selected randomly with different sampling rates. The boundary $M$ is chosen differently in each activity stratum in order to obtain optimum precision. In these three waves, approximately 6,000 establishments responded. The response rate for large companies, which all had to be surveyed, was close to $71 \%$ and was higher than the rate for small and medium-size companies. It was decided after the fact to treat some very large companies separately according to the "surprise" stratum methodology of Hidiroglou and Srinath (1981), considering that the response rate for the largest companies may well be better because they have an administrative structure better suited to responding to the survey questions. If they were assigned a weight equal to that of other large companies, this would introduce a bias as well as excessive variability. The "surprise" poststrata contain the $5 \%$ largest companies in the survey file. The latter were then considered as having, in effect, all been surveyed, and they received a weight of 1 . No other treatment (calibration, robustification) was applied to them. The take-some strata consisting of small, medium-size and large companies were updated and some strata (size classes) containing few companies were later collapsed. If we accept the hypothesis that the very large companies were all taken, then the resulting estimator is unbiased and the variance related to
very large companies is nil. We can therefore calculate only the variance in the other, updated strata.

During the survey, companies were again asked their category of economic activity. The estimates are based on these reported NOGA2s not on the NOGA2s in the sample frame. A calibration on the number of full-time equivalents (FTEs) provided by the business register is then conducted using a quotient estimator for the "reported" NOGA2 domains.

Finally, a robustification technique was used to lop the distribution of certain variables in the sample of small, medium-size and large companies (see Hulliger 1999; Peters, Renfer and Hulliger 2001). The weights of establishments whose values are considered outliers were modified by a factor $u_{k j}(s)$ included between 0 and 1 . This factor is equal to 1 for companies that have values that are considered normal.

### 6.2 Variance of the change in value added

The objective is to estimate correctly the variance of estimators of change in value added (see Renfer 2000; Peters et al. 2001). In computing variances according to the hypothesis of independence of the samples, we largely overestimate the variance of changes, because the "value added" variables in times $t_{1}$ and $t_{2}$ are positively correlated. Correctly taking account of all aspects of the sampling design and the adjustment should provide better variance estimates. The study focuses on the 1999, 2000 and 2001 waves of the survey. Between these three dates, the raw sample was not modified. The fact that the sample remained fixed should make it possible to reliably estimate changes, but a response rate hovering around $50 \%$ may cause us to lose the benefit of the panel, if the number of respondents common to successive waves is low. The case of change between two survey waves where the sample has been updated, and where there are therefore two different raw samples and reference populations, is an entirely different problem.

In the present case, the fact that low variances were obtained can be attributed to the combined effect of several factors:

1. Optimal design: The sampling design was optimized. According to the optimal stratification, large companies have higher probabilities of inclusion. The stratum of companies contributing the most to value added is a take-all stratum. For this reason, the cross-sectional estimators have a low variance.
2. High response fraction: In the take-all stratum of large companies, the response rate approaches $70 \%$. The finite population correction $(N-n) / N$
can therefore divide the variance by 3 compared to the case of an infinite population.
3. Panel effect: The sample is a panel, which is the best strategy for estimating changes over time.
4. Correlation of non-response: The non-response in one wave is strongly related to the previous wave and therefore does not greatly degrade the panel.
5. Correlation of variables between waves: The value added variables in times $t$ and $t+1$ are highly correlated, since they are the same variable estimated at two different points in time.
6. Calibration: The estimators are calibrated in the strata on a variable related to the variable of interest; the variance of the estimators can then be written as a residual variance.

Of the 11,210 companies selected in 1999, approximately 5,200 responded in 1999 and 2000, and 5,300 responded in the 2000 and 2001 waves. Thus the size of the panel is relatively modest, and the treatment of nonresponse will therefore have a major impact on the results. To make variance estimates, we have assumed that nonresponse is ignorable (missing completely at random) within the take-some strata.

In each wave, estimates are made in the reported NOGA2 domains. This implies the possibility of a change of domain on the part of companies, and it is necessary to try to factor this into longitudinal estimates. We decided to ignore the impact of these changes initially, and to consider for the estimation of covariance that the domains are fixed and given by the value reported in the first of the two consecutive waves. This simplification is not inappropriate, since only 30 companies changed domain between 1999 and 2000, and only 25 did so between 2000 and 2001, representing respectively less than $0.5 \%$ and $0.2 \%$ of the FTEs in the sample. Calibration is carried out each year, and it can be taken into account using a residuals technique. As with estimating the variance of the cross-sectional estimators, robustification is taken into account by reweighting the survey variables.

With realistic assumptions, all components of the variance may be taken into account by means of the general expression (7). This expression is applied within each stratum and it covers all the components of the survey of value added: the panel effect, non-response, stratification, calibration and robustification. The estimators for the survey of value added are ratio estimators, and in this case the calculation of residuals is simplified. This is because in the case of the ratio, the regression coefficients given in (5) and (6) are calculated having only one dummy variable, and therefore $\mathbf{z}_{k j}=z_{k j}$ is scalar. Also, we take $q_{k j}=1 / z_{k j}$, for
$j=1,2$, and with robustification taken into account, we thus obtain:

$$
\begin{aligned}
& e u_{k 1}=u_{k 1} x_{k}-\hat{B}_{1} u_{k 1} z_{k 1}, \\
& e u_{k 2}=u_{k 2} y_{k}-\hat{B}_{2} u_{k 2} z_{k 2},
\end{aligned}
$$

where

$$
\begin{aligned}
& \hat{B}_{1}=\frac{\sum_{D_{1}} u_{k 1} x_{k} / \pi_{k 1}}{\sum_{D_{1}} u_{k 1} z_{k 1} / \pi_{k 1}} \\
& \hat{B}_{2}=\frac{\sum_{D_{2}} u_{k 2} y_{k} / \pi_{k 2}}{\sum_{D_{2}} u_{k 2} z_{k 2} / \pi_{k 2}}
\end{aligned}
$$

### 6.3 Variance estimation of changes

We made estimates of the standard deviations of changes in gross output values and value added figures calculated by the Swiss Federal Statistical Office. These estimates take into consideration all the aspects described above. We compared them with the estimated standard deviations that would have been obtained under the assumption that the draws for the different waves are independent. Over the various activity strata, the standard deviations that take account of the correlation between the survey waves are $41 \%$ lower than those based on the assumption of independence. This makes it possible to have much smaller confidence intervals than those calculated before this study, which were more quickly obtained but less precise. However, the gain is not the same in all activity strata. The following tables show standard deviations (SDs), calculated for the five largest activity strata (NOGA), of changes over time in the value of gross output ( $\Delta O V$ ) and of value added ( $\Delta V A$ ) between 1999 and 2000. The standard deviation that would have been obtained by ignoring the correlation between samples $\left(S D_{\text {ind }}\right)$ is also included in the tables, along with the "gain" in precision realized by taking this correlation into account.

Table 1
Change in gross output value between 1999 and 2000 and standard deviations (in billions of Swiss francs)

| Stratum | $\boldsymbol{\Delta} \boldsymbol{O} \boldsymbol{V}$ | $\boldsymbol{S} \boldsymbol{D}_{\text {ind }}$ | $\mathbf{S D}$ | Gain (\%) |
| :---: | ---: | :---: | :---: | :---: |
| 1 | 3.31 | 2.35 | 0.87 | 63 |
| 2 | -0.77 | 4.38 | 1.98 | 55 |
| 3 | 3.07 | 2.11 | 0.94 | 56 |
| 4 | 4.33 | 1.10 | 1.00 | 09 |
| 5 | -0.09 | 0.81 | 0.53 | 35 |

Table 2
Change in value added between 1999 and 2000 standard deviations (in billions of Swiss francs)

| Stratum | $\boldsymbol{\Delta} \boldsymbol{V} \boldsymbol{A}$ | $\boldsymbol{S D}_{\text {ind }}$ | SD | Gain (\%) |
| :---: | :---: | :---: | :---: | :---: |
| 1 | 1.96 | 0.91 | 0.32 | 65 |
| 2 | 0.68 | 2.99 | 1.04 | 65 |
| 3 | 1.90 | 1.47 | 0.72 | 51 |
| 4 | 0.36 | 0.47 | 0.45 | 05 |
| 5 | -0.36 | 0.59 | 0.43 | 27 |

## Acknowledgements

This study was carried out under an agreement between the University of Neuchâtel and the Swiss Federal Statistical Office. The findings published in this article are those of the authors alone and in no case do they commit the Federal Statistical Office. We wish to thank Paul-André Salamin for his contribution to this study.

## Appendix

## Demonstration of proposition 1

It is well known that

$$
\operatorname{var}\left(\hat{X}_{1}\right)=N^{2}\left(\frac{1}{n_{1}}-\frac{1}{N}\right) S_{x}^{2}
$$

and

$$
\operatorname{var}\left(\hat{Y}_{2}\right)=N^{2}\left(\frac{1}{n_{2}}-\frac{1}{N}\right) S_{y}^{2} .
$$

It is thus sufficient to calculate $\operatorname{cov}\left(\hat{X}_{1}, \hat{Y}_{2}\right)$. We note

$$
\begin{array}{ll}
\bar{x}_{A}=\frac{1}{n_{A}} \sum_{k \in s_{A}} x_{k}, & \bar{x}_{C}=\frac{1}{n_{C}} \sum_{k \in s_{C}} x_{k}, \\
\bar{y}_{B}=\frac{1}{n_{B}} \sum_{k \in s_{B}} y_{k}, & \bar{y}_{C}=\frac{1}{n_{C}} \sum_{k \in s_{C}} y_{k}, \\
\bar{x}_{1}=\frac{n_{A} \bar{x}_{A}+n_{C} \bar{x}_{C}}{n_{1}}, & \bar{y}_{2}=\frac{n_{B} \bar{y}_{B}+n_{C} \bar{y}_{C}}{n_{2}},
\end{array}
$$

and therefore $\hat{X}_{1}=N \bar{x}_{1}$ and $\hat{Y}_{2}=N \bar{y}_{2}$. We must still calculate

$$
\begin{aligned}
\operatorname{cov}\left(\bar{x}_{1}, \bar{y}_{2}\right)= & E \operatorname{cov}\left(\bar{x}_{1}, \bar{y}_{2} \mid n_{A}, n_{B}, n_{C}\right) \\
& +\operatorname{cov}\left[E\left(\bar{x}_{1} \mid n_{A}, n_{B}, n_{C}\right), E\left(\bar{y}_{2} \mid n_{A}, n_{B}, n_{C}\right)\right] .
\end{aligned}
$$

Since $\bar{x}_{1}$ and $\bar{y}_{2}$ are unbiased conditional on $n_{A}, n_{B}$, and $n_{C}$,

$$
\operatorname{cov}\left[E\left(\bar{x}_{1} \mid n_{A}, n_{B}, n_{C}\right), E\left(\bar{y}_{2} \mid n_{A}, n_{B}, n_{C}\right)\right]=\operatorname{cov}(\bar{X}, \bar{Y})=0
$$

We therefore obtain

$$
\operatorname{cov}\left(\bar{x}_{1}, \bar{y}_{2}\right)=E \operatorname{cov}\left(\bar{x}_{1}, \bar{y}_{2} \mid n_{A}, n_{B}, n_{C}\right)
$$

Conditional on $n_{A}, n_{B}$, and $n_{C}$, we are in case A of Tam (1984, theorem 1). The conditional variance is equal to

$$
\operatorname{cov}\left(\bar{x}_{1}, \bar{y}_{2} \mid n_{A}, n_{B}, n_{C}\right)=\left(\frac{n_{C}}{n_{1} n_{2}}-\frac{1}{N}\right) S_{x y}
$$

and therefore

$$
\operatorname{cov}\left(\bar{x}_{1}, \bar{y}_{2}\right)=\left(\frac{E\left(n_{C}\right)}{n_{1} n_{2}}-\frac{1}{N}\right) S_{x y}
$$

Now,

$$
\operatorname{cov}\left(\hat{X}_{1}, \hat{Y}_{2}\right)=N^{2} \operatorname{cov}\left(\bar{x}_{1}, \bar{y}_{2}\right)
$$

enabling us to obtain the result (1).

## References

Ardilly, P., and Tillé, Y. (2003). Exercices corrigés de méthodes de sondage. Paris: Ellipses.

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. Canadian Journal of Statistics, 32, 4, 451467.

Binder, D.A., and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. Journal of the American Statistical Association, 89, 1035-1043.

Caron, N., and Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête Emploi en continu. Technical report, 0005. Méthodologie Statistique, INSEE, Paris.
Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. Survey Methodology, 25, 193-203.
Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. Survey Methodology, 27, 45-51.
Goga, C. (2003). Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques. Ph.D. Dissertation, Université de Rennes II, Haute Bretagne, France.
Hidiroglou, M., Särndal, C.-E. and Binder, D. (1995). Weighting and Estimation in Business Surveys. Business Survey Methods, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge and P.S. Kott), New York: John Wiley \& Sons, Inc., 477-502.
Hidiroglou, M.A., and Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. Journal of the American Statistical Association, 76, 690695.

Holmes, D.J., and Skinner, C.J. (2000). Variance Estimation for Labour Force Survey Estimates of Level and Change. Technical report, Government Statistical Service Methodology Series, 21, London, England.

Hulliger, B. (1999). Simple and robust estimators for sampling. Proceedings of the Section on Survey Research Methods, American Statistical Association, 54-63.
Kish, L. (1965). Survey Sampling. New York: John Wiley \& Sons, Inc.
Laniel, N. 1988. (1988). Variances for a rotating sample from a changing population. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 246-250.

Nordberg, L. (2000). On variance estimation for measure of change when samples are coordinated by the use of permanent random numbers. Journal of Official Statistics, 16, 363-378.
Peters, R., Renfer, J.-P. and Hulliger, B. (2001). Statistique de la valeur ajoutée : procédure d'extrapolation des données. Technical report, Swiss Federal Statistical Office.

Renfer, J.-P. (2000). Enquête sur la production et la valeur ajoutée : échantillonnage complémentaire. Technical report, Swiss Federal Statistical Office.

Sen, A.R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics, 29, 381385.

Tam, S.M. (1984). On covariances from overlapping samples. The American Statistician, 38, (4), 288-289.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Spinger-Verlag.

Woodruff, R.S. (1971). A simple method for approximating de variance of a complicated estimate. Journal of the American Statistical Association, 66, 411-414.

# PSU masking and variance estimation in complex surveys 

Inho Park ${ }^{1}$


#### Abstract

The analysis of stratified multistage sample data requires the use of design information such as stratum and primary sampling unit (PSU) identifiers, or associated replicate weights, in variance estimation. In some public release data files, such design information is masked as an effort to avoid their disclosure risk and yet to allow the user to obtain valid variance estimation. For example, in area surveys with a limited number of PSUs, the original PSUs are split or/and recombined to construct pseudo-PSUs with swapped second or subsequent stage sampling units. Such PSU masking methods, however, obviously distort the clustering structure of the sample design, yielding biased variance estimates possibly with certain systematic patterns between two variance estimates from the unmasked and masked PSU identifiers. Some of the previous work observed patterns in the ratio of the masked and unmasked variance estimates when plotted against the unmasked design effect. This paper investigates the effect of PSU masking on variance estimates under cluster sampling regarding various aspects including the clustering structure and the degree of masking. Also, we seek a PSU masking strategy through swapping of subsequent stage sampling units that helps reduce the resulting biases of the variance estimates. For illustration, we used data from the National Health Interview Survey (NHIS) with some artificial modification. The proposed strategy performs very well in reducing the biases of variance estimates. Both theory and empirical results indicate that the effect of PSU masking on variance estimates is modest with minimal swapping of subsequent stage sampling units. The proposed masking strategy has been applied to the 2003-2004 National Health and Nutrition Examination Survey (NHANES) data release.


Key Words: Disclosure control; Stratified multistage sampling; Subsequent stage sampling unit swapping; Design effect; Intracluster correlation coefficient (ICC); Sample mean.

## 1. Introduction

The analysis of stratified multistage sample data requires the use of design information such as stratum and primary sampling unit (PSU) identifiers, or associated replicate weights, in variance estimation. In large surveys, PSUs often consist of single or multiple counties. Some external sources that are publicly available such as Census data can provide extremely detailed PSU-level demographics. Even with their name suppressed, inclusion of PSU identifiers in public release data files alone can pose an identification risk by allowing their linkage to external sources. Thus, PSU identifiers are often masked as an effort (1) to reduce the risk of data disclosure and (2) yet to allow the user to obtain valid variance estimation. Mayda, Mohl and Tambay (1996) addressed the potential risk of data disclosure that is associated with the inclusion of the original PSU identifiers in the public release data files and considered the stratumcollapsing method by Rust (1986) for balancing out the aforementioned two needs. Due to a potential inconsistency of the variance estimation under the stratum-collapsing method indicated by Valliant (1996), Yung (1997) suggested constructing a set of average bootstrap replicate weights. Lu (2004) demonstrated that supplying replicate weights and giving the stratum and PSU identifiers are practically equivalent in the viewpoint of confidentiality, since one can be easily obtained from the others. Shah
(2001) discussed ways to create pseudo-strata and pseudoPSUs given a set of balanced repeated replication weights. Eltinge (1999) proposed a method similar to the stratumcollapsing methods. Lu, Brick and Sitter (2006) also established conditions for the consistency of the variance estimator under the stratum-collapsing method and also proposed stratum-grouping algorithms yielding efficient and consistent stratum-collapsed variance estimators.

With a limited number of PSUs in the sample, the stratum-collapsing method is not appealing due to insufficient degrees of freedom for variance estimation. Dohrmann, Curtin, Mohadjer, Montaquila and Le (2002), Dohrmann, Lu, Park, Sitter and Curtin (2005) dealt with such situations and considered two PSU masking methods. The first method splits each PSU into two pseudo-PSUs (sets of ultimate sampling units within the PSUs), arbitrarily doubling degrees of freedom for variance estimation. The second method constructs the pseudo-PSUs by swapping second-stage sampling units (SSUs) between the original PSUs, retaining the original degrees of freedom for variance estimation. That is, the PSU and stratum assignments of all ultimate sampling units in one SSU are switched to those in the matched SSU. This method can be generalized so that the original PSUs are divided into one or more splits and are recombined to construct pseudo-PSUs with swapped PSU splits. This approach is different from data swapping (Dalenius and Reiss 1982), which is often used for

[^6]protecting confidentiality in a way that values of sensitive survey variables are switched among individual records. Because of the resulting distortion in the clustering structure of the sample design, the two PSU masking methods can result in biased variance estimates possibly with certain systematic patterns between two variance estimates from the unmasked and masked PSU identifiers. Dohrmann et al. (2005) observed decreasing funnel-shape curvature patterns in the ratio of the masked and unmasked variance estimates of a sample mean when plotted against the design effect. They explained such patterns based on an approximate relationship of the variance estimate that is monotone in the intracluster correlation coefficient (ICC) of Kish's design effect formula.

This paper focuses on the issues related to the second PSU masking method that swaps subsequent stage sampling units among the original PSUs and discusses its effect on variance estimates. Section 2 deals with the effect of PSU masking on the variance regarding aspects of the clustering structure such as ICC and means and sizes of PSU-splits for swapping under a single-stage cluster sample design. Section 3 investigates how the degree of swapping in PSU masking is related to the bias in the variance utilizing a parametric model for cluster sampling. Section 4 considers a PSU masking strategy through SSU swapping that helps reduce the PSU masking effects on variance estimates and thus the resulting biases under complex surveys. Section 5 briefly reviews the recent work by Dohrmann et al. (2002, 2005) and also presents application results of the proposed masking strategy to data from the National Health Interview Survey (NHIS) with some artificial modification. Finally, Section 6 includes some discussions.

## 2. Effect of distortion in clustering structure on variance of sample mean

Cluster sampling, often used in surveys for its cost and logistic reasons, is a major source of the increase in the variance of an estimator compared with a simple random sample due to the similarity of sampling units within the clusters. Standard sampling texts such as Särndal, Swensson and Wretman (1992, Section 8.7) provide formulae for the variance of a sample mean in terms of the ICC, cluster sizes and means of a survey variable $y$. It indicates that clustering in the sample design should reveal its impact on the variance through them. In this section, we examine how the distortion in the clustering structure of the sample design affects the variance of a sample mean when the PSUs are masked through swapping their splits between the two PSUs. For our discussion in this section, we consider a single-stage probability-proportional-to-size (PPS) sampling of PSUs. This sampling scheme is rather simple but still
complex enough to reveal the effect of PSU masking on the variance in relation to these three components.

### 2.1 Variance under single-stage PPS cluster sampling

Suppose that a population $U$ of $M$ units is grouped into $N$ PSUs of $M_{i}$ units each. A random sample of $n$ PSUs is drawn with probabilities $p_{i}\left(\sum_{i=1}^{N} p_{i}=1\right)$ and every unit in a sampled PSU is included in the sample. For simplicity, we assume the selection of PSUs is with replacement. The weighted sample mean $\hat{\bar{Y}}=\left(\sum_{i=1}^{n} \sum_{j=1}^{M_{i}} w_{i j}\right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_{i}} w_{i j} y_{i j}$ is an estimator of the population mean $\bar{Y}=M^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M_{i}} y_{i j}$ of survey variable $y$, where $w_{i j}=\left(n p_{i}\right)^{-1}$ and $y_{i j}$ denote the sampling weight and the value of $y$ for the $j^{\text {th }}$ unit of PSU $i$, respectively. Let $m=\sum_{i=1}^{n} M_{i}, S_{y}^{2}=(M-1)^{-1}$ $\sum_{i=1}^{N} \sum_{j=1}^{M_{i}}\left(y_{i j}-\bar{Y}\right)^{2}, \bar{Y}_{i}=M_{i}^{-1} \sum_{j=1}^{M_{i}} y_{i j}$ denote the sample size, the population variance and the PSU means of $y$, respectively. Assuming $N$ is large so that $N /(N-1) \doteq 1$, its approximate variance can be written as

$$
\begin{align*}
V(\hat{\bar{Y}} \mid S) \doteq & \doteq m^{-1} S_{y}^{2}\left[1+(\bar{M}-1) \rho_{y U}\right] \\
& +(m N)^{-1} \sum_{i=1}^{N} p_{i}^{-1} M_{i}\left(M_{i} M^{-1}-p_{i}\right)\left(\bar{Y}_{i}-\bar{Y}\right), \\
\doteq & (m M N)^{-1} \sum_{i=1}^{N} p_{i}^{-1} M_{i}^{2}\left(\bar{Y}_{i}-\bar{Y}\right)^{2}, \tag{1}
\end{align*}
$$

where $S$ denotes the sample index set, $\rho_{y U}=1-S_{y w}^{2} / S_{y}^{2}$ is the ICC and $S_{y w}^{2}=(M-N)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M_{i}}\left(y_{i j}-\bar{Y}_{i}\right)^{2}$ is the within-PSU mean square deviation. The derivation of (1) is given in the Appendix.

For a common special case of $p_{i} \propto M_{i}$, that is, PPS sampling, (1) is simplified as

$$
\begin{align*}
V(\hat{\bar{Y}} \mid S) & \doteq m^{-1} S_{y}^{2}\left[1+(\bar{M}-1) \rho_{y U}\right] \\
& \doteq(m N)^{-1} \sum_{i=1}^{N} M_{i}\left(\overline{Y_{i}}-\bar{Y}\right)^{2} \tag{2}
\end{align*}
$$

and the ICC is expressed as

$$
\begin{equation*}
\rho_{y U} \doteq(\bar{M}-1)^{-1}\left[\frac{V(\hat{\bar{Y}} \mid S)-m^{-1} S_{y}^{2}}{m^{-1} S_{y}^{2}}\right] . \tag{3}
\end{equation*}
$$

The second approximation in (2) indicates that PSUs with larger $M_{i}\left(\bar{Y}_{i}-\bar{Y}\right)^{2}$ contribute more to the variance. The ICC in (3) reveals the precision loss (in a rough sense) of per-cluster relative increase in the variance of $m^{-1} S_{y}^{2}$, the variance of the simple sample mean $\bar{y}=m^{-1} \sum_{(i j) \in S} y_{i j}$ that could have been obtained from the same sized withreplacement simple random samples.

A complex survey often involves the above single-stage PPS sampling or other (additional) complex design (e.g., stratification, multi-stage sampling and unequal selection probabilities) or estimation features (e.g., nonresponse
adjustments and calibration adjustments). For example, if $p_{i}$ had been disproportional to size or further subsampling had been involved to induce unequal weights, then the corresponding complex feature might have come into the picture in variance estimation. The associated impact on variance estimation of complex samples will be discussed in detail in Section 4.

### 2.2 Means and sizes of PSU-splits

To mask the PSUs, consider that the first two PSUs in the sample are each split into two sets of units, $U_{1}=$ $U_{11} \cup U_{12}$ and $U_{2}=U_{21} \cup U_{22}$ say, and the two pseudoPSUs, $\quad U_{1}^{*}=U_{11} \cup U_{22} \quad$ and $\quad U_{2}^{*}=U_{21} \cup U_{12} \quad$ are constructed by swapping $U_{12}$ and $U_{22}$ between the two PSUs. Let $S$ and $S^{*}$ denote the unmasked and masked sample index sets, respectively. Let $V\left(\hat{\bar{Y}} \mid S^{*}\right)$ denote the variance of $\hat{\bar{Y}}$ associated with the pseudo-PSUs (also assuming the other non-sampled PSUs in $U$ remain the same). Also, let $\bar{Y}_{i}^{*}, Y_{i}^{*}, M_{i}^{*}$ denote the mean and total of $y$ and the size of the $i^{\text {th }}$ pseudo-PSU, respectively. Assuming $\bar{Y}=0$ without loss of generality, the difference between the masked and unmasked variances is written from (2) as
$V\left(\hat{\bar{Y}} \mid S^{*}\right)-V(\hat{\bar{Y}} \mid S)$
$\doteq(m N)^{-1} \sum_{i=1,2}\left(M_{i}^{*} \bar{Y}_{i}^{* 2}-M_{i} \bar{Y}_{i}^{2}\right)$,
$\doteq(m N)^{-1} \sum_{i=1,2}\left(\sqrt{M_{i}^{*}} \bar{Y}_{i}^{*}+\sqrt{M_{i}} \bar{Y}_{i}\right)\left(\sqrt{M_{i}^{*}} \bar{Y}_{i}^{*}-\sqrt{M_{i}} \bar{Y}_{i}\right)$.
Expression (4) shows that the difference in variance due to PSU masking depends upon the changes in PSU quantities $\sqrt{M_{i}} \bar{Y}_{i}$ 's. If $Y_{i l}$ and $M_{i l}$ denote the total and size of PSU split $U_{i l}$, respectively, for $i, l=1,2$, then $Y_{i}=Y_{i 1}+$ $Y_{i 2}, Y_{i}^{*}=Y_{i 1}+Y_{i^{\prime} 2}, M_{i}=M_{i 1}+M_{i 2}$ and $M_{i}^{*}=M_{i 1}+M_{i^{\prime} 2}$ for $i \neq i^{\prime}=1$, 2. It is clear from (4) that the variance will not change under PSU masking if the following condition holds:

$$
\begin{equation*}
\sqrt{M_{i}} \bar{Y}_{i}=\sqrt{M_{i}^{*}} \bar{Y}_{i}^{*} \quad \text { or } \quad \frac{Y_{i 1}+Y_{i 2}}{\sqrt{M_{i 1}+M_{i 2}}}=\frac{Y_{i 1}+Y_{i^{\prime} 2}}{\sqrt{M_{i 1}+M_{i^{\prime} 2}}} . \tag{5}
\end{equation*}
$$

This result is a bit surprising since by naive intuition one may think that PSU-splits for swapping with $\bar{Y}_{i}=\bar{Y}_{i}^{*}$ will preserve the variance. To better understand (5), consider the following three cases. If $M_{12}=M_{22}$, then (5) implies $Y_{12}=Y_{22}$ or $\bar{Y}_{12}=\bar{Y}_{22}$, where $\bar{Y}_{i l}=Y_{i l} / M_{i l}$ denotes the mean of $U_{i l}$. If $Y_{12}=Y_{22}$, then (5) implies $M_{12}=M_{22}$ or $\bar{Y}_{12}=\bar{Y}_{22}$. If $\bar{Y}_{12}=\bar{Y}_{22}$, then (5) can be written as $M_{i 1} \bar{Y}_{i 1}\left(\sqrt{M_{i}^{*} / M_{i}}-1\right)+\bar{Y}_{i 2}\left(M_{i 2} \sqrt{M_{i}^{*} / M_{i}}-M_{i^{\prime} 2}\right)=0$ for $i=1,2$, holding when $M_{12}=M_{22}$. It is clearly demonstrated from all of the three cases that the variance
will not change if the PSU-splits for swapping are formed equal in both size and mean.

### 2.3 Change in ICC

The effect of the clustering structure distortion on variance can also be investigated through the ratio of the masked and unmasked variances. Let $\rho_{y U}^{*}$ denote the masked ICC, that is, the ICC defined with the masked PSU identifiers. From (3), it is clear that the difference between the masked and unmasked ICCs is proportional to the difference between the corresponding variances, that is,

$$
\rho_{y U}^{*}-\rho_{y U} \doteq\left[m^{-1}(\bar{M}-1) S_{y}^{2}\right]^{-1}\left[V\left(\hat{\bar{Y}} \mid S^{*}\right)-V(\hat{\bar{Y}} \mid S)\right] .
$$

The second approximation in (2) indicates that the change in ICC depends upon how the PSU-splits are formed for swapping, that is, the change in $M_{i}\left(\bar{Y}_{i}-\bar{Y}\right)^{2}$. From (2), the ratio of the masked and unmasked variances is given as

$$
\begin{equation*}
R V\left(\hat{\bar{Y}} \mid S, S^{*}\right) \doteq \frac{\rho_{y U}^{*}+(\bar{M}-1)^{-1}}{\rho_{y U}+(\bar{M}-1)^{-1}} . \tag{6}
\end{equation*}
$$

See, also, Dohrmann et al. (2005, equation 8). Under the relationship of $\rho_{y U}^{*}=c_{y} \rho_{y U}$ for any given $c_{y}>0$, (6) is monotone in $\rho_{y U}$. Also, the ratio is very unstable when $\rho_{y U}$ or $\rho_{y U}^{*}$ is near $-(\bar{M}-1)^{-1}$, the lower bound of the ICC, because both numerator and denominator with their ICCs being near the lower bound are all close to zero. It indicates that any variable of such kind will be greatly influenced by PSU masking.

In general, surveys collect more than one variable and thus PSU masking based on one variable may not preserve well the ICC and thus not the variance of other variables. To better understand such an aspect, consider situations where the PSU masking results in both fixed and random distortion of the ICC written as $\rho_{y U}^{*}=c_{y} \rho_{y U}+e$ for $-0.02<\rho_{y}<0.2$, where $c_{y}=0.7,1.0,1.3, e \sim N\left(0,0.05^{2}\right)$ and $m=\bar{M}=$ $S_{y}^{2}=100$. The constant coefficient $c_{y}$ and the error term $e$ in the model, respectively, allow deterministic and random perturbation in the ICC of the corresponding variable due to masking. Figure 1 displays the resulting ratio of the masked and unmasked standard errors (square-root of variances) against the ICC of the sample design. Three scatter plots in Figure 1 are all similar in their funnel shape with a wide variation for very small $\rho_{y U}$. However, their generic patterns depend on the magnitude of $c_{y}$. For example, $c_{y}<1$ produces a decreasing pattern, $c_{y}>1$ an increasing pattern, and $c_{y}=1$ a non-monotonic pattern, respectively. As will be discussed in Section 3.2, the case of $c_{y}>1$ may rarely occur.

The above discussion may not be extended straightforwardly to other complex survey situations, mainly because surveys often involve complex sample design features
such as stratification, three or higher-stage selection and unequal probability sampling. Under such circumstances, the ICC may not be easily defined and the variance may not be approximated well in the form of (2) (see, e.g., Park (2004) and references cited therein). Nonetheless, the discussion in this section is still helpful to understand the effect of PSU masking on variance estimates in general.

## 3. Effect of degree of PSU masking on variance of sample mean

The more the clustering structure is distorted, the larger the bias in variance estimation. To study such a relationship, we consider a parametric model used for two-stage sampling. Suppose that two-stage sampling selects $n$ PSUs and $m$ units within each sampled PSU. Following Valliant, Dorfman and Royall (2000, page 248), we assume a sampled value $y_{i j}$ for the $j^{\text {th }}$ unit of PSU $i$ is generated from the following model:
$\xi: E_{\xi}\left(y_{i j}\right)=\mu_{y i} \& \operatorname{Cov}_{\xi}\left(y_{i j}, y_{i j^{\prime}}\right)= \begin{cases}\sigma_{y i}^{2} & \text { if } i=i^{\prime}, j=j^{\prime}, \\ \sigma_{i i}^{2} \rho_{y i} & \text { if } i=i^{\prime}, j \neq j^{\prime}, \\ 0 & \text { otherwise, }\end{cases}$
where $\mu_{y i}, \sigma_{y i}^{2}$ and $\rho_{y i}$ are the mean, variance and correlation of units within PSU $i$, respectively. The variance of a sample mean $\bar{y}_{2 s t}=(\mathrm{nm})^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{i j}$ is written as

$$
\begin{equation*}
V_{\xi}\left(\bar{y}_{2 s t} \mid S\right)=(n m)^{-1} \sigma_{y u}^{2}\left[1+(m-1) \rho_{y u}\right], \tag{7}
\end{equation*}
$$

where $\sigma_{y u}^{2}=n^{-1} \sum_{i=1}^{n} \sigma_{y i}^{2}, \rho_{y u}=\sum_{i=1}^{n}\left(\sigma_{y i}^{2} \rho_{y i}\right) / \sum_{i=1}^{n} \sigma_{y i}^{2}$ and $S$ denotes the sample index set. Note that $\rho_{y u}$ can be interpreted as the (pooled or $\sigma_{y i}^{2}$ - weighted) ICC under the model $\xi$.

Let $\beta$ denote the relative size of PSU splits to be swapped between the PSUs $i_{1}$ and $i_{2}$. For simplicity, we assume $m \beta$ to be an integer value, which is the number of units in each split for swapping. Let $S^{*}$ denote the masked sample index set. The variance of $\bar{y}_{2 s t}$ with $S^{*}$ can be written as

$$
\begin{equation*}
V_{\xi}\left(\bar{y}_{2 s t} \mid S^{*}\right)=V_{\xi}\left(\bar{y}_{2 s t} \mid S\right)+n^{-2}(\gamma-1)\left(\sigma_{y_{i}}^{2} \rho_{y_{i}}+\sigma_{y_{i}}^{2} \rho_{y_{i}}\right), \tag{8}
\end{equation*}
$$

for $\gamma=\beta^{2}+(1-\beta)^{2}$. The proof of ( 8 ) is given in the Appendix. Note that $-1<\gamma-1<0$ for $0<\beta<1$. The ratio of the masked and unmasked variances is written as

$$
\begin{equation*}
R V_{\xi}\left(\bar{y}_{2 s t} \mid S, S^{*}\right)=1+m(\gamma-1) \frac{\left(\sigma_{y i}^{2} \rho_{y i_{i}}+\sigma_{y i}^{2} \rho_{y y_{i}}\right)}{\sigma_{y u}^{2}\left[1+(m-1) \rho_{y u}\right]} . \tag{9}
\end{equation*}
$$



Figure 1 Ratios of the masked and unmasked standard errors against original intracluster correlation coefficient
 $c_{y}=0.7,1.0,1.3$ and $m=\bar{M}=S_{y}^{2}=100$

The variance will not change if swapping is done such that $\sigma_{y i_{1}}^{2} \rho_{y i_{1}}+\sigma_{y i_{2}}^{2} \rho_{y i_{2}}=0$, that is, the correlations within the corresponding PSU being opposite in their direction. Otherwise, the change in variance will be at the rate of $m(\gamma-1)<0$ for $\sigma_{y i_{1}}^{2} \rho_{y i_{1}}+\sigma_{y i_{2}}^{2} \rho_{y i_{2}} \neq 0$.

In general, units tend to be more similar within a PSU than across PSUs with $\rho_{y i}$ being small and positive in many populations (e.g., Valliant et al. 2000, Section 8.2.3). Thus, it is more likely that $\sigma_{y i_{1}}^{2} \rho_{y i_{1}}+\sigma_{y i_{2}}^{2} \rho_{y i_{2}}>0$ unless $\sigma_{y i}^{2} \doteq 0$ for all $i=i_{1}, i_{2}$ and the masked variance is prone to be smaller than the unmasked variance, that is, $R V_{\xi}\left(\bar{y}_{2 s t} \mid S, S^{*}\right)<1$. Figure 2 depicts the change in standard error against the unmasked (or baseline) ICC $\rho_{y u}$ with varying the proportion of units to be swapped between the two PSUs. Figure 2 shows that the more units that are swapped, the more the variance is changed, indicating that minimal swapping (i.e., PSU masking) should be done in order to not induce serious bias in the variance. Also, Figure 2 exhibits the L-shape decreasing pattern of the standard error ratio in the ICC, that is, indicating overestimation for negative ICCs but underestimation for positive ICCs. Therefore, under PSU masking, we can expect patterns of either kind $c_{y}=0.7$ (decreasing but random) or $c_{y}=1.0$ (pure random) in Figure 1, with the latter being the best results attainable with minimal masking. In Section 4, we propose a PSU masking strategy through SSU swapping that helps produce a pattern of the second kind in the resulting variance ratios. In Section 5, we apply the proposed strategy to artificial survey data with varying proportions of swapping.


Figure 2 Ratios of the masked and unmasked standard errors $\sqrt{R V_{\xi}}\left(\bar{y}_{2 s t} \mid S, S^{*}\right)$ against the ICC with varying the proportion of swapping units from each PSU. $\sigma_{y u}^{2} \equiv \sigma_{y i}^{2}=$ $25, \rho_{y u} \equiv \rho_{y i}$ for all $i, n=10, m=16$, for $\beta=$ $(0.1,0.2,0.3)$ and $-0.5(m-1)^{-1} \leq \rho_{y u} \leq 0.2$

## 4. PSU masking strategy for limiting biases in variance estimation

Many large-scale surveys involve several stages of sampling with unequal selection probabilities. Under such circumstances, the second stage or subsequent stage sampling units can be a natural choice for swapping to create pseudo-PSUs for operational reasons. For example, in the recent data releases of the National Health and Nutrition Examination Survey (NHANES) (Dohrmann et al. 2005) are included the pseudo-PSU identifiers constructed by swapping SSUs between the original PSUs. In this section, we consider SSU swapping for the purpose of PSU masking under stratified multi-stage sampling and their effect on variance estimates. We suggest a SSU swapping strategy based on the contribution of SSUs to variance estimates.

### 4.1 SSUs in variance estimation under stratified multistage sampling

Suppose that a finite population $U$ of $M$ units is partitioned into $N$ PSUs and similar PSUs in a number of characteristics are grouped to form a total of $H$ strata. Suppose also that each stratum consists of $N_{h}$ PSUs and each PSU contains $N_{h i}$ SSUs with $N_{h i j}$ ultimate sampling units, where $N=\sum_{h=1}^{H} N_{h}$ and $M=\sum_{h=1}^{H} \sum_{i=1}^{N_{h}} \sum_{j=1}^{N_{h i}} N_{h i j}$. Assume that the first stage sampling selects $n_{h}=2$ PSUs within each stratum independently across strata and the second stage and subsequent stage sampling select, in turn, $n_{h i}$ SSUs within each sampled PSU (hi) and $n_{h i j}$ ultimate units within each sampled SSU (hij), where $h=1, \ldots, H$, $i=1, \ldots, n_{h}$ and $j=1, \ldots, n_{h i}$. Associated with the sampled ultimate unit $(h i j k) \in S$ is the observed value $y_{h i j k}$ of survey variable $y$ and the sample weight $w_{h i j k}$, where $k=1, \ldots, n_{h i j}$ and $S$ denotes the sample index set. The population total $Y=\sum_{h=1}^{H} \sum_{i=1}^{N_{h}} \sum_{j=1}^{N_{h i}} \sum_{k=1}^{N_{h j}} y_{h j j k}$ and size $M$ are estimated by $\hat{Y}=\sum_{(h i j k) \in S} w_{h i j k} y_{h i j k}$ and $\hat{M}=\sum_{(h i j k) \in S} w_{h i j k}$, respectively. Also, the population mean $\bar{Y}=Y / M$ is estimated by $\hat{Y}=\hat{Y} / \hat{M}$ and its Taylor series variance estimator (e.g., Shao and Tu 1995) is given by

$$
\begin{equation*}
v(\hat{\bar{Y}} \mid S)=\sum_{h=1}^{H}\left(\frac{z_{h 1}-z_{h 2}}{2}\right)^{2} \tag{10}
\end{equation*}
$$

where $z_{h i}=z_{h i}(y)=\sum_{j=1}^{n_{h i}} \sum_{k=1}^{n_{h i j}} 2 w_{h i j k} z_{h i j k}$ are the estimated stratum totals of $z_{\text {hijk }}=z_{h i j k}(y)=\hat{M}^{-1}\left(y_{h i j k}-\hat{\bar{Y}}\right)$ for PSU (hi).

Writing $z_{h i}$ in (10) in the units of SSUs, we can see SSUs' contribution to the variance estimate, thus helping find better SSU swapping strategies to limit biases in the variance estimates. If $w_{h i j}$ and $w_{k \mid h i j}$ denote the SSU sampling weights and the conditional ultimate sampling unit weights, respectively, then $w_{h i j k}=w_{h i j} \times w_{k \mid h i j}$. Let $\hat{N}_{h i j}=$ $\sum_{k=1}^{n_{h i j}} w_{k \mid h i j}$ and $\hat{\bar{Y}}_{h i j}=\hat{N}_{h i j}^{-1} \sum_{k=1}^{n_{h i j}} w_{k \mid h i j} y_{h i j k}$ denote the estimated
size and sample mean of SSU (hij), respectively. The quantities $z_{h i}$ in (10) can now be written as

$$
\begin{equation*}
z_{h i}=\sum_{j=1}^{n_{h i}} 2 w_{h i j} z_{h i j} \tag{11}
\end{equation*}
$$

where

$$
z_{h i j}=\sum_{k=1}^{n_{h i j}} w_{k \mid h i j} z_{h i j k}=\hat{M}^{-1} \hat{N}_{h i j}\left(\hat{\bar{Y}}_{h i j}-\hat{\bar{Y}}\right) .
$$

It is clear from (10) and (11) that the contribution of the sampled SSUs to the variance estimate is through three components $\left\{w_{h i j}, \hat{N}_{h i j}, \hat{\bar{Y}}_{h i j}\right\}$ of SSU (hij). In Section 4.2, we will examine closely the effect of PSU masking on variance estimates through SSU swapping.

### 4.2 Effect of SSU swapping on variance estimates

We now assume that two SSUs $\left(h_{a} i_{a} j_{a}\right)$ and $\left(h_{b} i_{b} j_{b}\right)$ are to be swapped between two PSUs $\left(h_{a} i_{a}\right) \neq\left(h_{b} i_{b}\right)$. Then, the masked variance estimate can be written from (10) as

$$
\begin{align*}
& v\left(\hat{\bar{Y}} \mid S^{*}\right)= \\
& v(\hat{\bar{Y}} \mid S)+\sum_{h \in\left\{h_{a}, h_{b}\right\}}\left[\left(\frac{z_{h 1}^{*}-z_{h 2}^{*}}{2}\right)^{2}-\left(\frac{z_{h 1}-z_{h 2}}{2}\right)^{2}\right], \tag{12}
\end{align*}
$$

where $z_{h i}^{*}$ denotes the quantity $z_{h i}$ in (11) with the sample index set $S^{*}$ altered due to swapping. Let $i_{a}^{\prime}$ and $i_{b}$ denote the other PSUs in strata $h_{a}$ and $h_{b}$, respectively, and define $z_{h i(j)}=z_{h i}-2 w_{h i j} z_{h i j}=\sum_{l \neq j} 2 w_{h i l} z_{h i l}$. Then, (12) can be written as

$$
\begin{equation*}
v\left(\hat{\bar{Y}} \mid S^{*}\right)=v(\hat{\bar{Y}} \mid S)+e_{0}(y) g_{0}(y) \tag{13}
\end{equation*}
$$

where $\quad e_{0}(y)=2\left(w_{h_{a} i_{a} j_{a}} z_{h_{a} i_{a} j_{a}}-w_{h_{b} i_{b} j_{b}} z_{h_{b} i_{b} j_{b}}\right) \quad \underset{\hat{M}}{\text { is }} \quad$ the difference in the quantity $2 w_{h i j} z_{h i j}=2 w_{h i j} M^{-1}\left(\hat{\bar{Y}}_{h i j}-\hat{\bar{Y}}\right)$ of the two SSUs to be swapped and

$$
g_{0}(y)= \begin{cases}{\left[z_{h_{a} i_{a}\left(j_{a}\right)}-z_{h_{b} i_{b}\left(j_{b}\right)}\right]} & \text { if } h_{a}=h_{b}, \\ 2^{-1}\left[\left(z_{h_{a_{a}^{\prime}}}-z_{h_{a} i_{a}\left(j_{a}\right)}\right)-\left(z_{h_{b} i_{b}^{\prime}}-z_{h_{b} i_{b}\left(j_{b}\right)}\right)\right] & \text { if } h_{a} \neq h_{b},\end{cases}
$$

is a function of $2 w_{h i j} z_{h i j}$ of the SSUs to be retained in the original PSUs. Note that, for $h_{a}=h_{b}, r_{0}$ can also be expressed $g_{0}=2^{-1}\left[\left(z_{h_{b} i_{b}\left(j_{b}\right)}-z_{h_{a} i_{a}\left(j_{a}\right)}\right)-\left(z_{h_{a} i_{a}\left(j_{a}\right)}-z_{h_{b_{b}}\left(j_{b}\right)}\right)\right]$. It shows that the effect of SSU swapping on the variance estimate will be negligible if the two SSUs for swapping are paired in such a way that the product of $e_{0}(y)$ and $g_{0}(y)$ is close to zero. In other words, the change in the variance estimate under SSU swapping can be controlled when a segment pair is formed taking into account all three components $\left\{w_{h i j}, \hat{N}_{h i j}, \hat{\bar{Y}}_{h i j}\right\}$ so as to minimize $e_{0}(y) \times g_{0}(y)$ as
similar to the case under single-stage PPS cluster sampling in Section 2.2.

In addition, by writing
$g_{0}=g\left(e_{0}\right)= \begin{cases}\left(z_{h_{a} 1}-z_{h_{a} 2}\right)-e_{0} & \text { if } h_{a}=h_{b}, \\ 2^{-1}\left[\left(z_{h_{a} 1}-z_{h_{a} 2}\right)-\left(z_{h_{b} 1}-z_{h_{b}}\right)-e_{0}\right] & \text { if } h_{a} \neq h_{b},\end{cases}$
(13) can be expressed as a quadratic function of $e_{0}(y)$ for given $\left\{z_{h i}: h=\left\{h_{a}, h_{b}\right\}, i=1,2\right\}$. For $h_{a}=h_{b}$, we can show that $v\left(\hat{\bar{Y}} \mid S^{*}\right)>v(\hat{\bar{Y}} \mid S)$ only for $e_{0}$ in between zero and $\left(z_{h_{a} 1}-z_{h_{b} 2}\right)$. When $\underset{\hat{\bar{Y}}}{k_{a} 1}{ }^{2}-z_{h_{b} 2} \doteq 0$, it may be more likely that $v\left(\bar{Y} \mid S^{*}\right)=v(\bar{Y} \mid S)$. Similar arguments can be made for $h_{a} \neq h_{b}$.

### 4.3 Sequential SSU swapping with multiple matching characteristics

Suppose that there are a total of $n_{J}$ SSUs in the sample and only $R$ of them are chosen to form pairs for swapping, where $n_{J}=\sum_{h=1}^{H} \sum_{i=1}^{n_{h}} n_{h i}$ and $1 \leq R<n_{J}$. Assume that a fixed number of $R$ SSUs is chosen in accordance with a certain data risk-utility tradeoff consideration. See, for example, Gomatam, Karr and Sanil (2005) for some related discussion concerning data swapping. In addition, assume that their sequential order for the matching process is given as $j_{1}, j_{2}, \ldots, j_{R}$ say. For example, at first, all possible pairs are formed for each of the $R$ SSUs and the best pair is picked based on a certain distance measure such as (12). The order of the $R$ SSUs for the (main) matching process is then determined according to the ascending order of the distances of the $R$ best pairs.

Let $S^{r-1}$ denote the altered sample index set after the $(r-1)^{\text {th }}$ SSU pair has been formed and swapped, where $r=1, \ldots, R$ and $S^{0} \equiv S$. Let $S_{(j)}^{r-1}$ denote the sample index set with SSUs $j_{r}$ and $j$ being swapped for $S^{r-1}$. Then, the change in the variance estimate caused by swapping the $r^{\text {th }}$ SSU $j_{r}$ and any other SSU that was not involved in the ( $r-1$ ) previous match(es) can be written as

$$
\begin{align*}
\delta_{r}(y, j) & =v\left(\hat{\bar{Y}} \mid S_{(j)}^{r-1}\right)-v\left(\hat{\bar{Y}} \mid S^{r-1}\right) \\
& =e_{r-1}(y, r) g_{r-1}(y, r) \tag{14}
\end{align*}
$$

where $e_{r-1}(y, r)$ and $g_{r-1}(y, r)$ are defined similarly as in (13) but with $S^{r-1}$ and $S_{(j)}^{r-1}$. Clearly, the choice of the best match for the $r^{\text {th }}$ SSU depends on the $(r-1)$ previous match(es) and thus the matching process should be viewed as a sequential process. Note that those SSUs that were matched and swapped in the previous match(es) should be excluded in the current search.

In addition, more than one characteristic can be considered for matching, with the hope that they will be related to many other survey variables so as to minimize the bias in the associated variance estimate. Suppose that $q$
matching characteristics are chosen with care, say $\mathbf{x}=$ $\left(x_{1}, x_{2}, \ldots, x_{q}\right)^{\prime}$ (e.g., Dohrmann et al. 2005, for some related discussion). To measure the distance between SSUs $j_{r}$ and $j$, any distance measure of the form

$$
\begin{equation*}
D_{r}(j \mid \mathbf{x})=\sum_{l=1}^{q} c_{l}\left|v\left[\hat{\bar{X}}_{l} \mid S_{(j)}^{r-1}\right]-v\left[\hat{\bar{X}}_{l} \mid S^{r-1}\right]\right| \tag{15}
\end{equation*}
$$

or

$$
\begin{equation*}
\Delta_{r}(j \mid \mathbf{x})=\sum_{l=1}^{q} c_{l}\left|v\left[\hat{\bar{X}}_{l} \mid S_{(j)}^{r-1}\right]-v\left[\hat{\bar{X}}_{l} \mid S\right]\right| \tag{16}
\end{equation*}
$$

can be considered with any reasonable choice of positive coefficients $c_{l}$. For example, $c_{l} \equiv 1$ simply considers the absolute difference in the variance estimates of $\hat{X}_{l}, c_{l}=$ $v\left(\hat{\bar{X}}_{l} \mid S\right)^{-1}$ the absolute difference in variance estimates relative to the original variance estimates, $c_{l}=X_{l}^{-1}$ the absolute difference in relative variance estimates. The first distance measure (15) considers the change in the variance estimate due to swapping segments of the $r^{\text {th }}$ pair. The second distance measure (16) takes into account the cumulative swapping effect of all the $r$ segment pairs.

Matching constraints can be set, for example, to prohibit the pairing of SSUs from the same PSU and to apply a threshold of the proportion of SSUs from each PSU to be swapped (Lu 2004). Let $J_{A}=\left\{j_{1}, \ldots, j_{R}\right\}$ denote the index set of $R$ SSUs that are considered for forming swapping pairs and let $J_{B}$ denote all possible SSUs that can be matched satisfying a given set of matching constraints. For simplicity, consider that the pairing of SSUs is not allowed within $J_{A}$, that is, $J_{A} \cap J_{B}=\varnothing$. If $D_{r}^{*}(j \mid \mathbf{x})$ denotes the chosen distance measure for SSUs $j_{r}$ and $j$, then a sequential SSU swapping algorithm for limiting the biases of variance estimates can be given as follows:

Step 1. Set $r=1, J_{A}^{r}=J_{A}$ and $J_{B}^{r}=J_{B}$;
Step 2. For each of the $(R-r+1)$ SSUs in $J_{A}^{r}$, compute $D_{r}^{*}(\cdot \mid \mathbf{x})$ for all SSUs in $J_{B}^{r}$;
Step 3. Choose the best match with the smallest $D_{r}^{*}(\cdot \mid \mathbf{x})$, that is, find $j_{r}^{\prime}$ such that $D_{r}^{*}\left(j_{r}^{\prime} \mid \mathbf{x}\right)=$ $\min _{j \in J_{B}^{r}} D_{r}^{*}(j \mid \mathbf{x}) ;$

Step 4. Set $r=r+1$, and drop the chosen pair from the searching pool, that is, set $J_{A}^{r}=J_{A}^{r-1} \backslash\left\{j_{r}\right\}$ and update $J_{B}^{r}$ accordingly, where $J_{B}^{r} \subseteq J_{B}^{r-1} \backslash\left\{j_{r}^{\prime}\right\}$;
Step 5. If $r=R+1$, then stop; otherwise repeat Steps 2-4.
This SSU matching (or swapping) approach basically searches for the pair at the $r^{\text {th }}$ matching that is best in a sense of minimizing the change in variance estimates due to the corresponding SSU swapping. With a large number of SSUs, this method will lead to a scatter plot similar to that of $c_{y}=1.0$ in Figure 1 (i.e., a random perturbation with a funnel-shape pattern).

A choice of more sophisticated optimality criterion applied to $\left\{c_{l}: l=1, \ldots, q\right\}$ may help improve the above method to reduce the magnitude of such random perturbation in variance estimates. Also, if one uses multivariate techniques such as principal component analysis to develop some kind of scores (e.g., one or more principal component axes) from a larger number of continuous characteristics, the magnitude of such random perturbations in the variance estimates may be further reduced. In Section 5, we give examples regarding SSU swapping.

## 5. Examples

### 5.1 Previous work

For a sample design with no stratification but a small number of PSUs, Dohrmann et al. (2002) considered various methods of splitting PSUs into pseudo-PSUs in order to use the delete-one jackknife variance estimation method. Their basic idea is to double the number of masked PSUs by keeping the split PSUs as separate masked PSUs, thus hoping to reduce data disclosure risk as a result of the broken linkage between the true and masked PSUs. In their empirical study, noticeable underestimation patterns were present for the resulting variance estimates for variables with large design effects, which resemble the plot of $c_{y}=0.7$ in Figure 1. Let $S$ and $S^{\dagger}$ denote the unmasked and masked sample index sets respectively. Let $w_{i j}$ denote the sample weight and let $y_{i j}$ denote the observed value of $y$ for the $j^{\text {th }}$ sampled unit in PSU $i$. To explain the observed underestimation patterns, Dohrmann et al. (2005) derived the following relationship

$$
v\left(\hat{\bar{Y}} \mid S^{\dagger}\right)=\frac{n-1}{2 n-1} v(\hat{\bar{Y}} \mid S)+\frac{1}{n(2 n-1)} \sum_{i=1}^{n}\left(z_{1, i}-z_{2, i}\right)^{2},
$$

where $z_{g, i}=\sum_{j \in S_{g, i}} 2 w_{i j} z_{i j}$ are the PSU-split totals of $z_{i j}=\left(\sum_{i j} w_{i j}\right)^{-1}\left(y_{i j}-\sum_{i j} w_{i j} y_{i j} / \sum_{i j} w_{i j}\right)$ and $S_{u, i}$ are the index sets of the $u^{\text {th }}$ split of PSU $i$ for $i=1, \ldots, n$ and $u=1$, 2. It indicates that the resulting variance estimate is about a half of the unmasked one plus a positive value reflecting the between PSU-split totals of $z_{i j}$ within the PSUs. If $S_{u, i}$ are formed such that $z_{1, i} \doteq z_{2, i}$, this PSUsplitting method leads to about a half of the unmasked variance estimate and thus the masked variance estimate could be doubled to get close to the unmasked value.

For the two-PSU-per-stratum design, Dohrmann et al. (2005) considered an alternative approach under which the pseudo-PSUs are constructed by swapping SSUs between the PSUs. As discussed in Section 1, this approach can be viewed as dividing the PSUs into one or more splits and recombining them to construct pseudo-PSUs with swapped PSU splits. For simplicity, we assume that each PSU is divided into two splits $S_{1, h i}$ and $S_{2, h i}$. If it is done so with
$S_{h 1}^{\dagger}=S_{1, h 1} \cup S_{1, h 2} \quad$ and $\quad S_{h 2}^{\dagger}=S_{2, h 1} \cup S_{2, h 2}$, then the masked variance estimate can be written as

$$
\begin{equation*}
v\left(\hat{\bar{Y}} \mid S^{\dagger}\right)=v(\hat{\bar{Y}} \mid S)+\sum_{h=1}^{H} e_{h}(y) g_{h}(y), \tag{17}
\end{equation*}
$$

where $e_{h}(y)=z_{1, h 2}-z_{2, h 1}$ is the difference between the PSU-split totals of $S_{1, h 2}$ and $S_{2, h 1}$ to be swapped and $g_{h}(y)=z_{1, h 1}-z_{2, h 2}$ is the difference between the PSU-split totals of $S_{1, h 1}$ and $S_{2, h 2}$ to be retained in the original PSUs. The proof of (17) is given in the Appendix. Equation (17) indicates that a similar strategy for splitting PSUs would help preserve the magnitude of the original variance estimate. Dohrmann et al. (2005) adopted a probabilitybased record linkage technique (Fellegi and Sunter 1969) to form pairs of SSUs for swapping that are similar in their means $\hat{\bar{Y}}_{h i j}$ for several characteristics, with the hope that the terms $e_{h}$ in (17) are all close to zero. Dohrmann et al. (2005) demonstrated that the PSU-recombination method can help reduce the biases of variance estimates and the resulting underestimation patterns to some degree as compared to the PSU-split method used in Dohrmann et al. (2002). To increase speed and flexibility, Lu (2004) developed SSU-swapping algorithms based on sequentially evaluating distance measures between SSU means without directly considering the impact of successive swapping on the bias of the variance estimate. As discussed in Section 4.2, the effect of SSU swapping on variance estimates can be further reduced by direct consideration of SSU's contribution to the variance estimates. In the next section, we apply both strategies, one by Dohrmann et al. (2005) and the other proposed in Section 4.2, to artificial data from a complex survey.

### 5.2 Data example

To illustrate the effect of PSU masking on variance estimates of sample means, we used real survey data from the 1993 National Health Interview Survey (NHIS) Year 2000 Health objectives Public Use Data File (PUF) with some artificial modification. The NHIS is an annual household health interview survey of the civilian noninstitutionalized population of the United States. The NHIS involves a typical multistage, stratified sample design, with the first stage PSUs consisting of counties or metropolitan areas and the second stage SSUs consisting of segments (that is, a small number of households in a small geographic area) within sampled PSUs. This specific Year 2000 topic questionnaire was administered to one adult sample person per family only in the last half of 1993. The NHIS data used here and its documentation are available from National Center for Health Statistics (1994) or the United States Centers for Disease Control National Center for Health

Statistics website (http://www.cdc.gov/nchs/about/major/ nhis/quest_data_related_doc.htm).

This PUF contains the stratum and PSU identifiers, and sample person's final weights for the purpose of variance estimation. For our example, we used only ten strata but limited their number of PSUs to two per stratum. Two of the selected strata, 110 and 520 , were restricted to their two largest PSUs, $(181,410)$ and $(048,233)$, respectively, and the other eight strata, 102, 142, 192, 211, 261, 300, 561 and 571 contain only two PSUs. The PUF also includes the SSU identifiers but not their sample weights. To generate SSU sample weights $w_{h i j}$, we employed a two-way nested random effects model to fit $\log w_{h j k}=\log w_{h i j}+\log w_{k \mid h i j}$ such that $\log w_{h i j}=\mu+\alpha_{h i}+\beta_{j(h i)}$ and $\log w_{k \mid h i j}=\varepsilon_{k(h i j)}$, where $\mu$ is a common value, $\alpha_{h i}$ is the random effect of PSU (hi), $\beta_{j(h i)}$ is the random effect of SSU $j$ nested within PSU (hi) and $\varepsilon_{k(h i j)}$ is the random effect of sampled person $k$ within SSU (hij). We restricted our study to include only those SSUs with five or more sampled persons, giving a total of 293 SSUs in the analysis. The resulting weight decomposition ( $w_{h i j}, w_{k \mid h i j}$ ) may involve possible model misspecification but it suffices our need for the illustration, since both $w_{h j j}$ and $w_{k \mid h i j}$ are all positive under the model. To obtain SSU pairs for swapping, we used six socio-demographic variables, denoted as $x_{1}, x_{2}, \ldots, x_{6}$. They are listed in Table 1 with their description, definition, overall sample mean and squared root of design effect (i.e., design factor or deft in short). The sample means of these variables range from 0.05 to 0.63 and the design factors from 1.285 to 8.511 .

Table 1
Variables used for matching

| Variable | Description | Definition | Sample mean | Design factor |
| :---: | :---: | :---: | :---: | :---: |
| $x_{1}$ | Male | SEX=1 | 0.49 | 1.285 |
| $x_{2}$ | Hispanicity | HISPANIC | 0.14 | 8.511 |
|  |  | $=00,01, \ldots, 08$ | 0.14 | 8.511 |
| $x_{3}$ | Married couple | MARSTAT $=1,2$ | 0.63 | 3.209 |
| $x_{4}$ | College or higher education | EDUCR $=4,5,6$ | 0.45 | 2.902 |
| $x_{5}$ | High family income of | INCFAMR $=8$ | 0.23 | 3.558 |
| $x_{6}$ | \$50k or higher Has household air been tested for Radon? | TESTRDN = 1 | 0.05 | 2.191 |

We applied the two SSU matching strategies discussed in Section 4 and Section 5.1, respectively. The first strategy employs a distance measure (15) for any SSU pair $\left(r_{a}, r_{b}\right)$ with $c_{l} \equiv 1$ for all $l=1, \ldots, 6$. Let $S^{r-1}$ and $S^{r}$ denote the two sample index sets after the $(r-1)^{\text {th }}$ and $r^{\text {th }}$ swapping, respectively. Then the distance of the $r^{\text {th }}$ matching pair of the first strategy (variance-matching) is written as

$$
D_{r}(v \mid \mathbf{x})=\sum_{l=1}^{6}\left|v\left(\hat{\bar{X}}_{l} \mid S^{r}\right)-v\left(\hat{\bar{X}}_{l} \mid S^{r-1}\right)\right|,
$$

where $v\left(\hat{\bar{X}}_{l} \mid S^{r}\right)$ and $v\left(\hat{\bar{X}}_{l} \mid S^{r-1}\right)$ represent the variance estimates of $\hat{X}_{l}$ for the $l^{\text {th }}$ matching characteristic with $S^{r}$ and $S^{r-1}$ respectively. The smaller the distance is, the smaller the biases of variance estimates arises from swapping the $r^{\text {th }}$ matching pair. The second strategy by Dohrmann et al. (2005) is to pair SSUs that are similar in their sample means of the six matching characteristics. This strategy (mean-matching) defines the distance of the $r^{\text {th }}$ matching pair as:

$$
d_{r}(\mu \mid \mathbf{x})=\sum_{l=1}^{6}\left|\hat{\bar{X}}_{l, r_{a}}-\hat{\bar{X}}_{l, r_{b}}\right|,
$$

where $\hat{\bar{X}}_{l, r_{i}}$ represents the SSU sample mean of SSU $r_{i}$ ( $i=a, b$ ) for matching characteristic $x_{l}$.
Table 2 lists standard error ratios of the six matching characteristics at each matching in the sequential order for each strategy with 18 swapping pairs (representing about $12 \%$ of the SSUs in the study). The first strategy, shown in the left panel of the table, gave a moderate but slightly increasing range of variations in standard error ratios over the sequence of the 18 swapping pairs. The second strategy, shown in the right panel of the table, produced a rather wider range of variation in standard error ratios over the sequence with its dramatic changes from the thirteenth and higher pairs in the swapping sequence. Although both strategies tend to lose their control over the biases in the
variance estimates for higher orders of the swapping sequence, the first strategy was quite successful in controlling the biases of the variance estimates for a relatively large number of swapping pairs.

Figure 3 plots the standard error ratios against the design factors for the two strategies varying the number of SSUs swapped. These three sets included $6(4 \%), 12(8 \%)$ and 18 ( $12 \%$ ) SSU pairs (percentage of SSUs involved in swapping), respectively. Each plot includes two sets of characteristics, 6 matching characteristics marked with the corresponding numbers as listed in Table 1 and 92 characteristics marked with $\times$ that are not used in matching. For the scenario with only $4 \%$ of the SSUs swapped, the difference between the two strategies is negligible for both sets of characteristics. However, as the percentage of SSUs swapped increases, the perturbation in the variance estimates becomes greater for both strategies and both sets of characteristics. This result indicates that a small percentage of swapping should occur, reinforcing the findings of Section 3. In addition, the standard error ratios are clustered more closely to the line of one (i.e., small biases of the masked variance estimates) with the first strategy than with the second strategy. The second strategy produced a rather steeply decreasing pattern over the design factor even for the six matching characteristics. That is, the mean-matching strategy is seen more poignant for the variables used for matching.

Table 2
Standard error ratios by swapping sequence: Comparison of the two matching criteria with $12 \%$ ( 18 pairs) SSU swapping

| Variance-matching |  |  |  |  |  |  | Mean-matching |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Swapping Sequence | $x_{1}$ | $x_{2}$ | $x_{3}$ | $\boldsymbol{x}_{4}$ | $x_{5}$ | $\boldsymbol{x}_{6}$ | $x_{1}$ | $x_{2}$ | $x_{3}$ | $x_{4}$ | $x_{5}$ | $x_{6}$ |
| 1 | 0.999 | 1.000 | 1.000 | 0.998 | 1.000 | 0.998 | 0.998 | 1.000 | 1.000 | 1.002 | 1.002 | 1.001 |
| 2 | 1.002 | 1.000 | 1.000 | 1.000 | 1.000 | 0.996 | 0.999 | 1.000 | 1.000 | 1.002 | 1.001 | 1.001 |
| 3 | 1.004 | 1.000 | 1.000 | 1.001 | 1.000 | 0.996 | 0.998 | 1.000 | 0.999 | 0.997 | 0.994 | 1.001 |
| 4 | 1.012 | 1.001 | 0.999 | 1.000 | 1.000 | 0.989 | 1.025 | 1.000 | 0.999 | 1.001 | 0.994 | 1.016 |
| 5 | 1.009 | 1.001 | 1.000 | 0.998 | 1.001 | 0.988 | 1.021 | 1.004 | 0.964 | 0.968 | 0.951 | 1.013 |
| 6 | 1.007 | 1.000 | 1.000 | 1.000 | 1.003 | 0.988 | 1.020 | 1.004 | 0.964 | 0.968 | 0.955 | 1.015 |
| 7 | 1.011 | 1.000 | 1.000 | 1.002 | 1.003 | 1.008 | 1.020 | 1.004 | 0.964 | 0.970 | 0.954 | 1.017 |
| 8 | 1.016 | 1.000 | 0.997 | 1.002 | 1.003 | 1.026 | 1.022 | 0.998 | 0.957 | 0.963 | 0.964 | 1.005 |
| 9 | 1.009 | 0.998 | 1.000 | 1.002 | 1.003 | 1.020 | 1.021 | 0.997 | 0.955 | 0.965 | 0.982 | 1.034 |
| 10 | 1.007 | 0.996 | 1.001 | 1.010 | 1.006 | 1.014 | 1.019 | 0.997 | 0.931 | 0.960 | 0.972 | 1.033 |
| 11 | 1.014 | 0.994 | 1.005 | 1.010 | 1.001 | 1.012 | 1.020 | 0.995 | 0.946 | 0.953 | 0.989 | 1.034 |
| 12 | 1.029 | 0.995 | 1.003 | 1.013 | 1.011 | 1.036 | 1.021 | 0.991 | 0.946 | 0.953 | 0.987 | 1.035 |
| 13 | 1.064 | 0.992 | 1.003 | 1.001 | 1.008 | 1.047 | 1.035 | 0.990 | 0.946 | 0.932 | 0.967 | 1.114 |
| 14 | 1.008 | 0.991 | 1.000 | 1.007 | 1.022 | 1.044 | 1.031 | 0.955 | 0.946 | 0.929 | 0.952 | 1.103 |
| 15 | 1.042 | 0.988 | 0.984 | 1.017 | 1.015 | 1.044 | 1.052 | 0.955 | 0.946 | 0.922 | 0.952 | 1.124 |
| 16 | 1.012 | 0.982 | 0.986 | 1.024 | 1.041 | 1.042 | 1.107 | 0.939 | 0.936 | 0.920 | 0.942 | 1.128 |
| 17 | 0.987 | 0.978 | 1.000 | 1.016 | 1.009 | 1.021 | 1.107 | 0.878 | 0.936 | 0.927 | 0.935 | 1.123 |
| 18 | 1.029 | 0.943 | 1.000 | 0.970 | 0.947 | 1.042 | 1.014 | 0.538 | 0.946 | 0.841 | 0.945 | 1.106 |

See Table 1 for the description of the six matching characteristics $\left(x_{1}, \ldots, x_{6}\right)$.


Figure 3 Ratio of Standard Errors vs. Baseline Design Factors. Six numbers represent the points of the corresponding matching characteristics and $\times$ marks represent those of 92 characteristics not used in matching

## 6. Discussion

In this paper, we investigated the effect of PSU masking on variance estimates in complex surveys. Obviously, PSU masking distorts the clustering structure of the original sample design, possibly yielding systematic biases in the analysis of the resulting data as seen in Sections 2, 3 and 5.2. The proposed PSU masking strategy in Section 4 can help reduce such biases but still leave a random perturbation in the variance estimation and thus a loss of inferential
efficiency. Research on the effect of PSU masking would be interesting on other types of complex data analyses such as regression and multivariate analyses. Although PSU masking can provide disclosure control, the degree of masking should be minimal to limit the resulting biases of variance estimates as discussed in Sections 3 and 5.2.

In addition, the reduction of the identification risk incurred by SSU masking may be better understood by writing the distance between the masked sample PSU mean and the PSU mean in the population as follows:

$$
\begin{equation*}
\hat{\bar{Y}}_{h i \mid S^{\dagger}}-\bar{Y}_{h i \mid U}=\left(\hat{\bar{Y}}_{h i \mid S^{\dagger}}-\hat{\bar{Y}}_{h i \mid S}\right)+\left(\hat{\bar{Y}}_{h i \mid S}-\bar{Y}_{h i U}\right), \tag{18}
\end{equation*}
$$

where $\hat{\bar{Y}}_{h i \mid S^{\dagger}}$ and $\hat{\bar{Y}}_{h i S}$ denote the masked and unmasked PSU means in the sample, respectively, and $\bar{Y}_{h i U}$ denote the PSU mean in the population that may be available to an intruder (i.e., a malicious data user) from external sources such as Census data. One can show easily that the first term in the right-hand side of (18) is not equal to zero, in general, with PSU masking. $\hat{\bar{Y}}_{h i \mid S^{\dagger}}$ and $\bar{Y}_{h i U}$, together with nonneglible sample variation of the second term in the righthand side of (18), are never equal except by rare chance. Dohrmann et al. (2005) compare $\left\{\hat{\bar{Y}}_{h i \mid S^{\dagger}}\right\}$ of the sample to $\left\{\bar{Y}_{h i U}\right\}$ of the population by a stylish stem-and-leaf diagram to demonstrate how hard it would be for an intruder to identify a sampled PSU in the public release data files in association with two aspects: 1) few pairs of ( $\hat{Y}_{h i \mid S^{\dagger}}, \bar{Y}_{h i U}$ ) being close to each other; and 2) many unsampled PSU's with population values similar to $\left\{\hat{\bar{Y}}_{h i S^{\dagger}}\right\}$ or $\left\{\bar{Y}_{h i \mid U}\right\}$ of the sampled PSUs. Some forms of probabilistic measurements may be interesting to evaluate identification risk reduction (e.g., Eltinge 1999) but are beyond the scope of this paper. The proposed masking strategy has been applied to the 2003-2004 National Health and Nutrition Examination Survey (NHANES) release (Park, Dohrmann, Montaquila, Mohadjer and Curtin 2006).

## Appendix

## Proofs

## Proof of equation (1)

From Park and Lee (2004, Section 4.2),

$$
\begin{aligned}
V(\hat{\bar{Y}} \mid S)= & \frac{1}{m} S_{y}^{2} \times \operatorname{Deft}^{2}(\hat{\bar{Y}} \mid S) \\
& \doteq \frac{1}{m} S_{y}^{2}\left[1+(\bar{M}-1) \rho_{y U}\right] \\
& +\frac{1}{m N} \sum_{i=1}^{N} \frac{M_{i}}{p_{i}}\left(\frac{M_{i}}{p_{i}}-p_{i}\right)\left(\bar{Y} \bar{Y}_{i}-\bar{Y}\right)^{2} \\
\doteq & \frac{1}{m(N-1)} \sum_{i=1}^{N} M_{i}(\bar{Y}-\bar{Y})^{2} \\
& +\frac{1}{m N} \sum_{i=1}^{N} \frac{M_{i}}{p_{i}}\left(\frac{M_{i}}{M}-p_{i}\right)\left(\overline{Y_{i}}-\bar{Y}\right)^{2}, \\
& \doteq \frac{1}{m M N} \sum_{i=1}^{N} \frac{M_{i}^{2}}{p_{i}}(\bar{Y}-\bar{Y})^{2},
\end{aligned}
$$

where $\operatorname{Deft}^{2}(\hat{\bar{Y}} \mid S)$ represents the design effect of $\hat{\bar{Y}}$ for a given $S$, the second and the last approximations follow from $(N-1) / N \doteq 1$ and the third equation from
$S_{y}^{2}\left[1+(\bar{M}-1) \rho_{y U}\right] \doteq(N-1)^{-1} \sum_{i=1}^{N} M_{i}\left(\bar{Y}_{i}-\bar{Y}\right)^{2}, \quad$ which completes the proof.

## Proof of equation (8)

By definition, the variance of the sample PSU total $y_{i}=\sum_{j} y_{i j} \quad$ is $\quad V_{\xi}\left(y_{i} \mid S\right)=m \sigma_{y i}^{2}+m(m-1) \sigma_{y i}^{2} \rho_{y i} \quad$ for $i=1, \ldots, n$. Suppose that $\left\{y_{i j}: j=m(\beta-1)+1, \ldots, n\right\}$ for the two PSUs $i=a$ and $b$ are to be switched between the PSUs. Then these two PSUs have their variance changed to $V_{\xi}\left(y_{a} \mid S^{*}\right)=m(1-\beta) \sigma_{a}^{2}+m(1-\beta)[m(1-\beta)-1] \sigma_{a}^{2} \rho_{a}+$ $m \beta \sigma_{b}^{2}+m \beta(m \beta-1) \sigma_{b}^{2} \rho_{b}$ and to $V_{\xi}\left(y_{b} \mid S^{*}\right)$ being the same with switching the indices $a$ and $b$. Since $m(1-\beta)[m(1-\beta)-1]+m \beta(m \beta-1)=m(m \gamma-1)$, the proof is completed from observing

$$
\begin{aligned}
(n m)^{2} V_{\xi}\left(\bar{y}_{2 s t} \mid S^{*}\right)= & \sum_{i \neq a, b}^{n}\left[m \sigma_{y i}^{2}+m(m-1) \sigma_{y i}^{2} \rho_{y i}\right] \\
& +\sum_{i=a, b}\left[m \sigma_{y i}^{2}+m(m \gamma-1) \sigma_{y i}^{2} \rho_{y i}\right] \\
= & (n m)^{2} V_{\xi}\left(\bar{y}_{2 s t} \mid S\right) \\
& +\sum_{i=a, b} m^{2}(\gamma-1) \sigma_{y i}^{2} \rho_{y i} .
\end{aligned}
$$

## Proof of equation (13)

Suppose that two PSUs ( $h 1$ ) and ( $k 1$ ) from two different strata $h \neq k$ are to be reconstructed by swapping each of their SSUs, $\left(h 1 j_{a}\right)$ and $\left(k 1 j_{b}\right)$. Let $e_{h k}=$ $2\left(w_{h 1 j_{a}} y_{h 1 j_{a}}-w_{k 1 j_{b}} y_{k 1 j_{b}}\right)$ denote the difference between the contributions of the two SSUs to $z_{h i}$ in (11). Let $z_{h 1\left(j_{a}\right)}=$ $\sum_{j \neq j_{a}} 2 w_{h 1 j} z_{h 1 j}$ and $z_{k 1\left(j_{b}\right)}=\sum_{j \neq j_{b}} 2 w_{k 1 j} z_{k 1 j}$ denote, respectively, $z_{h i}$ excluding the contributions from the SSUs to be swapped. By noting that $z_{h 1}^{*}=z_{h 1}-e_{h k}, z_{h 2}^{*}=z_{h 2}, z_{k 1}^{*}=$ $z_{k 1}+e_{h k}$, it follows from (12) that

$$
\begin{aligned}
4\left[v\left(\hat{\bar{Y}} \mid S^{*}\right)-v(\hat{\bar{Y}} \mid S)\right]= & \left(z_{h 1}^{*}-z_{h 2}^{*}\right)^{2}-\left(z_{h 1}-z_{h 2}\right)^{2} \\
& +\left(z_{k 1}^{*}-z_{k 2}^{*}\right)^{2}-\left(z_{k 1}-z_{k 2}\right)^{2} \\
& =2 e_{h k}\left[e_{h k}-\left(z_{h 1}-z_{h 2}\right)+\left(z_{k 1}-z_{k 2}\right)\right]
\end{aligned}
$$

and thus, (13) holds with $g_{h k}=2^{-1}\left\{\left[z_{h 2}-z_{h 1\left(j_{a}\right)}\right]\right.$ $\left.\left[z_{k 2}-z_{k 1\left(j_{b}\right)}\right]\right\}$. The proofs for the other three cases are similar. When $h=k$, we have $e_{h h}=2\left(w_{h 1 j_{a}} z_{h 1 j_{a}}-\right.$ $\left.w_{h 1 j_{b}} z_{h 1 j_{b}}\right), z_{h 1}^{*}=z_{h 1}-e_{h h}$ and $z_{h 2}^{*}=z_{h 2}-e_{h h}$. The proof is completed by letting $g_{h h}=2^{-1}\left\{\left[z_{h 2\left(j_{b}\right)}-z_{h 1\left(j_{a}\right)}\right]-\left[z_{h 1\left(j_{a}\right)}-\right.\right.$ $\left.\left.z_{h 2\left(j_{b}\right)}\right]\right\}=z_{h 1\left(j_{b}\right)}-z_{h 2\left(j_{b}\right)}$.

## Proof of equation (17)

By definition, we have $z_{h i}=z_{1, h i}+z_{2, h i}$ and $z_{y i}^{\dagger}=z_{i, h i}+$ $z_{i, h i}$ for any $h$ and $i$. Thus, observing $z_{h 1}^{\dagger}-z_{h 2}^{\dagger}=z_{h 1}-$ $z_{h 2}+2\left(z_{1, h 2-z_{2, h 1}}\right)$ and $z_{h 1}-z_{h 2}+z_{1, h 2}-z_{2, h 1}=z_{1, h 1}-$ $z_{2, h 2}$, we have

$$
\begin{aligned}
v\left(\hat{\bar{Y}} \mid S^{\dagger}\right)= & \sum_{h=1}^{H}\left(\frac{z_{h 1}^{\dagger}-z_{h 2}^{\dagger}}{2}\right)^{2} \\
= & \sum_{h=1}^{H}\left[\left(\frac{z_{h 1}-z_{h 2}}{2}\right)+\left(z_{1, h 2}-z_{2, h 1}\right)\right]^{2} \\
= & \sum_{h=1}^{H}\left(\frac{z_{h 1}^{\dagger}-z_{h 2}^{\dagger}}{2}\right)^{2} \\
& +\sum_{h=1}^{H}\left(z_{1, h 2}-z_{2, h 1}\right)\left(z_{h 1}-z_{h 2}+z_{1, h 2}-z_{2, h 1}\right) \\
= & v(\hat{\bar{Y}} \mid S)+\sum_{h=1}^{H}\left(z_{1, h 1}-z_{2, h 2}\right)\left(z_{1, h 2}-z_{2, h 1}\right),
\end{aligned}
$$

which completes the proof.

## Acknowledgements

This work was done while the author was at Westat, Inc., U.S.A. The author thank Leyla Mohadjer, Sylvia Dohrmann, Jill Montaquila and Lexter R. Curtin for their support to this research. The author is also grateful to Barry Graubard, the associate editor and two anonymous reviewers for their helpful comments, which helped improve the paper.

## References

Dalenius, T., and Peiss, S.P. (1982). Data-swapping: A technique for disclosure control. Journal of Statistical Planning and Inferences, 6, 73-85.

Dohrmann, S., Curtin, L.R., Mohadjer, L., Montaquila, J. and Le, T. (2002). National Health and Nutrition Examination Survey limiting the risk of data disclosure using replication techniques in variance estimation. Proceedings of the Survey Research Methods Section, American Statistical Association, 807-812.

Dohrmann, S., Lu, W., Park, I., Sitter, R. and Curtin, L.R. (2005). Variance estimation and data disclosure issues in the National Health and Nutrition Examination Survey limiting the risk of data disclosure using replication techniques in variance estimation. Submitted to a journal.

Eltinge, J.L. (1999). Use of stratum mixing to reduce primary-unitlevel identification risk in public-use survey datasets. Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference.

Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64, 1183-1210.

Gomatam, S., Karr, A.F. and Sanil, A.P. (2005). Data swapping as a decision problem. Journal of Official Statistics, 21, 635-655.

Lu, W. (2004). Confidentiality and Variance Estimation in Complex Surveys. Unpublished Ph.D. dissertation, Simon Fraser University, Department of Statistics and Actuarial Science.

Lu, W., Brick, M.J. and Sitter, R.R. (2006). Algorithms for constructing combined strata grouped jackknife and balanced repeated replications with domains. Journal of the American Statistical Association, 101, 1680-1692.

Mayda, J.E., Mohl, C. and Tambay, J.-L. (1996). Variance estimation and confidentiality: They are related! Proceedings of the Survey Methods Section, Statistical Society of Canada, 135-141.

National Center for Health Statistics (1994). Data file document, National Health Interview Survey of Topics Related to the Year 2000 Health Objectives, 1993 (machine readable data file and documentation), National Center for Health Statistics, Hyattsville, Maryland.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. and Curtin, L.R. (2006) Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. Proceedings of the Survey Research Methods Section, American Statistical Association, 1761-1767.

Park, I. (2004). Assessing complex sample designs via design effect decompositions. Proceedings of the Survey Research Methods Section, American Statistical Association, 4135-4142.

Park, I., and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. Survey Methodology, 30, 183-193.

Rust, K.F. (1986). Efficient replicated variance estimation. In the Proceedings of the Survey Research Methodos Section, American Statistical Association, 81-87.

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

Shah, B. (2001). A Method to Create Pseudo Strata and PSU's based on BRR weights. Unpublished manuscript, Research Triangle Institute.

Shao, J., and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer-Verlag.

Valliant, R. (1996). Limitations of balanced half-sampling. Journal of Official Statistics, 12, 225-240.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). Finite Population Sampling and Inference: A Predication Approach. New York: John Wiley \& Sons, Inc.

Yung, W. (1997). Variance estimation for public use files under confidentiality constraints. Proceedings of the Survey Research Methods Section, American Statistical Association, 434-439.

# A tree-based approach to forming strata in multipurpose business surveys 

Roberto Benedetti, Giuseppe Espa and Giovanni Lafratta ${ }^{1}$


#### Abstract

The design of a stratified simple random sample without replacement from a finite population deals with two main issues: the definition of a rule to partition the population into strata, and the allocation of sampling units in the selected strata. This article examines a tree-based strategy which plans to approach jointly these issues when the survey is multipurpose and multivariate information, quantitative or qualitative, is available. Strata are formed through a hierarchical divisive algorithm that selects finer and finer partitions by minimizing, at each step, the sample allocation required to achieve the precision levels set for each surveyed variable. In this way, large numbers of constraints can be satisfied without drastically increasing the sample size, and also without discarding variables selected for stratification or diminishing the number of their class intervals. Furthermore, the algorithm tends not to define empty or almost empty strata, thus avoiding the need for strata collapsing aggregations. The procedure was applied to redesign the Italian Farm Structure Survey. The results indicate that the gain in efficiency held using our strategy is nontrivial. For a given sample size, this procedure achieves the required precision by exploiting a number of strata which is usually a very small fraction of the number of strata available when combining all possible classes from any of the covariates.


Key Words: Multivariate stratification; Optimal multipurpose sample allocation; Farm Structure Survey; Sample design.

## 1. Introduction

Many business surveys employ stratified sampling procedures in which simple random sampling without replacement is executed within each stratum (see, e.g., Sigman and Monsour 1995, and, for farm surveys, Vogel 1995). Usually the list frame from which units are selected is set up using administrative or census information, represented by a rich data base of auxiliary variables, each of which can be potentially exploited to form strata. Furthermore, such surveys are often also multipurpose, and given precision levels must be achieved in estimating multiple variables under study.

The goal of satisfying such a large number of constraints without drastically increasing the sample size is commonly considered as strictly related to the choice of the number of stratifying variables and of their class intervals (Kish and Anderson 1978). This is due to the well known fact that finer partitions of the population introduce more information useful for the reduction of estimation variances, but, on the other hand, their application implies higher risks for units to become jumpers.

Let us indicate as the atomised stratification that one obtained forming strata by combination of all possible classes from any of the covariates in use. If the corresponding number of such basic strata, or atoms, exceeds a given threshold imposed by practical restrictions,
it seems unavoidable to redesign the survey selecting a smaller number of stratifying variables or creating fewer classes from each of them. Notwithstanding, it can be noted that another way of obviating such an unsatisfactory situation can be based on the following argument: the atomised stratification can really be interpreted as an extreme solution to the problem of strata formation, since, between the cases of no stratification and using the atomised stratification, there exists a full range of opportunities to select a stratification whose subpopulations can be obtained as unions of atoms.

Our proposal is to accomplish this selection through the definition of a tree-based stratified design. We form strata by means of a hierarchical divisive algorithm that selects finer and finer partitions by minimizing, at each step, the sample allocation required to achieve the precision levels set for each surveyed variable. The procedure is sequential, and determines a path from the null stratification, i.e., that one whose single stratum matches the population, to the atomised one. At each step, we select which variable is to be used to define the new, more disaggregated partition: each stratum in the current partition is split on any covariate, using in turn all of its available classes, and the one that better decreases the global allocation size is selected.

Bloch and Segal (1989) discussed the application of classification tree methods (see, e.g., Breiman, Friedman, Olshen and Stone 1984) to strata formation, but their focus

[^7]was mainly on strata interpretation about the relationships between the covariates and a unique outcome variable. Instead, our rules to partition the population are directly oriented to the optimal allocation of sampling units in the selected strata. The classical methods which deal with the univariate case (Dalenius and Hodges 1959; Singh 1971; Lavallée and Hidiroglou 1988; Hedlin 2000; Lu and Sitter 2002; Gunning and Horgan 2004; for a review see Horgan 2006) can't be easily extended to cover the case where one seeks to exploit multiple covariates for stratification. The solutions proposed in this literature are, as a consequence, of poor practical value if the survey is multipurpose and information on multiple covariates is available. In such a context, methods to satisfy a large number of constraints on errors when minimizing the sample size were proposed by Bethel $(1985,1989)$ and Chromy $(1987)$. Valliant and Gentle (1997) also approached the problem for two-stage sampling frameworks. For a given stratification, we choose to apply the Bethel's allocation rule and henceforth the procedure selects subsequent partitions by minimizing the survey cost function corresponding to the stratifications consisting of the currently unsplit strata and of the available split substrata.

According to what we have said before, our position in the grand picture of multivariate stratification follows the goal by Kish and Anderson (1978), namely bringing some results in the field of stratified sampling towards the needs of survey practice. Practitioners daily perform multivariate (several variables available for stratification) and multipurpose (several variables and many other statistics are the main objectives of survey efforts) surveys. Thus, the aim of our approach consists in giving the possibility of combining stratification and sample allocation. This means that we are concerned with the choice of the number of stratifying variables, of the number of class intervals for each variable and of the optimal Bethel's allocation to strata. As of this choice, our methodology cannot be reduced to the standard solution of the multivariate stratification problem, i.e., the use of multivariate techniques such as cluster analysis and principal components (see, for example, Mulvey 1983, Pla 1991 and Jarque 1981). As a matter of fact, this branch of literature does not use (or uses only indirectly) the variables of interest, but only the auxiliary variables, and the allocation issue is neglected. It would be even less justifiable to reduce our approach to the ones reviewed by Särndal, Swensson and Wretman (1992, section 12.6 and 12.7): the techniques presented in section 12.6 combine stratification and multivariate sample allocation, but are not multipurpose, whereas the methods of section 12.7 are multipurpose but are based on predetermined strata.

The paper is organized as follows. Section 2 introduces the procedure we propose for the computation of
stratification trees. We thoroughly describe the algorithm used to generate the sequence of stratifications, and we show how it can be represented as a classification tree. Stopping criteria are also discussed to determine how they can affect the optimal number of strata. In Section 3 we examine how a stratification tree can be exploited to design the European Community survey on the structure of agricultural holdings, also known as Farm Structure Survey (FSS). We illustrate our stratification technique identifying a tree-based set of strata and allocations using a basic set of atoms defined by means of multivariate information collected during the fifth Agricultural General Census held in Italy in the year 2000. Finally, Section 4 is devoted to some concluding remarks, focusing on issues regarding the practice of forming strata by trees and discussing how the procedure can be used to better manage multipurpose surveys based on stratified designs.

## 2. A procedure to generate multivariate stratification trees

Consider a finite population $P$ of $N$ units, on which variables $y_{1}, \ldots, y_{g}, \ldots, y_{G}$ are to be surveyed to estimate their totals using a stratification on $P$, i.e., a collection $F$ of $H_{F}$ nonempty subpopulations, called strata, partitioning $P$. Our problem is how to select $F$ in order to minimize the corresponding overall sample allocation $n_{F}$ in a way such that, for $g=1, \ldots, G$, the coefficient of variation $\left(\mathrm{CV}_{g}\right)$ corresponding to the $g^{\text {th }}$ variate of interest is not greater than the desired level of precision, say $\varepsilon_{g}>0$.

For a given $F$, such minimization is executed by computing the Bethel's (1985) sample allocation rule. More thoroughly, let us indicate by $n_{h}, h=1, \ldots, H_{F}$, the sample allocation in stratum $h$. The global survey cost corresponding to $F$ can thus be given as follows

$$
f\left(n_{1}, \ldots, n_{H_{F}}\right)=c_{F}+\sum_{h=1}^{H_{F}} c_{h} n_{h},
$$

where $c_{F}$ is a fixed cost independent from $\mathbf{n}_{F}=$ $\left(n_{1}, \ldots, n_{H_{F}}\right)^{\prime}$, and $c_{h}$ represents the cost to sample one unit in stratum $h$. Furthermore, let $Y_{g}$ be the total in $P$ of the $g^{\text {th }}$ response variable, $N_{h}$ the size of the $h^{\text {th }}$ stratum of $F$, and $S_{h, g}^{2}$ the variance of $y_{g}$ in stratum $h$. Then the $g^{\text {th }}$ constraint on the required precision can be expressed as:

$$
\begin{aligned}
\left(\mathrm{CV}_{g}\right)^{2} \leq \varepsilon_{g}^{2} & \equiv \sum_{h=1}^{H_{F}} \frac{N_{h}^{2} S_{h, g}^{2}}{n_{h}}-\sum_{h=1}^{H_{F}} N_{h} S_{h, g}^{2} \leq Y_{g}^{2} \varepsilon_{g}^{2} \\
& \equiv \sum_{h=1}^{H_{F}} \frac{N_{h}^{2} S_{h, g}^{2}}{\left(Y_{g}^{2} \varepsilon_{g}^{2}+\sum_{h=1}^{H_{F}} N_{h} S_{h, g}^{2}\right) n_{h}} \leq 1,
\end{aligned}
$$

so that, if we consider the following quantities, referred to as the standardized precision units,

$$
\xi_{h, g}=N_{h}^{2} S_{h, g}^{2} /\left(Y_{g}^{2} \varepsilon_{g}^{2}+\sum_{h=1}^{H_{F}} N_{h} S_{h, g}^{2}\right),
$$

the problem of optimal allocation for $F$ can be expressed as follows:

$$
\begin{array}{lll}
\min & f\left(\mathbf{n}_{F}\right) & \\
\text { subject to } & \sum_{h=1}^{H_{F}} \xi_{h, g} / n_{h} \leq 1, & g=1, \ldots, G, \\
& 1 / n_{h}>0, & h=1, \ldots, H_{F} .
\end{array}
$$

Bethel (1989) derived the solution to such problem, say $n_{h}^{*}$, $h=1, \ldots, H_{F}$, as follows:
$1 / n_{h}^{*}=$
$\begin{cases}\sqrt{c_{h}} /\left(\sqrt{\sum_{g=1}^{G} \alpha_{g}^{*} \xi_{h, g}} \sum_{l=1}^{H_{F}} \sqrt{c_{l} \sum_{g=1}^{G} \alpha_{g}^{*} \xi_{l, g}}\right) & \text { if } \sum_{g=1}^{G} \alpha_{g}^{*} \xi_{h, g}>0, \\ +\infty & \text { otherwise },\end{cases}$
where $\alpha_{g}^{*}=\lambda_{g} / \sum_{g=1}^{G} \lambda_{g}$, and $\lambda_{g}$ is the Lagrangian multiplier of the constraint on the maximum error allowed estimating the $g^{\text {th }}$ surveyed variable, and indicates whether the $g^{\text {th }}$ constraint is "active" in the allocation problem solution (namely, if $\alpha_{g}^{*}=0$, then the constraint is not active). The corresponding global optimal allocation is thus given by setting $n_{F}=\sum_{h=1}^{H_{F}} n_{h}^{*}$.

Let us now assume that total estimates and their variances are available for any of a given set of $M>1$ basic strata $A_{1}, \ldots, A_{m}, \ldots, A_{M}$, so that we can rely on two $M \times G$ matrices, respectively of totals $\mathbf{T}=\left(Y_{m, g}\right)$ and estimation variances $\mathbf{V}=\left(S_{m, g}^{2}\right)$, and the sizes $N_{m}$, $m=1, \ldots, M$. The definition of such strata, which in the sequel will be referred to as atoms, is based on a set of covariates $X_{1}, \ldots, X_{k}, \ldots, X_{K}$ as follows. Let $x_{i, k}$ be the value of $X_{k}$ measured on unit $i \in P$, and consider the set of distinct values observed for $X_{k}$ in $P, \Xi_{k}=$ $\left\{x \in \mathbb{R}: \exists i \in P: x=x_{i, k}\right\}$. We build $\quad M=\prod_{k=1}^{K}\left|\Xi_{k}\right|$ atoms, one for every vector ( $a_{m, 1}, \ldots, a_{m, K}$ ) in the Cartesian product $\Xi=\otimes_{k=1}^{K} \Xi_{k}$, by setting for $m=1, \ldots, M$

$$
A_{m}=\bigcap_{k=1}^{K} A_{m, k}
$$

where $A_{m, k}=\left\{i \in P: x_{i, k}=a_{m, k}\right\}$. In the case where the covariates $X_{k}$ are continuous, the set $\Xi_{k}$ will contain $N$ not empty atoms. As the algorithm is hierarchical divisive, the number of final atoms does not affect at all the steps of the algorithm. Only the initial phase of construction of the aggregate statistics and, at most, the memory allocation are impacted. Our empirical experience suggests that the computing times do not change much if the size of the atoms is equal to 1 . On the contrary, continuous or ordered variables speed up the algorithm, as the number of possible binary partitions is, if the number of values is the same, much smaller with respect to the case of categorical variables. One verifies that this construction does yield a
stratification: each unit of the population appears in one atom, and in one only. To illustrate our definitions, let us refer to the data shown in Table 1, where a simple example is described in which a set of $M=9$ atoms (obtained exploiting $K=2$ covariates both having 3 distinct values, namely 1,2 , and 3 ) is assumed to constitute the basic stratification to survey $G=2$ variables, whose totals and estimation variances are also reported, together with atom sizes. In this context, we have $\Xi_{1}=\Xi_{2}=\{1,2,3\}$, and, for example, $A_{8}$ is the subpopulation whose elements $i$ are such that $x_{i, 1}=a_{8,1} \equiv 3$ and $x_{i, 2}=a_{8,2} \equiv 2$.

Table 1
Example data for 9 atoms and 2 surveyed variates

| Atoms |  |  |  | Surveyed Variates |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Id | Definition |  | $\begin{gathered} \hline \text { Sizes } \\ N_{m} \end{gathered}$ | Totals |  | Variances |  |
| m | $a_{m, 1}$ | $a_{m, 2}$ |  | $\boldsymbol{Y}_{m, 1}$ | $\boldsymbol{Y}_{m, 2}$ | $S_{m, 1}^{2}$ | $S_{m, 2}^{2}$ |
| 1 | 1 | 1 | 1,000 | 10 | 10 | 16 | 25 |
| 2 | 1 | 2 | 1,000 | 10 | 10 | 16 | 4 |
| 3 | 1 | 3 | 1,000 | 10 | 10 | 16 | 4 |
| 4 | 2 | 1 | 1,000 | 10 | 10 | 16 | 25 |
| 5 | 2 | 2 | 1,000 | 10 | 10 | 16 | 4 |
| 6 | 2 | 3 | 1,000 | 10 | 10 | 16 | 4 |
| 7 | 3 | 1 | 1,000 | 10 | 10 | 4 | 25 |
| 8 | 3 | 2 | 1,000 | 10 | 10 | 4 | 16 |
| 9 | 3 | 3 | 1,000 | 10 | 10 | 4 | 16 |

The procedure we propose generates a sequence of stratifications which can be represented as a classification tree. Define the level $l$ of a given node $v$ in the tree as the number of arcs in the (unique) chain connecting node $v$ to the root node, and let us indicate with $r_{l}$ the number of nodes sharing the same level $l$. Since only one node will be split at each level, we have $r_{l}=l+1$ for every $l$. At each level $l \geq 0$ the procedure determines a class $F_{l}$ of $r_{l}$ nonempty subpopulations in which $P$ can be partitioned, putting them in a one-to-one correspondence with the nodes of level $l$. The strata in $F_{l}$ are all candidates for being split on any given covariate $X_{k}$, and, following Bethel (1989), the sample allocation is computed which optimally minimizes the survey cost function for the stratification consisting of the unsplit strata in $F_{l}$ and the two substrata which define the current split. The best split at level $l$ is identified as the most favorable in terms of decreasing sample allocation, with respect to that characterizing $F_{l}$, than any other possible split on any of the covariates in use. The optimal allocation corresponding to the stratification defined by such best split, indicated by $n_{b, l+1}$, is taken as the optimal sample size at level $l+1$, and is considered as an upper bound value constraining allocations in the successive level of classification. At initialization, we set $F_{0}=\{P\}$, whose single stratum is thus equivalent to the entire population, and the best sample size $n_{b, 0}$ is computed
as the maximum among those optimal sizes obtained taking into account, separately, every single precision level $\varepsilon_{g}$ set about the $g^{\text {th }}$ surveyed variate:

$$
n_{b, 0}=\max _{g=1, \ldots, G} \frac{N^{2} S_{g}^{2}}{Y_{g}^{2} \varepsilon_{g}^{2}+N S_{g}^{2}},
$$

where $Y_{g}$ is the total estimate for $y_{g}$ on $P$ and $S_{g}^{2}$ is the corresponding variance (see, for the optimum allocation with only one item, Cochran 1977, pages 97-106, and Särndal et al. 1992, pages 104-109).

When $l>0$, the set of strata $F_{l-1}$, optimal at step $l-1$, is analyzed. The best sample allocation at step $l, n_{b, l}$, is initially set equal to $n_{b, l-1}$, and, for each stratum $U \in F_{l-1}$ and every auxiliary variable $X_{k}$, the following algorithm is executed. Let $A_{U}$ be the set of atoms contained in the current stratum $U$, so that $U=\cup A_{U}$ holds true, and let $m(A)$ be a function returning the index assigned to any atom $A\left(m(A)=m_{0}\right.$ if and only if $\left.A=A_{m_{0}}\right)$, then we can express the set of values taken on by $X_{k}$ for units contained in any atom of $A_{U}$ as follows:

$$
Q_{k}=\left\{q \in \mathbb{R}: \exists A \in A_{U}: q=a_{m(A), k}\right\} .
$$

If $X_{k}$ is an ordered variate, for every $q$ in $Q_{k}$ other than $\max \left(Q_{k}\right)$ the stratum $U$ is partitioned into sets $U_{1}=U_{q, 1}$ and $U_{2}=U_{q, 2}$ as follows:

$$
U_{q, 1}=\bigcup\left\{A \in A_{U}: a_{m(A), k} \leq q\right\}
$$

and $U_{q, 2}$ is the relative complement of $U_{q, 1}$ in $U$, i.e., the set of all $i \in U$ which are not in $U_{q, 1}$ :

$$
U_{q, 2}=U \backslash U_{q, 1} .
$$

In our example, for a stratum $U$ defined as $A_{1} \cup A_{2} \cup A_{8}$ we have $A_{U}=\left\{A_{1}, A_{2}, A_{8}\right\}, Q_{1}=\{1,3\}$ and $Q_{2}=\{1,2\}$ (see Table 1), so that our algorithm would try to split $U$ in $U_{1}=A_{1} \cup A_{2}$ and $U_{2}=A_{8}$ using $X_{1}$, and in $U_{1}=A_{1}$ and $U_{2}=A_{2} \cup A_{8}$ using $X_{2}$. If, on the contrary, $X_{k}$ is unordered, $U$ is instead partitioned in sets $U_{1}$ and $U_{2}$ for every proper subset $U_{1}$ of $U$, with $U_{2}=U \backslash U_{1}$.

We thus have a corresponding candidate stratification, namely

$$
C=\left(F_{l-1} \mid\{U\}\right) \cup\left\{U_{1}\right\} \cup\left\{U_{2}\right\},
$$

which includes all the strata in $F_{l-1}$ other than $U$, and, in addition, $U_{1}$ and $U_{2}$. For every stratum $C$ in the collection $C$, the total estimates of $Y_{g}, g=1, \ldots, G$,

$$
Y_{C, g}=\sum_{A \in A_{C}} Y_{m(A), g},
$$

and their corresponding variances

$$
\begin{aligned}
S_{C, g}^{2}=\left(N_{C}-1\right)^{-1} & \left(\sum_{A \in A_{C}}\left(N_{A}-1\right) S_{m(A), g}^{2}\right. \\
& \left.+\sum_{A \in A_{C}} N_{A}\left(N_{A}^{-1} Y_{m(A), g}-N_{C}^{-1} Y_{C, g}\right)^{2}\right),
\end{aligned}
$$

are computed, and the sample allocation $n_{C}$ is thus obtained applying the Bethel's rule. If $n_{C}<n_{b, l}$, then the split $\left(U_{1}, U_{2}\right)$ becomes the current best one, the best stratification candidate $C^{*}$ becomes $C$ and $n_{b, l}$ is updated to $n_{c}$. In this way, the divisive procedure which achieves the best result, i.e., the smallest sample size, is selected to generate the next optimal strata:

$$
F_{l}=C^{*} .
$$

In the framework of our example, for precision levels $\varepsilon_{1}=\varepsilon_{2}=0.1$, let us describe the optimal split at level $l=1$, i.e., that one splitting the entire population in two strata. Using the data described in Table 1, the algorithm indicates that the best split of $U=P$ is based on variable $X_{2}$, and is obtained by setting

$$
\begin{aligned}
U_{1} & =\bigcup\left\{A \in A_{P}: a_{m(A), k} \leq 2\right\} \\
& =A_{1} \cup A_{2} \cup A_{4} \cup A_{5} \cup A_{7} \cup A_{8},
\end{aligned}
$$

and correspondingly $U_{2}=P \backslash U_{1}=A_{3} \cup A_{6} \cup A_{9}$. Such optimal division is represented in Figure 1, where, for every stratum, its size, its definition in terms of included atoms, the current allocation, and the estimation statistics are thoroughly reported.

Issues concerning the optimal number of strata are taken into account by defining the stopping criteria of the tree generating procedure. We decide to stop the algorithm if the relative difference between the optimal sample size at the current level and the optimal one at the previous level is smaller than a given parameter $\delta>0$ :

$$
\begin{equation*}
\delta>\left(n_{b, l-1}-n_{b, l}\right) / n_{b, l-1} . \tag{1}
\end{equation*}
$$

Since the Bethel's algorithm converges to a vector whose range is $(0,+\infty)^{l+1}$, its entries must be rounded to the corresponding nearest integers towards infinity; as a consequence, especially in presence of many small strata, a given allocation is likely to yield a sample size greater than the previous one. Also, in this case, we decided to stop our procedure. To avoid too small and henceforth statistically unstable strata, additional rules can be set to avoid further disaggregations of current strata if the corresponding substrata have cardinalities smaller than a predefined minimum stratum size. Complexities in survey management can also be easily mitigated by imposing a maximum number of strata.


Figure 1 The first optimal split for the example data

This approach, performing an exhaustive search in each single split, guarantees that the corresponding stratification and allocation are optimal, but only conditionally to the splits previously executed. We know that monotonicity of solutions and conditional optimality of each sub-tree obtained by splitting recursively each node are necessary but not sufficient conditions for a binary tree to be optimal. In order to guarantee the overall optimality, to these conditions we should add the requirement that an optimal stratification in, say, $H$ strata, can only be obtained by partitioning one of the nodes of the optimal stratification in $H-1$ strata. In other words, we should assume that an optimal partition in $H$ strata is a subspace of the optimal partition in $H-1$ strata, which implies that partitioning a given stratum will not modify the objective function - i.e., the allocation - in the remaining $H-1$ strata. However, this assumption is rarely true in practical survey applications, since splitting a stratum usually induces a modification of the optimal allocations in all the remaining unsplit strata.

The proposed algorithm, inspired by the sequential and recursive nature of binary trees, can be considered as an heuristic approach to the problem of multivariate stratification, which enables us to detect good, nearly-optimal, strata at the cost of a reasonable computational burden. As a result, this technique is effective in partitioning populations making use of large sets of both continuous and qualitative auxiliary stratifying variables. In addition, the simple structure of binary trees implies a great flexibility in the
introduction of any number of additional constraints, such as lower limits on the number of units in each stratum.

## 3. Forming strata for the Italian Farm Structure Survey

For the requirements of European Community agricultural policies, the Farm Structure Survey (FSS) is executed, every two years, as a census update (Council Regulation (EEC) No 70/66), collecting data on technoeconomic variables characterizing EU farms. It represents the primary source of information for the EUROFARM project (Council Regulation (EEC) No 571/88), a set of data banks to be used for processing Community surveys on the structure of agricultural holdings. Member States are responsible for taking all appropriate steps to carry out the FSS in their territories, and they are also free to select a sampling criterion, but the questionnaire and the precision required, at a national level, for the estimates of the study variables are fixed by Community regulations (see EC Regulations No 837/90 and No 959/93, and subsequent Commission Decisions 1998/377/EC and 2000/115/EC).

To illustrate our stratification technique, we execute the algorithm described in Section 2 to design the italian FSS and identify a tree-based set of strata and allocations using multivariate information. All the algorithms have been implemented by one of the authors in MATLAB language; a Win32 console application has also been developed in C++ to enable software execution in batch mode.The design
exploits the frame of farms listed during the fifth Agricultural General Census held in Italy in the fall of 2000. ISTAT, the Italian national statistical institute, is responsible for updates of such frame based on integration of administrative records, but were not available at the moment of this writing. For the procedure to be initialised, we need a set of atoms into which the population of the italian agricultural holdings must be partitioned. This set of basic strata is obtained by aggregation of farms sharing the same classes of seven covariates. We select four variables related to land use and livestocks, namely utilised agricultural area (UAA), number of bovine animals (NBA), number of pigs (NP), and number of sheep and goats (NSG). To take into account the geographical characteristics of the holdings, we also added, as a stratification variable, the altitude of the farm (ALT). Finally, we collected information about holding administration and organization by means of two variables referred to as legal personality of the holder (LP), and type of tenure of the holding (TT).

Ranges of the covariates concerning the farming structure are divided into four classes for number of bovine animals $(\mathrm{NBA}=0,1 \leq \mathrm{NBA}<10,10 \leq \mathrm{NBA}<50,50 \leq$ NBA), number of pigs ( $\mathrm{NP}=0,1 \leq \mathrm{NP}<500,500 \leq \mathrm{NP}<$ $1,000,1,000 \leq \mathrm{NP})$, and number of sheep and goats $(\mathrm{NSG}=$ $0,1 \leq \mathrm{NSG}<250,250 \leq \mathrm{NSG}<500,500 \leq \mathrm{NSG}$ ), and into seven classes for utilised agricultural area $(\mathrm{UAA}=0$, $0<\mathrm{UAA}<1,1 \leq \mathrm{UAA}<5,5 \leq \mathrm{UAA}<10,10 \leq \mathrm{UAA}<50$, $50 \leq \mathrm{UAA}<100, \mathrm{UAA} \geq 100 \mathrm{ha}$ ). The range of altitude values is divided into five classes: inland mountains, coastal mountains, inland hills, coastal hills, and flat lands. Classes for the legal personality of the holder are defined in order to discriminate among sole holders, legal persons (companies) and groups of physical persons (partnership) in a group holding, cooperative enterprises, associations of holders, public institutions, and, finally, legal personalities other than the previous ones (e.g., consortia), which will be referred to as the residual ones. Holdings are also stratified taking into account their type of tenure, by discerning among ownerfarmed (with further subclasses based on farm labour force categories: family labour, prevalent family labour, prevalent non-family labour), tenant-farmed, shared-farmed agricultural areas, and modes of tenure other than the previous ones. Combining all possible classes from any of the selected covariates leads to 2,964 nonempty atoms, the starting point of the procedure.

We put under study 12 land use variables, whose list is reported in Table 2. For every surveyed variable, totals and variances in each atom are computed elaborating the available Census data, enabling us to execute the Bethel's algorithm at each step of our procedure. Additional parameters needed to identify our stopping criteria are set as follows. The maximum number of strata is defined as 300,
and we decide to disallow strata having a size smaller than 10. A tolerance about the relative difference between optimal sample sizes at subsequent levels is introduced setting $\delta=0$ in equation (1), so the algorithm is stopped if $n_{b, l-1}<n_{b, l}$ for some level $l \geq 0$.

Table 2
Surveyed variables in the Italian farm structure survey and their precision levels

| Surveyed variable | Requested <br> by FSS |  | Required CV <br> Achieved by |  |
| :--- | :---: | :---: | :---: | :---: |
|  |  |  | Atomised <br> stratification | Stratification <br> tree |
| Cereals | 1.00 |  | 0.98 | 0.98 |
| Vineyards | 3.00 |  | 1.38 | 1.38 |
| Olive plants | 3.00 |  | 1.11 | 1.11 |
| Fodder roots and brassicas | 3.00 |  | 2.39 | 2.40 |
| Industrial plants | 3.00 |  | 2.22 | 2.23 |
| Forage plants | 3.00 |  | 1.37 | 1.39 |
| Vegetables | 3.00 |  | 3.03 | 3.03 |
| Fallow land | 3.00 |  | 2.69 | 2.78 |
| Number of Bovine Animals | 1.00 |  | 0.99 | 1.00 |
| Number of Pigs | 2.00 |  | 0.80 | 0.82 |
| Number of Sheep | 2.00 |  | 1.99 | 2.01 |
| Number of Goats | 2.00 |  | 1.92 | 1.98 |

Convergence was achieved since the maximum number of strata was reached and no other stopping rule was activated for $l<300$. Figure 2 shows the optimal allocations $n_{b, l}$ plotted as a function of the number of strata $r_{l}, l=1, \ldots, 300$ on a logarithmic scale, i.e., against $\log \left(n_{b, l}\right)$. It can be noted that the relative difference between subsequent allocations rapidly decreases, with the first ten splits being the more important with respect to such behaviour: in fact, by setting $\delta=10 \%$ the procedure would reach convergence at step $l=7$.

Figure 3 displays a diagram of the stratification tree generated up to level 7. In order to optimize the global allocation, our splitting criterion recursively created smaller and smaller strata. The first split is on the legal personality of the holder, LP, and atoms have been included in the left daughter stratum if the class of variable LP they assume was sole holder, public institution, or a residual one. Such split is the optimal split at level 1 , since it corresponds to a partition of the entire population, the only stratum available at level 0 , that best decreases the sample allocation. This mainly indicates that farms organized by sole holders behave differently from those managed by more complex legal persons, such as companies, partnerships, associations, or cooperative enterprises. The second split is on the number of bovine animals, NBA. It creates two new substrata of stratum 2 (see the bottom side of Figure 3), namely strata 4 and 5 , as follows: the new stratum 4 is defined as the union of such atoms in stratum 2 for which condition NBA> 10
holds true, while stratum 5 is the relative complement of stratum 4 in stratum 2 . In this way, the algorithm detects the best decrement of the overall sample size (passing from $1,570,313$ to 689,404 sampled units, see the right side of Figure 3) by recognizing that farms characterized by medium or large bovine livestocks need to be treated separately for sole held farms. The third split is instead on the utilized agricultural area, UUA. Here, stratum 4 is partitioned between atoms for which variable UUA is less than 100 ha (stratum 6) and remaining ones (stratum 7). Both these new strata are also divided, in successive steps, namely steps 4 to 7 (see the left side of Figure 3), on variables NP and NSG: more thoroughly, the procedure suggests to distinguish farms having no sheep or goat livestocks ( $\mathrm{NSG}=0$ ), or characterized by large livestocks of pigs ( $\mathrm{NP} \geq 500$ ).


Figure 2 Step by Step Sample Sizes. The optimal allocations $n_{b, l}$ are shown as a function of the number of strata $r_{l}$ exploited by the treebased sampling design at steps $l=0, \ldots, 299$. A logarithmic scale is applied to the horizontal axis, so that $n_{b, l}$ is plotted against $\log \left(r_{l}\right)$. As the number of strata increases, the tree-based stratification design attains its goals using a rapidly decreasing global sample size, since the procedure greatly improves the sampling efficiency in its first ten steps of execution

To evaluate the efficiency of the tree-based sampling design, we calculate the best allocation corresponding to the atomised stratification, which determined a sample of 89,522 units. By inspecting the stratification tree, it can be noted that a very similar overall allocation corresponds to the best stratification obtained at level $l=102$ : in fact, for such partition of 103 strata the sample size is equal to 89,509 . This means that, for the same sample size, our algorithm achieves the precision requested for the survey by
exploiting a number of strata, 103, which is a very small fraction of 2,964 , the number of available atoms, henceforth enabling an easier organization of the survey. Another noticeable advantage of our procedure consists in avoiding unstable strata: it is worth noting that 1,618 of the 2,964 atoms have a size equal to or less than 5 , while the minimum size of any of the optimal strata at level 102 is 16 , so that, as a consequence, there is no need to introduce any strata collapsing procedure. Further comparisons can be obtained contrasting the levels of precision achieved implementing, respectively, the atomised stratification and the stratification tree at step 102. Such levels, as reported in Table 2, can be considered very similar for the two designs. In fact, we observed that, for the atomised stratification, the Bethel's allocation was actively constrained on the precision regarding three surveyed variates, namely Cereals, Vegetables and Number of Sheep. With respect to the strata corresponding to level 102 of the tree, the previous constraints also happened to be active, even if another constraint, that on variable Number of Goats, also resulted tight for the optimization, with achieved precision levels increased from $1.92 \%$ to $1.98 \%$. Such findings suggest that, with respect to the atomised partition, the tree can be used to detect a more compact stratification of the population, still preserving the achieved precision levels and the overall sample size.

## 4. Concluding remarks

The tree-based strategy for multipurpose surveys examined in this article is planned to jointly define a rule to partition the population and to allocate sampling units in strata formed exploiting multivariate information, quantitative or qualitative. A hierarchical divisive algorithm selects finer partitions by minimizing, at each step, the sample allocation needed to achieve the required precision levels. In this way, large numbers of constraints can be satisfied without drastically increasing the number of strata. In addition, variables selected for stratification are not discarded merely on the basis of practical considerations, nor the number of their class intervals is diminished. Furthermore, the algorithm avoids creating empty or almost empty strata, thus excluding the need for ex post strata aggregations aimed at a better evaluation of in stratum estimation variances.

Notwithstanding, some points of criticism can be raised about our proposal. Theoretically, our procedure cannot be considered as a multiresponse generalization of the well known classification regression tree method, where the aim is that of exploiting the relationships between the covariates and a unique outcome variable. In fact, even if we deal with
multipurpose surveys, our approach consists in partitioning the available information so as to optimise only one variable, namely the sampling allocation in strata. Furthermore, the sampling strategy obtained through our methodology does not necessarily represent a global optimum: in fact, the procedure constitutes a forward strata selection algorithm, and, as a consequence, the search for optimality at a given step is conditioned on the stratification currently in use, i.e., that one based upon the splits previously executed: there is no guarantee that the stratification selected by the procedure at a certain step $l$ will be the optimal one, even solely among all the possible partitions in $l+1$ subsets of the population. In some situations, the use of other methods such as dynamic programming can be used for conducting an efficient exhaustive search for the globally optimal stratification (see Bühler and Deutler 1975, and Lavallée 1988).

The procedure was applied to redesign the Italian Farm Structure Survey. The results indicate gains in efficiency held using our strategy: for a given sample size, our
procedure achieves the requested precision by exploiting a number of strata which is usually a very small fraction of the number of strata available when combining all possible classes from any of the covariates. In addition, allowing for more strata, the algorithm detects further sampling strategies for which the constraints are satisfied with sample sizes smaller than the one corresponding to the atomised stratification. The final sampling choice obviously depends upon the survey overall cost function. For this purpose, stratification trees can be applied to take into consideration the fact that an increasing number of strata usually implies larger costs due to survey organization issues, but also corresponds to smaller sample sizes, which lead to decreasing unitary costs. Forming strata by trees can thus be useful to manage the survey in an easier way, as a tool to assist the selection of the stratified sampling design which is suited to collect information about the multivariate phenomenon under study.


Figure 3 Stratification Tree Diagram. The bottom side of the horizontal axis is labeled with the stratum identifier, a number that uniquely represents the corresponding subpopulation inside the stratification procedure. Sizes of such strata are reported on the top side. The left side of the vertical axis displays the sequence of steps from 0 to 7, while the right side accounts for the global optimal allocations corresponding to such steps. Double bordered blocks represent split strata. Daughter strata are linked to their parents through elbow lines, and, when not further split in subsequent steps, they are shown as single bordered blocks. For left daughter strata, the covariate on which the split happened and the condition it satisfied when defining the left substratum are reported above the corresponding elbow line. The number inside a given block is the sample allocation the procedure assigns, to the corresponding stratum, during the step at which the block is positioned. Since a stratum can remain unsplit in steps successive to that in which it is created, but its sample allocation can vary from one step to the other, dashed blocks are used to report modifications of stratum sample sizes

## Acknowledgements

The authors wish to thank an Associate Editor and two referees for their valuable comments, which led to a significant improvement of the paper.

## References

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. In Proceedings of the Surveys Research Methods Section, American Statistical Association, 209-212.

Bethel, J. (1989). Sample allocation in multivariate surveys. Survey Methodology, 15, 47-57.

Bloch, D.A., and Segal, M.R. (1989). Empirical comparison of approaches to forming strata - Using classification trees to adjust for covariates. Journal of the American Statistical Association, 84, 408, 897-905.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). Classification and Regression Trees, Belmont, CA: Wadsworth International Group.

Bühler, W., and Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. Metrika, 22, 161-175.

Chromy, J. (1987). Design optimization with multiple objectives. In Proceedings of the Surveys Research Methods Section, American Statistical Association, 194-199.

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley \& Sons, Inc.

Dalenius, T., and Hodges, J.L. Jr. (1959). Minimum variance stratification. Journal of the American Statistical Association. 54, 285, 88-101.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. Survey Methodology, 30, 2, 159-166.

Hedlin, D. (2000). A procedure for stratification by the extended ekman rule. Journal of Official Statistics, 16, 1, 15-29.

Horgan, J.M. (2006). Stratification of skewed populations: A review. International Statistical Review. 74, 1, 67-76.

Jarque, C.M. (1981). A solution to the problem of optimum stratification in multivariate sampling. Applied Statistics. 30, 2, 163-169.

Kish, L., and Anderson, D.W. (1978). Multivariate and multipurpose stratification. Journal of the American Statistical Association, 73, 361, 24-34.

Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. Proceedings of the Section on Survey Research Methods, American Statistical Association. Virginia, August 1988, 646-651.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.

Lu, W., and Sitter, R.R. (2002). Multi-way stratification by linear programming made practical. Survey Methodology, 28, 2, 199207.

Mulvey, J.M. (1983). Multivariate stratified sampling by optimization. Management Science, 29, 6, 715-724.

Pla, L. (1991). Determining stratum boundaries with multivariate real data. Biometrics, 47, 4, 1409-1422.

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

Sigman, R.S., and Monsour, N.J. (1995). Selecting samples from list frames of businesses. In Business Survey Methods, (Eds. B.G. Cox, D.A. Binder, B. Nanjamma Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott). New York: John Wiley \& Sons, Inc., 133-152.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. Journal of the American Statistical Association, 66, 336, 829-833.

Valliant, R., and Gentle, J.E. (1997). An application of mathematical programming to sample allocation. Computational Statistics \& Data Analysis, 25, 337-360.

Vogel, F.A. (1995). The evolution and development of agricultural statistics at the United States department of agriculture. Journal of Official Statistics, 11, 2, 161-180.

# Determining the optimum strata boundary points using dynamic programming 

Mohammad G. Mostafa Khan, Niraj Nand and Nesar Ahmad ${ }^{1}$


#### Abstract

Optimum stratification is the method of choosing the best boundaries that make strata internally homogeneous, given some sample allocation. In order to make the strata internally homogenous, the strata should be constructed in such a way that the strata variances for the characteristic under study be as small as possible. This could be achieved effectively by having the distribution of the main study variable known and create strata by cutting the range of the distribution at suitable points. If the frequency distribution of the study variable is unknown, it may be approximated from the past experience or some prior knowledge obtained at a recent study. In this paper the problem of finding Optimum Strata Boundaries (OSB) is considered as the problem of determining Optimum Strata Widths (OSW). The problem is formulated as a Mathematical Programming Problem (MPP), which minimizes the variance of the estimated population parameter under Neyman allocation subject to the restriction that sum of the widths of all the strata is equal to the total range of the distribution. The distributions of the study variable are considered as continuous with Triangular and Standard Normal density functions. The formulated MPPs, which turn out to be multistage decision problems, can then be solved using dynamic programming technique proposed by Bühler and Deutler (1975). Numerical examples are presented to illustrate the computational details. The results obtained are also compared with the method of Dalenius and Hodges (1959) with an example of normal distribution.


Key Words: Stratified random sampling; Optimum stratification; Triangular distribution; Standard normal distribution; Mathematical programming problem; Multistage decision problem; Dynamic programming technique.

## 1. Introduction

The basic consideration involved in the determination of optimum strata boundaries (OSB) is that the strata should be internally as homogenous as possible, that is, the stratum variances $\sigma_{h}^{2}$ should be as small as possible, given some sample allocation. When a single characteristic is under study and the distribution of the study variable is available, the OSB can be determined by cutting the range of this distribution at suitable points. This problem of determining the OSB was first discussed by Dalenius (1950), when the study variable itself is used as stratification variable. He presented a set of minimal equations that could be solved for finding OSB. Unfortunately these equations could not usually be solved because of their implicit nature. Hence attempts have been made by several authors to obtain the approximate strata boundaries using classical methods. Given the number of strata, Dalenius and Gurney (1951) suggested that the strata boundaries be determined when $W_{h} \sigma_{h}$ remain constant, where $W_{h}$ is the weight of stratum h. Mahalanobis (1952) and Hansen and Hurwitz (1953) have suggested that the strata boundaries can be determined when $W_{h} \mu_{h}$ remain constant. Aoyama (1954) suggested an approximate rule and recommended to make strata of equal width $x_{h}-x_{h-1}$, where $x_{h-1}$ and $x_{h}$ are the boundaries of stratum $h$. Ekman (1959) determined the strata boundaries with the condition that $W_{h}\left(x_{h}-x_{h-1}\right)=$ constant. Dalenius
and Hodges (1959) recommended to construct the strata by taking equal intervals on the cumulative of $\sqrt{f(x)}$. Sethi (1963) proposed a method to work out the boundaries given by the calculus equations

$$
\frac{\left(x_{h}-\mu_{h}\right)^{2}+\sigma_{h}^{2}}{\sigma_{h}}=\frac{\left(x_{h+1}-\mu_{h+1}\right)^{2}+\sigma_{h+1}^{2}}{\sigma_{h+1}}
$$

for a standard continuous distribution resembling the study population.

In a comparison on some of the classical approximate methods, the Ekman method and the Dalenius and Hodges method are proved to work consistently well (see Cochran 1961, Hess, Sethi and Balakrishnan 1966, Murthy 1967) but the later is more convenient and easier to apply (see Nicoloni 2001).

Unnithan (1978) suggested an iterative method using Shanno's modified Newton method for determining the strata boundaries that leads to a local minimum of the variance for Neyman allocation, if a suitable initial solution is chosen. The procedure is proved to be faster than the Dalenius and Hodges iterative procedure. Later on Unnithan and Nair (1995) gave a method of selecting an appropriate starting point for modified Newton method that may lead to a global minimum of the variance.

Lavallée and Hidiroglou (1988) proposed an algorithm to construct stratum boundaries for a power allocated stratified sample of non-certainty sample units. Hidiroglou and

[^8]Srinath (1993) presented a more general form of the algorithm, which by assigning different values to operating parameters yields a power allocation, a Neyman allocation, or a combination of these allocations. Sweet and Sigman (1995) and Rivest (2002) reviewed Lavallée and Hidiroglou algorithm and proposed their modified versions of the algorithm that incorporate the different relationships between the stratification and study variables. Detlefsen and Veum (1991) investigated the Lavallée and Hidiroglou algorithm for several strata and observed that the algorithm's convergence was slow or non-existent. They also found that different starting points lead to different OSBs for the same population.

Niemiro (1999) proposed a random search method in the stratification problem but the algorithm did not guarantee that it leads to global optimum. Furthermore, it would go wrong in a case of a large population, as it requires too many iteration steps (see Kozak 2004).

Nicolini (2001) suggested a method, named Natural Class Method (NCM), to oppose the most utilized Dalenius and Hodges method but neither method was proved to be more efficient than other.

Lednicki and Wieczorkowski (2003) presented a method of stratification using the simplex method of Nelder and Mead (1965). Later Kozak (2004) presented the modified random search algorithm as a method of the optimal stratification. The Kozak algorithm was quite faster and efficient as compared to Rivest, and Lednicki and Wieczorkowski but it could not guarantee that the algorithm leads to the global optimum.

Bühler and Deutler (1975) formulated the problem of determining OSB as an optimization problem that can be solved by a dynamic programming technique. This approach is also used by Lavallée $(1987,1988)$ for determining the OSB which would divide the population domain of two stratification variables into distinct subsets such that the precision of the variables of interest is maximized.

Khan, Khan and Ahsan (2002) considered the problem of finding OSB as an equivalent problem of determining Optimum Strata Width (OSW). The authors formulated the problem of OSW as a Mathematical Programming Problem (MPP). Following the Bühler and Deutler's dynamic programming approach, they solve the MPP that gives exact solution, if the frequency distribution of the study variable is known and the number of strata is fixed in advance. Khan et al. (2002) applied their procedure to work out OSB to the population having uniform and right triangular distribution. Later Khan, Najmussehar and Ahsan (2005) extended this dynamic programming approach for determining the OSB for an exponential study variable also.

In this paper the problem of determining OSB for the study variables with Triangular and Standard Normal distributions are discussed. Viewing the fact that these problems are equivalent to the problems of determining OSW, we formulate the problems as MPPs and solve them by following Bühler and Deutler's dynamic programming approach. The formulated MPPs minimize the variance of the estimated population parameter under Neyman allocation subjected to a restriction that sum of the widths of all the strata is equal to the total range of the distribution of the study variable. In Section 2, a review of dynamic programming approach proposed by Bühler and Deutler (1975) is presented. In Section 3, the details of the formulation of the problems of OSW as MPPs are provided. The solution procedure using dynamic programming technique to solve the MPPs is discussed in Section 4. The computational details of the solution procedure is illustrated with numerical examples in Section 5. Finally, in Section 6, an investigation is carried out to compare the results obtained by the dynamic programming method and the cum $\sqrt{f}$ method of Dalenius and Hodges (1959) with an example from a population of normal distribution. It reveals that the proposed dynamic programming method yields a gain in efficiency over the cum $\sqrt{f}$ method.

## 2. Determination of OSB using dynamic programming techniques: A review of Bühler and Deutler's approach

Let $X$ be a random study variable, discrete or continuous, with probability density function $f(x), a \leq x \leq b$. To estimate the population mean $\mu$ by a stratified sample, $X$ is partitioned into $L$ strata $\left[a, x_{1}\right],\left(x_{1}, x_{2}\right], \ldots,\left(x_{L-1}, b\right]$ such that

$$
\begin{equation*}
a=x_{0} \leq x_{1} \leq x_{2} \leq, \ldots, \leq x_{L-1} \leq x_{L}=b . \tag{1}
\end{equation*}
$$

Suppose that from stratum $h(h=1,2, \ldots, L)$, which contains $N_{h}$ units, a sample of size $n_{h}$ with units $y_{h j}$ $\left(h=1,2, \ldots, L ; j=1,2, \ldots, n_{h}\right) \quad$ is selected. Then the stratified mean $\bar{x}_{s t}=\sum_{h=1}^{L} W_{h} \bar{x}_{h}$ is an unbiased estimate of $\mu$ with variance

$$
\begin{equation*}
V\left(\bar{x}_{s t}\right)=\sum_{h=1}^{L} W_{h} \sigma_{h}^{2}\left(\frac{W_{h}}{n_{h}}-\frac{1}{N}\right) \tag{2}
\end{equation*}
$$

where $W_{h}=N_{h} / N, \bar{x}_{h}=1 / n_{h} \sum_{j=1}^{n_{h}} y_{h j}, \sigma_{h}^{2}=\left[1 /\left(N_{h}-1\right)\right] \times$ $\sum_{j=1}^{N_{h}}\left(y_{h j}-\mu_{h}\right)^{2}$ and $\mu_{h}=1 / N_{h} \sum_{j=1}^{N_{h}} y_{h j}$.

When the frequency function $f(x)$ is known, the values of $W_{h}$ and $\sigma_{h}$ in (2) can be obtained by

$$
\begin{equation*}
W_{h}=\int_{x_{h-1}}^{x_{h}} f(x) d x \tag{3}
\end{equation*}
$$

$$
\begin{equation*}
\sigma_{h}^{2}=\frac{1}{W_{h}} \int_{x_{h-1}}^{x_{h}} x^{2} f(x) d x-\mu_{h}^{2}, \tag{4}
\end{equation*}
$$

where

$$
\begin{equation*}
\mu_{h}=\frac{1}{W_{h}} \int_{x_{h-1}}^{x_{h}} x f(x) d x \tag{5}
\end{equation*}
$$

is the mean and $\left(x_{h-1}, x_{h}\right)$ are the boundaries of $h^{\text {th }}$ stratum.

Then (2) reads as the function of strata boundary points and sample sizes, that is,

$$
V\left(\bar{x}_{s t}\right)=V\left(\bar{x}_{s t} \mid x_{1}, \ldots, x_{L-1}, n_{1}, \ldots, n_{L}\right)
$$

If $n_{h}$ are fixed, the objective of the optimum stratification is to determine stratum boundary points $\left(x_{1}, \ldots, x_{L-1}\right)$ such that $V\left(\bar{x}_{s t}\right)$ is minimum. Further, if the sampling ratios $n_{h} / N_{h}$ are small or the sampling is with replacement, then the following optimization problems are obtained, depending on the type of allocation of total sample size $\left(n=\sum_{h=1}^{L} n_{h}\right)$ to strata.

1. Proportional allocation $\left(n_{h}=n \cdot W_{h}\right)$

Minimize $\sum_{h=1}^{L} W_{h} \sigma_{h}^{2}$
subject to $a=x_{0} \leq x_{1} \leq x_{2} \leq, \ldots, \leq x_{L-1} \leq x_{L}=b$
2. Equal allocation $\left(n_{h}=n / L\right)$

Minimize $\sum_{h=1}^{L} W_{h}^{2} \sigma_{h}^{2}$
subject to $a=x_{0} \leq x_{1} \leq x_{2} \leq, \ldots, \leq x_{L-1} \leq x_{L}=b$
3. Neyman allocation $\left(n_{h}=n \cdot W_{h} \sigma_{h} / \sum_{h=1}^{L} W_{h} \sigma_{h}\right)$

Minimize $\sum_{h=1}^{L} W_{h} \sigma_{h}$
subject to $a=x_{0} \leq x_{1} \leq x_{2} \leq, \ldots, \leq x_{L-1} \leq x_{L}=b$.

The problems (6) to (8) have the following structure:

$$
\begin{align*}
& \text { Minimize } \sum_{h=1}^{L} \phi_{h}\left(x_{h-1}, x_{h}\right) \\
& \text { subject to } a=x_{0} \leq x_{1} \leq x_{2} \leq, \ldots, \leq x_{L-1} \leq x_{L}=b \tag{9}
\end{align*}
$$

Bühler and Deutler (1975) have suggested a recursive optimization method for solving (9) using a dynamic programming technique as follows:

Consider an optimization problem with the special structure:

$$
\begin{array}{ll}
\operatorname{Minimize} & \sum_{h=1}^{m} u_{h}\left(z_{h-1}, y_{h}\right) \\
\text { subject to } & z_{h}=v_{h}\left(z_{h-1}, y_{h}\right) \\
& z_{h} \in Z_{h} \\
& y_{h} \in S_{h}\left(z_{h-1}\right) \\
& z_{0}=z^{\prime} ; h=1,2, \ldots, m \tag{10}
\end{array}
$$

where $m=$ number of stages, $u_{h}=$ stage return functions, $v_{h}=$ stage transformation functions, $Z_{h}=$ state spaces, $S_{h}=$ decision spaces, and $z^{\prime}=$ initial state. Then a dynamic programming procedure using Bellman's principle of optimality (Bellman 1957) can be used to solve (10).

If $m=L, z_{0}=a, z_{L}=b, Z_{h}=[a, b], Z_{h-1}=\left[a, b-y_{h}\right]$, $S_{h}\left(z_{h-1}\right)=\left[0, b-z_{h-1}\right]$ with $z_{h-1} \in Z_{h-1}, u_{h}\left(z_{h-1}, y_{h}\right)=$ $\phi_{h}\left(z_{h-1}, y_{h}+z_{h-1}\right)$ with $y_{h} \in S_{h}\left(z_{h-1}\right), v_{h}\left(z_{h-1}, y_{h}\right)=$ $y_{h}+z_{h-1}$, then (10) is transformed to the following problem:

$$
\begin{array}{ll}
\text { Minimize } & \sum_{h=1}^{L} \phi_{h}\left(z_{h-1}, y_{h}+z_{h-1}\right) \\
\text { subject to } & z_{h}=y_{h}+z_{h-1} \\
& z_{h} \in[a, b] \\
& y_{h} \in\left[0, b-z_{h-1}\right] \\
& z_{0}=a, z_{L}=b ; h=1,2, \ldots, L \tag{11}
\end{array}
$$

The problem (11) is an equivalent problem of (9) as they hold the following results:

1. If $\left(x_{1}^{*}, \ldots, x_{L-1}^{*}\right)$ is an optimum solution of (9), then $y_{h}^{*}=x_{h}^{*}-x_{h-1}^{*}, z_{h}^{*}=x_{h}^{*}$ is an optimum of (11).
2. If $y_{h}^{*}(h=1, \ldots, L), z_{h}^{*}(h=1, \ldots, L-1)$ is an optimum solution of (11), then $x_{h}^{*}=z_{h}^{*}(h=1, \ldots, L-1)$ is an optimum solution of (9).

If $\Phi_{h}\left(z_{h-1}\right)$ is the optimum value of objective function at stage $h$ with the available state $z_{h-1}$, then the backward recursive equation to solve (11) using a dynamic programming technique is given by

$$
\begin{array}{r}
\Phi_{h}\left(z_{h-1}\right)=\min \left[\phi_{h}\left(z_{h-1}, y_{h}+z_{h-1}\right)+\Phi_{h+1}\left(z_{h}\right) \mid\right. \\
\left.z_{h}=y_{h}+z_{h-1}\right] \tag{12}
\end{array}
$$

on $y_{h} \in S_{h}\left(z_{h-1}\right)$ with initially $\Phi_{L+1} \equiv 0$.

## 3. Formulation of the problem of OSW as an MPP

In this section the Bühler and Deutler's approach discussed above is extended for a study variable with a continuous density function $f(x)$. The problem (11) is transformed into an equivalent problem of determining OSW by considering $y_{h}=z_{h}-z_{h-1}=x_{h}-x_{h-1}$ as strata widths and then the objective function and the constraints are constructed as functions of $y_{h}$. The MPP is treated as a multistage decision problem in which at each stage the value of the OSW and hence the OSB for a stratum is worked out using dynamic programming technique with a forward recursive equation.

Let $f(x)$ be the frequency function and $x_{0}$ and $x_{L}$ are the smallest and largest values of $x$. If the population mean is estimated under Neyman allocation, then the problem of determining the strata boundaries is to cut up the range,

$$
\begin{equation*}
x_{L}-x_{0}=d \tag{13}
\end{equation*}
$$

at intermediate points $x_{1} \leq x_{2} \leq, \ldots, \leq x_{L-1}$ such that $\sum_{h=1}^{L} W_{h} \sigma_{h}$ in (8) is minimum.

Consider that $f(x)$ has $n$ piece-wise continuous linear or non-linear functions as follows:

$$
f(x)=\left\{\begin{align*}
g_{1}(x) ; & x_{0}=a_{0} \leq x \leq a_{1}  \tag{14}\\
g_{2}(x) ; & a_{1}<x \leq a_{2} \\
& \vdots \\
g_{n}(x) ; & a_{n-1}<x \leq a_{n}=x_{L}
\end{align*}\right.
$$

Also assume that out of $L$ strata, $l_{i}$ be the number of strata to be formed under the density function $g_{i}(x)$; $i=1,2, \ldots, n$ and $\sum_{i=1}^{n} l_{i}=L$.

If $f(x)$ in (14) is integrable, using the expressions (3), (4) and (5), $W_{h}, \sigma_{h}^{2}$ and $\mu_{h}$ are obtained as a function of the boundary points $x_{h}$ and $x_{h-1}$. Thus the objective function in (8) could be expressed as a function of boundary points $x_{h}$ and $x_{h-1}$ only. Let

$$
\phi_{h}\left(x_{h}, x_{h-1}\right)=W_{h} \sigma_{h} .
$$

Note that the above function has different values for different density functions in (14).

Thus, the problem (8) can be treated as an optimization problem to find $x_{1}, x_{2}, \ldots, x_{L-1}$ as stated in (9).

Let $y_{h}=x_{h}-x_{h-1} \geq 0$ denote the width of the $h^{\text {th }}$ $(h=1,2, \ldots, L)$ stratum.

With the above definition of $y_{h}$, the range of the distribution given in (13) is expressed as the function of the stratum widths as:

$$
\begin{equation*}
\sum_{h=1}^{L} y_{h}=\sum_{h=1}^{L}\left(x_{h}-x_{h-1}\right)=x_{L}-x_{0}=d \tag{15}
\end{equation*}
$$

The $k^{\text {th }}$ stratification point $x_{k} ; k=1,2, \ldots, L-1$ is then expressed as:

$$
\begin{aligned}
x_{k} & =x_{0}+y_{1}+y_{2}+\ldots+y_{k} \\
& =x_{k-1}+y_{k},
\end{aligned}
$$

which is a function of $k^{\text {th }}$ stratum width and $(k-1)^{\text {th }}$ stratum boundary.

Considering $z_{k}=x_{k}$ and adding (15) as a constraint, the problem (11) can be rewritten as an equivalent problem of determining OSW as:

$$
\begin{align*}
& \text { Minimize } \sum_{h=1}^{L} \phi_{h}\left(y_{h}, x_{h-1}\right), \\
& \text { subject to } \sum_{h=1}^{L} y_{h}=d, \\
& \text { and } \quad y_{h} \geq 0 ; h=1,2, \ldots, L . \tag{16}
\end{align*}
$$

Initially, $x_{0}$ is known. Therefore, the first term, that is, $\phi_{1}\left(y_{1}, x_{0}\right)$ in the objective function of the MPP (16) is a function of $y_{1}$ alone. Once $y_{1}$ is known, the next stratification point $x_{1}=x_{0}+y_{1}$ will be known and the second term in the objective function $\phi_{2}\left(y_{2}, x_{1}\right)$ will become a function of $y_{2}$ alone.

Therefore, stating the objective function as a function of $y_{h}$ alone the MPP (16) is expressed as:

$$
\begin{array}{ll}
\text { Minimize } & \sum_{h=1}^{L} \phi_{h}\left(y_{h}\right) \\
\text { subject to } & \sum_{h=1}^{L} y_{h}=d, \\
\text { and } & y_{h} \geq 0 ; h=1,2, \ldots, L .
\end{array}
$$

The Sections 3.1 and 3.2 illustrate the formulation of the problem of determining OSW as an MPP for Triangular and Standard Normal study variables respectively.

### 3.1 MPP for triangular distribution

Let the study variable $x$ be following the Triangular distribution on the interval $[\mathrm{a}, \mathrm{b}]$ with the probability density function:

$$
f(x)=\left\{\begin{array}{cl}
\frac{2(x-a)}{(b-a)(c-a)} ; & a \leq x \leq c  \tag{18}\\
\frac{2(b-x)}{(b-a)(b-c)} ; & c<x \leq b
\end{array}\right.
$$

where $a$ is a location parameter, $b$ is a scale parameter and $c$ is the shape parameter.

It has two piece-wise functions.
When $a \leq x \leq c$, from (18) and using (3), (5) and (4), $W_{h}, \mu_{h}$, and $\sigma_{h}^{2}$ are obtained as:

$$
\begin{align*}
& W_{h}=\frac{y_{h}\left(y_{h}+2 a_{h}\right)}{(b-a)(c-a)},  \tag{19}\\
& \mu_{h}=\frac{\frac{2}{3} y_{h}^{2}+2 y_{h} x_{h-1}-a y_{h}+2 a_{h} x_{h-1}}{y_{h}+2 a_{h}},
\end{align*}
$$

and

$$
\begin{equation*}
\sigma_{h}^{2}=\frac{y_{h}^{2}\left[y_{h}^{2}+6 a_{h} y_{h}+6 a_{h}^{2}\right]}{18\left(y_{h}+2 a_{h}\right)^{2}}, \tag{20}
\end{equation*}
$$

where $y_{h}=x_{h}-x_{h-1}, a_{h}=x_{h-1}-a$ and $a \leq x_{h-1} \leq x_{h} \leq c$.
Thus from (19) and (20),

$$
\begin{equation*}
W_{h} \sigma_{h}=\frac{y_{h}^{2} \sqrt{y_{h}^{2}+6 a_{h} y_{h}+6 a_{h}^{2}}}{3 \sqrt{2}(b-a)(c-a)} . \tag{21}
\end{equation*}
$$

Similarly, when $c<x \leq b$, from (18) and using (3), (5) and (4), it can be demonstrated that

$$
\begin{align*}
& W_{h}=\frac{y_{h}\left(2 b_{h}-y_{h}\right)}{(b-a)(b-c)},  \tag{22}\\
& \mu_{h}=\frac{3 b_{h} y_{h}-3 y_{h} x_{h-1}+6 b_{h} x_{h-1}-2 y_{h}^{2}}{3\left(2 b_{h}-y_{h}\right)},
\end{align*}
$$

and

$$
\begin{equation*}
\sigma_{h}^{2}=\frac{y_{h}^{2}\left(6 b_{h}^{2}-6 b_{h} y_{h}+y_{h}^{2}\right)}{18\left(2 b_{h}-y_{h}\right)^{2}}, \tag{23}
\end{equation*}
$$

where $y_{h}=x_{h}-x_{h-1}, b_{h}=b-x_{h-1}$ and $c<x_{h-1} \leq x_{h} \leq b$.
Thus, from (22) and (23),

$$
\begin{equation*}
W_{h} \sigma_{h}=\frac{y_{h}^{2} \sqrt{6 b_{h}^{2}-6 b_{h} y_{h}+y_{h}^{2}}}{3 \sqrt{2}(b-a)(b-c)} . \tag{24}
\end{equation*}
$$

Let $\lambda_{1}$ and $\lambda_{2}$ be the last and the first stratum formed under the first and second piece-wise function of (18) respectively. If any stratum (say, $l$ ) falls under both functions, then $\lambda_{1}$ and $\lambda_{2}$ are not considered to be two different strata but the fractions of the same $l^{\text {th }}$ stratum. Then, using (21) and (24) the MPP (17) could be expressed as the problem of determining the OSW for the study variable with Triangular frequency function as:

$$
\begin{aligned}
\text { Minimize } & \left\{\sum_{h=1}^{\lambda_{1}} \frac{y_{h}^{2} \sqrt{y_{h}^{2}+6 a_{h} y_{h}+6 a_{h}^{2}}}{3 \sqrt{2}(b-a)(c-a)}\right. \\
& \left.+\sum_{h=\lambda_{2}}^{L} \frac{y_{h}^{2} \sqrt{6 b_{h}^{2}-6 b_{h} y_{h}+y_{h}^{2}}}{3 \sqrt{2}(b-a)(b-c)}\right\},
\end{aligned}
$$

subject to $\sum_{h=1}^{L} y_{h}=d$,

$$
\begin{equation*}
\text { and } \quad y_{h} \geq 0 ; h=1,2, \ldots, L \text {, } \tag{25}
\end{equation*}
$$

where $d=b-a$.

### 3.2 MPP for normal distribution

The study variable $x$ is said to have a Standard Normal distribution if its probability density function is given by

$$
f(x)=\frac{1}{\sqrt{2 \pi}} \exp \left(-\frac{x^{2}}{2}\right) ; \quad-\infty<x<\infty .
$$

As in section 3.1, using the definition (3), (5) and (4), it can be seen that

$$
\begin{align*}
W_{h} & =\frac{\operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)}{2},  \tag{26}\\
\mu_{h} & =\frac{\sqrt{2}\left[\exp \left(-\frac{x_{h-1}^{2}}{2}\right)-\exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right)\right]}{\sqrt{\pi}\left[\operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right]},
\end{align*}
$$

and

$$
\begin{align*}
\sigma_{h}^{2} & =\left\{\sqrt { 2 \pi } \left[x_{h-1} \exp \left(-\frac{x_{h-1}^{2}}{2}\right) \operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)\right.\right. \\
& -\left(y_{h}+x_{h-1}\right) \exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right) \\
& -x_{h-1} \exp \left(-\frac{x_{h-1}^{2}}{2}\right) \operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right) \\
& \left.+\left(y_{h}+x_{h-1}\right) \exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right] \\
& +\pi\left[\operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right]^{2} \\
& \left.-2\left[\exp \left(-\frac{x_{h-1}^{2}}{2}\right)-\exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right)\right]^{2}\right\} \\
& \div \pi\left[\operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right]^{2} \tag{27}
\end{align*}
$$

where $\quad y_{h}=x_{h}-x_{h-1}, \operatorname{erf}\left(x_{h}\right)-\operatorname{erf}\left(x_{h-1}\right)=(2 / \sqrt{\pi}) \times$ $\int_{x_{l-1}}^{x_{n}} \exp \left(-u^{2}\right) d u$ and $h=1,2, \ldots, L$.

Therefore, using the values in (26) and (27) the MPP (17) can be expressed as

$$
\begin{aligned}
\text { Minimize } & \sum_{h=1}^{L} \operatorname{Sqrt}\left\{\frac { 1 } { 2 \sqrt { 2 \pi } } \left[x_{h-1} \exp \left(-\frac{x_{h-1}^{2}}{2}\right) \operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)\right.\right. \\
& -\left(y_{h}+x_{h-1}\right) \exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right) \\
& -x_{h-1} \exp \left(-\frac{x_{h-1}^{2}}{2}\right) \operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right) \\
& \left.+\left(y_{h}+x_{h-1}\right) \exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right] \\
& +\frac{1}{4}\left[\operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right]^{2} \\
& \left.-\frac{1}{2 \pi}\left[\exp \left(-\frac{x_{h-1}^{2}}{2}\right)-\exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right)\right]^{2}\right\}
\end{aligned}
$$

subject to $\sum_{h=1}^{L} y_{h}=d$
and $\quad y_{h} \geq 0 ; h=1,2, \ldots, L$.

## 4. The solution procedure using dynamic programming technique

The MPP (17) is a multistage decision problem in which the objective function and the constraints are separable of $y_{h}$, which allow us to use a dynamic programming technique as illustrated by Bühler and Deutler (1975) for the problem (11).

Consider the following subproblem of (17) for first $k(<L)$ strata:

$$
\begin{array}{ll}
\text { Minimize } & \sum_{h=1}^{k} \phi_{h}\left(y_{h}\right), \\
\text { subject to } & \sum_{h=1}^{k} y_{h}=d_{k}, \\
\text { and } \quad y_{h} \geq 0 ; h=1,2, \ldots, k, \tag{29}
\end{array}
$$

where $d_{k}<d$ is the total width available for division into $k$ strata or the state value at stage $k$. Note that $d_{k}=d$ for $k=L$.

The transformation functions are given by

$$
\begin{aligned}
d_{k} & =y_{1}+y_{2}+\ldots+y_{k}, \\
d_{k-1} & =y_{1}+y_{2}+\ldots+y_{k-1}=d_{k}-y_{k}, \\
d_{k-2} & =y_{1}+y_{2}+\ldots+y_{k-2}=d_{k-1}-y_{k-1}, \\
& \vdots \quad \vdots \\
d_{2} & =y_{1}+y_{2}=d_{3}-y_{3}, \\
d_{1} & =y_{1}=d_{2}-y_{2} .
\end{aligned}
$$

Let $\Phi_{k}\left(d_{k}\right)$ denote the minimum value of the objective function of (29), that is,

$$
\begin{aligned}
& \Phi_{k}\left(d_{k}\right)=\min \left[\sum_{h=1}^{k} \phi_{h}\left(y_{h}\right) \mid \sum_{h=1}^{k} y_{h}=d_{k}\right. \\
&\text { and } \left.y_{h} \geq 0 ; h=1,2, \ldots, k\right] .
\end{aligned}
$$

With the above definition of $\Phi_{k}\left(d_{k}\right)$, the MPP (17) is equivalent to finding $\Phi_{L}(d)$ recursively by finding $\Phi_{k}\left(d_{k}\right)$ for $k=1,2, \ldots, L$ and $0 \leq d_{k} \leq d$.

We can write:

$$
\begin{array}{r}
\Phi_{k}\left(d_{k}\right)=\min \left[\phi_{k}\left(y_{k}\right)+\sum_{h=1}^{k-1} \phi_{h}\left(y_{h}\right) \sum_{h=1}^{k-1} y_{h}=d_{k}-y_{k},\right. \\
\text { and } \left.y_{h} \geq 0 ; h=1,2, \ldots, k-1 .\right]
\end{array}
$$

For a fixed value of $y_{k} ; 0 \leq y_{k} \leq d_{k}$,

$$
\begin{aligned}
\Phi_{k}\left(d_{k}\right)= & \phi_{k}\left(y_{k}\right) \\
& +\min \left[\sum_{h=1}^{k-1} \phi_{h}\left(y_{h}\right) \mid \sum_{h=1}^{k-1} y_{h}=d_{k}-y_{k},\right. \\
& \left.\quad \text { and } y_{h} \geq 0 ; h=1,2, \ldots, k-1\right] .
\end{aligned}
$$

Using the Bellman's principle of optimality, we write a forward recursive equation, instead of backward recursive equation as suggested by Bühler and Deutler in (12), for using dynamic programming technique as:

$$
\begin{equation*}
\Phi_{k}\left(d_{k}\right)=\min _{0 \leq y_{k} \leq d_{k}}\left[\phi_{k}\left(y_{k}\right)+\Phi_{k-1}\left(d_{k}-y_{k}\right)\right], k \geq 2 . \tag{30}
\end{equation*}
$$

For the first stage, that is, for $k=1$ :

$$
\begin{equation*}
\Phi_{1}\left(d_{1}\right)=\phi_{1}\left(d_{1}\right) \Rightarrow y_{1}^{*}=d_{1}, \tag{31}
\end{equation*}
$$

where $y_{1}^{*}=d_{1}$ is the optimum width of the first stratum. The relations (30) and (31) are solved recursively for each $k=1,2, \ldots, L$ and $0 \leq d_{k} \leq d$, and $\Phi_{L}(d)$ is obtained. From $\Phi_{L}(d)$ the optimum width of $L^{\text {th }}$ stratum, $y_{L}^{*}$, is obtained. From $\Phi_{L-1}\left(d-y_{L}^{*}\right)$ the optimum width of $(L-1)^{\text {th }}$ stratum, $y_{L-1}^{*}$, is obtained and so on until $y_{1}^{*}$ is obtained.

Note that depending upon the piece-wise function(s) in (14) under which the stratum is formed, $\phi_{k}\left(y_{k}\right)$ in (30) will take different value for each $y_{k}$ as follows:

$$
y_{k}=x_{k}-x_{k-1} \leq a_{i}-a_{0}, \text { for some } i(i=1,2, \ldots, n)
$$

and,

$$
x_{k} \in\left[a_{i-1}, a_{i}\right], \quad \text { for some } i(i=1,2, \ldots, n)
$$

## 5. Numerical illustrations

In this section the computational details of the solution procedure discussed in section 4 for the MPPs (25) and (28) are presented.

### 5.1 Triangular distribution

Let us assume that $a=x_{0}=0, c=1$ and $b=x_{L}=2$. This implies that $d=x_{L}-x_{0}=2$ and the MPP (25) is expressed as:

$$
\begin{aligned}
\text { Minimize } & \left\{\sum_{h=1}^{\lambda_{1}} \frac{y_{h}^{2} \sqrt{y_{h}^{2}+6 a_{h} y_{h}+6 a_{h}^{2}}}{6 \sqrt{2}}\right. \\
& \left.+\sum_{h=\lambda_{2}}^{L} \frac{y_{h}^{2} \sqrt{6 b_{h}^{2}-6 b_{h} y_{h}+y_{h}^{2}}}{6 \sqrt{2}}\right\},
\end{aligned}
$$

subject to $\sum_{h=1}^{L} y_{h}=2$,
and $\quad y_{h} \geq 0 ; h=1,2, \ldots, L$,
where $a_{h}=x_{h-1}$ and $b_{h}=2-x_{h-1}$.
Using (30) and (31), the recursive equations for solving MPP (32) can be stated as:

For the first stage $(k=1)$

$$
\begin{equation*}
\Phi_{1}\left(d_{1}\right)=\frac{d_{1}^{3}}{6 \sqrt{2}} \quad \text { at } \quad y_{1}=d_{1} \tag{33}
\end{equation*}
$$

and for the stages $(k \geq 2)$

$$
\Phi_{k}\left(d_{k}\right)=
$$

$$
\left\{\begin{array}{c}
\min \left[\frac{y_{k}^{2} \sqrt{y_{k}^{2}+6 a_{k} y_{k}+6 a_{k}^{2}}}{6 \sqrt{2}}+\Phi_{k-1}\left(d_{k}-y_{k}\right)\right] \\
\text { if } 0 \leq d_{k} \leq 1, \\
\min \left[\frac{y_{k}^{2} \sqrt{6 b_{k}^{2}-6 b_{k} y_{k}+y_{k}^{2}}}{6 \sqrt{2}}+\Phi_{k-1}\left(d_{k}-y_{k}\right)\right]  \tag{34}\\
\text { if } 1<d_{k} \leq 2
\end{array}\right.
$$

where the min function is on $0 \leq y_{k} \leq d_{k}, a_{k}=x_{k-1}=$ $d_{k}-y_{k}$ and $b_{k}=2-x_{k-1}=2-d_{k}+y_{k}$.

Substituting this values of $a_{k}$ and $b_{k}$, (34) becomes
$\Phi_{k}\left(d_{k}\right)=$

$$
\left\{\begin{array}{c}
\min \left[\frac{y_{k}^{2} \sqrt{y_{k}^{2}+6\left(d_{k}-y_{k}\right) d_{k}}}{6 \sqrt{2}}+\Phi_{k-1}\left(d_{k}-y_{k}\right)\right] \\
\text { if } 0 \leq d_{k} \leq 1, \\
\min \left[\frac{y_{k}^{2} \sqrt{y_{k}^{2}+6\left(2-d_{k}+y_{k}\right)\left(2-d_{k}\right)}}{6 \sqrt{2}}+\Phi_{k-1}\left(d_{k}-y_{k}\right)\right]  \tag{35}\\
\text { if } 1<d_{k} \leq 2,
\end{array}\right.
$$

where the min function is on $0 \leq y_{k} \leq d_{k}$.
Then solving the recursive equations (33) and (35) by executing a computer program developed for the solution procedure given in section 4 , the OSWs are obtained. The results of optimum strata widths $y_{h}^{*}$ and hence the optimum strata boundaries $x_{h}^{*}$ along with the values of the objective function $\sum_{h=1}^{L} \phi_{h}\left(y_{h}\right)$ for $L=2,3,4,5$ and 6 are presented in Table 1.

Table 1
Optimum strata widths and boundaries of triangular distribution


### 5.2 Normal distribution

Let $x$ follow the Standard Normal distribution in the interval $\left(x_{0}, x_{L}\right)$. For the purpose of illustration, we assume that $x_{0}=-4$ and $x_{L}=4$. Then $d=8$, which gives MPP (28) as:
$\operatorname{Minimize} \sum_{h=1}^{L}\left\{\operatorname{Sqrt}\left\{\frac{1}{2 \sqrt{2 \pi}} \times\right.\right.$
$\left[x_{h-1} \exp \left(-\frac{x_{h-1}^{2}}{2}\right) \operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)\right.$
$-\left(y_{h}+x_{h-1}\right) \exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)$
$-x_{h-1} \exp \left(-\frac{x_{h-1}^{2}}{2}\right) \operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)$
$\left.+\left(y_{h}+x_{h-1}\right) \exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right]$
$+\frac{1}{4}\left[\operatorname{erf}\left(\frac{y_{h}+x_{h-1}}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{x_{h-1}}{\sqrt{2}}\right)\right]^{2}$
$\left.\left.-\frac{1}{2 \pi}\left[\exp \left(-\frac{x_{h-1}^{2}}{2}\right)-\exp \left(-\frac{\left(y_{h}+x_{h-1}\right)^{2}}{2}\right)\right]^{2}\right\}\right\}$,
subject to $\quad \sum_{h=1}^{L} y_{h}=8$,
and $\quad y_{h} \geq 0 ; h=1,2, \ldots, L$.
We have

$$
\begin{aligned}
x_{k-1} & =x_{0}+y_{1}+y_{2}+\ldots+y_{k-1} \\
& =-4+y_{1}+y_{2}+\ldots+y_{k-1} \\
& =d_{k-1}-4 \\
& =d_{k}-y_{k}-4 .
\end{aligned}
$$

Substituting this value of $x_{k-1}$ in (36) and using (30) and (31), the recursive equations for solving MPP (36) are obtained as:

For first stage $(k=1)$ :

$$
\begin{align*}
\Phi_{1}\left(d_{1}\right) & =\left\{\operatorname { S q r t } \left\{\frac { 1 } { 2 \sqrt { 2 \pi } } \left[-\exp \left(-\frac{1}{2}\right) \operatorname{erf}\left(\frac{\left(d_{1}-4\right)}{\sqrt{2}}\right)\right.\right.\right. \\
& -\left(d_{1}-4\right) \exp \left(-\frac{\left(d_{1}-4\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{\left(d_{1}-4\right)}{\sqrt{2}}\right) \\
& +\exp \left(-\frac{1}{2}\right) \operatorname{erf}\left(-\frac{1}{\sqrt{2}}\right) \\
& \left.+\left(d_{1}-4\right) \exp \left(-\frac{\left(d_{1}-4\right)^{2}}{2}\right) \operatorname{erf}\left(-\frac{1}{\sqrt{2}}\right)\right] \\
& +\frac{1}{4}\left[\operatorname{erf}\left(\frac{\left(d_{1}-4\right)}{\sqrt{2}}\right)-\operatorname{erf}\left(-\frac{1}{\sqrt{2}}\right)\right]^{2} \\
& \left.\left.-\frac{1}{2 \pi}\left[\exp \left(-\frac{1}{2}\right)-\exp \left(-\frac{\left(d_{1}-4\right)^{2}}{2}\right)\right]^{2}\right\}\right\} \tag{37}
\end{align*}
$$

at $y_{1}=d_{1}$,
and for the stages $k \geq 2$ :

$$
\begin{align*}
& \Phi_{k}\left(d_{k}\right)=\min _{0 \leq y_{k} \leq d_{k}}\left\{\operatorname { S q r t } \left\{\frac{1}{2 \sqrt{2 \pi}} \times\right.\right. \\
& {\left[\left(d_{k}-y_{k}-4\right) \exp \left(-\frac{\left(d_{k}-y_{k}-4\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{d_{k}-4}{\sqrt{2}}\right)\right.} \\
& -\left(d_{k}-4\right) \exp \left(-\frac{\left(d_{k}-4\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{d_{k}-4}{\sqrt{2}}\right) \\
& -\left(d_{k}-y_{k}-4\right) \exp \left(-\frac{\left(d_{k}-y_{k}-4\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{\left(d_{k}-y_{k}-4\right)}{\sqrt{2}}\right) \\
& \left.+\left(d_{k}-4\right) \exp \left(-\frac{\left(d_{k}-4\right)^{2}}{2}\right) \operatorname{erf}\left(\frac{\left(d_{k}-y_{k}-4\right)}{\sqrt{2}}\right)\right] \\
& +\frac{1}{4}\left[\operatorname{erf}\left(\frac{d_{k}-4}{\sqrt{2}}\right)-e r f\left(\frac{\left(d_{k}-y_{k}-4\right)}{\sqrt{2}}\right)\right]^{2} \\
& \left.-\frac{1}{2 \pi}\left[\exp \left(-\frac{\left(d_{k}-y_{k}-4\right)^{2}}{2}\right)-\exp \left(-\frac{\left(d_{k}-4\right)^{2}}{2}\right)\right]^{2}\right\} \\
& \left.+\Phi \Phi_{k-1}\left(d_{k}-y_{k}\right)\right\} . \tag{38}
\end{align*}
$$

Solving the recursive equations (37) and (38), the optimum strata widths $y_{h}^{*}$ and hence the optimum strata boundaries $x_{h}^{*}$ are obtained. Table 2 shows these results along with the values of the objective function $\sum_{h=1}^{L} \phi_{h}\left(y_{h}\right)$ for $L=2,3,4,5$ and 6 .

Table 2
Optimum strata widths and boundaries of standard normal distribution

| No. of strata L | Optimum Strata Widths (OSW) $\left(y_{h}^{*}\right)$ | Optimum Strata Boundaries (OSB) $\left(x_{h}^{*}=x_{h-1}^{*}+y_{h}^{*}\right)$ | Optimum values of the objective function $\sum_{h=1}^{L} \phi_{h}\left(y_{h}\right)=\sum_{h=1}^{L} W_{h} \sigma_{h}$ |
| :---: | :---: | :---: | :---: |
| 2 | $\begin{aligned} & y_{1}^{*}=4.000000 \\ & y_{2}^{*}=4.000000 \end{aligned}$ | $x_{1}^{*}=0.000000$ | 0.6021710931 |
| 3 | $\begin{aligned} y_{1}^{*} & =3.450300 \\ y_{2}^{*} & =1.099400 \\ y_{3}^{*} & =3.450300 \end{aligned}$ | $\begin{array}{r} x_{1}^{*}=-0.549700 \\ x_{2}^{*}=0.549700 \end{array}$ | 0.4265717619 |
| 4 | $\begin{aligned} y_{1}^{*} & =3.124570 \\ y_{2}^{*} & =0.875430 \\ y_{3}^{*} & =0.875430 \\ y_{4}^{*} & =3.124570 \end{aligned}$ | $\begin{array}{r} x_{1}^{*}=-0.875430 \\ x_{2}^{*}=0.000000 \\ x_{3}^{*}=0.875430 \end{array}$ | 0.3297899642 |
| 5 | $\begin{aligned} & y_{1}^{*}=2.896360 \\ & y_{2}^{*}=0.767900 \\ & y_{3}^{*}=0.671480 \\ & y_{4}^{*}=0.767900 \\ & y_{5}^{*}=2.896360 \end{aligned}$ | $\begin{aligned} x_{1}^{*} & =-1.103640 \\ x_{2}^{*} & =-0.335740 \\ x_{3}^{*} & =0.335740 \\ x_{4}^{*} & =1.103640 \end{aligned}$ | 0.2686646379 |
| 6 | $\begin{aligned} & y_{1}^{*}=2.722440 \\ & y_{2}^{*}=0.702200 \\ & y_{3}^{*}=0.575360 \\ & y_{4}^{*}=0.575360 \\ & y_{5}^{*}=0.702200 \\ & y_{6}^{*}=2.722440 \end{aligned}$ | $\begin{aligned} x_{1}^{*} & =-1.277560 \\ x_{2}^{*} & =-0.575360 \\ x_{3}^{*} & =0.000000 \\ x_{4}^{*} & =0.575360 \\ x_{5}^{*} & =1.277560 \\ x_{6}^{*} & =4.000000 \end{aligned}$ | 0.2265979522 |

Table 3
Frequency distribution of $x$ and cum $\sqrt{f(x)}$

| Class | Frequency $\boldsymbol{f ( x )}$ | Cum $\sqrt{\boldsymbol{f ( x )}}$ |
| :---: | :---: | :---: |
| $(-3.98)-(-3.58)$ | 2 | 1.4 |
| $(-3.58)-(-3.18)$ | 6 | 3.8 |
| $(-3.18)-(-2.78)$ | 23 | 8.6 |
| $(-2.78)-(-2.38)$ | 59 | 16.3 |
| $(-2.38)-(-1.98)$ | 155 | 28.7 |
| $(-1.98)-(-1.58)$ | 296 | 45.9 |
| $(-1.58)-(-1.18)$ | 630 | 71.0 |
| $(-1.18)-(-0.783)$ | 1,015 | 102.9 |
| $(-0.783)-(-0.383)$ | 1,361 | 139.8 |
| $(-0.383)-0.017$ | 1,551 | 179.2 |
| $0.017-0.417$ | 1,495 | 217.9 |
| $0.417-0.817$ | 1,315 | 254.2 |
| $0.817-1.22$ | 1,003 | 285.9 |
| $1.22-1.62$ | 613 | 310.7 |
| $1.62-2.02$ | 285 | 327.6 |
| $2.02-2.42$ | 128 | 338.9 |
| $2.42-2.82$ | 38 | 345.1 |
| $2.82-3.22$ | 18 | 349.3 |
| $3.22-3.62$ | 7 | 351.9 |

## 6. Discussion

In this section we will undertake a numerical investigation into the effectiveness of the dynamic programming method to the Dalenius and Hodges' cum $\sqrt{f}$ method, which is the most frequently used and better known method. For this purpose, we have generated data of size $N=10,000$ for a population with standard normal density function $f(x)=(1 / \sqrt{2 \pi}) \exp \left(-x^{2} / 2\right)$, which have been grouped into 19 equal classes. In Table 3 the class frequencies are given in column 2 while their cumulative roots are given in column 3 .

For this example the smallest and the largest values of $x$ are $x_{0}=-3.98$ and $x_{L}=3.62$ respectively. Therefore, the range of the distribution $d=7.60$.

The OSB are determined for this distribution by using cum $\sqrt{f}$ method and also dynamic programming method. For each $L=2,3,4,5$ and 6 the variance $\sum_{h=1}^{L} W_{h} \sigma_{h}$ is calculated, which is used for the efficiency of the two methods of stratification. The results of this investigation are given in Table 4. From the last column of table it can be seen that the OSB obtained by dynamic programming method are more efficient for all $L=1,2, \ldots, 6$. Although, the efficiency of cum $\sqrt{f}$ method depends on the initial choice of the number of classes but there is no theory which gives the best number of classes (see Hedlin 2000).

Table 4
Relative efficiency of dynamic programming method

| $L$ | (Cum | $\sqrt{\boldsymbol{f}}$ method) | Dynamic programming <br> method | Relative <br> efficiency |  |
| :---: | ---: | :--- | ---: | :--- | :--- |
|  | OSB | $\sum_{\boldsymbol{h}=\mathbf{1}}^{\boldsymbol{L}} \boldsymbol{W}_{\boldsymbol{h}} \sigma_{\boldsymbol{h}}$ | OSB | $\sum_{\boldsymbol{h}=\mathbf{1}}^{\boldsymbol{L}} \boldsymbol{W}_{\boldsymbol{h}} \sigma_{\boldsymbol{h}}$ |  |
| 2 | -0.017 | 0.60131 | -0.00034 | 0.60126 | 100.00832 |
| 3 | -0.783 | 0.43177 | -0.55015 | 0.42576 | 101.41159 |
|  | 0.417 |  | 0.54884 |  |  |
| 4 | -0.783 | 0.33067 | -0.87593 | 0.32905 | 100.49233 |
|  | -0.017 |  | -0.00081 |  |  |
|  | 0.817 |  | 0.87395 |  |  |
| 5 | -1.18 | 0.27066 | -1.10418 | 0.26799 | 100.99631 |
|  | -3.83 |  | -0.33656 |  |  |
|  | 0.417 |  | 0.33452 |  |  |
|  | 1.22 |  | 1.10147 |  | 107.27498 |
| 6 | -1.18 | 0.24242 | -1.27813 | 0.22598 |  |
|  | -0.783 |  | -0.57619 |  |  |
|  | -0.017 |  | -0.00115 |  |  |
|  | 0.417 |  | 0.57369 |  |  |
|  | 1.22 |  | 1.27462 |  |  |

Finally, the other methods available in the literature such as Aoyama (1954), Ekman (1959), Sethi (1963), etc. are mostly classical methods to obtain approximate strata boundaries. Many authors such as Unithan (1978), Lavallée and Hidiroglou (1988), Sweet and Sigman (1995), Rivest (2002), etc. suggested iterative procedures. These iterative procedure require initial approximate solutions. Also there is no guarantee that an iterative procedure will give the global minimum in the absence of a suitable approximate initial
solution and the variance function have more than one local minimum. The advantage of the dynamic programming method is that it gives the global minimum of the objective function and it does not require any initial approximate solutions.

## Acknowledgement

The authors are grateful to the Associate Editor and referees for their invaluable comments and suggestions to improve the manuscript.

## References

Aoyama, H. (1954). A study of stratified random sampling. Annals of the Institute of Statistical Mathematics, 6, 1-36.

Bellman, R.E. (1957). Dynamic Programming. Princetown University Press, New Jersey.

Bühler, W., and Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. Metrika, 22, 161-175.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. Bulletin of the International Statistical Institute, 38, 2, 345-358.

Dalenius, T. (1950). The problem of optimum stratification-II. Skand. Aktuartidskr, 33, 203-213.

Dalenius, T., and Gurney, M. (1951). The problem of optimum stratification. Skand. Aktuartidskr, 34, 133-148.

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. Journal of the Americal Statistical Association, 54, 88-101.

Detlefsen, R.E., and Veum, C.S. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. Proceedings of the Survey Research Methods Section, American Statistical Association, 214-219.

Ekman, G. (1959). Approximate expression for conditional mean and variance over small intervals of a continuous distribution. Annals of the Institute of Statistical Mathematics, 30, 1131-1134.

Hansen, M.H., and Hurwitz, W.N. (1953). On the theory of sampling from finite population. The Annals of Mathematical Statistics, 14, 333-362.

Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. Journal of Official Statistics, 16, 15-29.

Hess, I., Sethi, V.K. and Balakrishnan, T.R. (1966). Stratification: A practical investigation. Journal of the Americal Statistical Association, 61, 71-90.
Hidiroglou, M.A., and Srinath, K.P. (1993). Problems associated with designing subannual bussiness surveys. Journal of Bussiness and Economic Statistics, 11, 397-405.

Khan, E.A., Khan, M.G.M. and Ahsan, M.J. (2002). Optimum stratification: A mathematical programming approach. Culcutta Statistical Association Bulletin, 52 (special), 205-208.

Khan, M.G.M., Najmussehar, and Ahsan, M.J. (2005). Optimum stratification for exponential study variable under Neyman allocation. Journal of Indian Society of Agricultural Statistics, 59(2), 146-150.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. Statistics in Transition, 6, 5, 797-806.
Lavallée, P. (1987). Some contributions to optimal stratification. Master Thesis, Carleton University, Ottawa, Canada.
Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. Proceedings of the Section on Survey Research Methods, American Statistical Association, Virginia, 646-651.

Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. Survey Methodology, 14, 33-43.

Lednicki, B., and Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. Statistics in Transition, 6, 287-306.

Mahalanobis, P.C. (1952). Some aspects of the design of sample surveys. Sankhyā, 12, 1-7.

Murthy, M.N. (1967). Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. Computer Journal, 7, 308-313.

Nicolini, G. (2001). A method to define strata boundaries. Departmental Working Papers 2001-01, Department of Economics, University of Milan, Italy (www.economia.unimi.it/ pubb/wp83.pdf).
Niemiro, W. (1999). Konstrukcja optymalnej stratyfikacja metoda poszukiwan losowych. (Optimal Stratification using Random Search Method). Wiadomosci Statystyczne, 10, 1-9.

Rivest, L.-P. (2002). A Generalization of Lavallée and Hidiroglou algorithm for stratification in business survey. Survey Methodology, 28, 191-198.

Sethi, V.K. (1963). A note on optimum stratification of population for estimating the population mean. Australian Journal of Statistics, 5, 20-33.

Sweet, E.M., and Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. Proceedings of the Survey Research Methods Section, American Statistical Association, 491-496.

Unnithan, V.K.G. (1978). The minimum variance boundry points of stratification. Sankhyā, 40, C, 60-72.

Unnithan, V.K.G., and Nair, N.U. (1995). Minimum variance stratification. Communications in Statistics, 24(1), 275-284.

# Multi-objective optimisation for optimum allocation in multivariate stratified sampling 

José A. Díaz-García and Liliana Ulloa Cortez ${ }^{1}$


#### Abstract

This paper considers the optimum allocation in multivariate stratified sampling as a nonlinear matrix optimisation of integers. As a particular case, a nonlinear problem of the multi-objective optimisation of integers is studied. A full detailed example including some of proposed techniques is provided at the end of the work.


Key Words: Multivariate stratified sampling; Optimum allocation; Multi-objective optimisation.

## 1. Introduction

One of the areas of statistics that is most commonly used in all fields of scientific investigation is that of probabilistic sampling. An effective sampling technique within a population represents an appropriate extraction of useful data which provides meaningful knowledge of the important aspects of the population. Stratified sampling is one of the classical methods for obtaining such information. This method considers the computation of the stratum sample size, which can be computed by various procedures, but optimum allocation has been found to be a useful approach. Optimum allocation is considered as a non-linear optimisation problem in which the objective function is the variance subject to a cost restriction, or vice versa. Traditionally, this problem has been solved by using the Cauchy-Schwarz (Stuart 1954) inequality, cited in Cochran (1977) or Lagrange's multiplier method, see Sukhatme, Sukhatme, Sukhatme and Asok (1984).

Classical sampling theory considers a single decision variable or parameter; for example, in our case, univariate stratified sampling studies one parameter, the sample size and its strata allocation, see Cochran (1977), Sukhatme et al. (1984) and Thompson (1997). Moreover, in the context of stratified sampling, some multivariate approaches have been proposed whereby the sample size and its allocation within strata take diverse characteristics into consideration, see Sukhatme et al. (1984) and Arthanari and Dodge (1981), among others.

When the optimum allocation is performed, and the cost function is the objective function, subject to certain variance restrictions in the different characteristics, then the problem can be reduced to a question of classical mathematical programming, and for this purpose there are two wellknown approaches: Arthanari and Dodge (1981), from a
deterministic point of view; and Prékopa (1978), from a stochastic position. In the latter case, the problem can be solved by using any of the techniques presented in DíazGarcía and Garay (2007).

Alternatively, if we wish to minimise the variances subject to a cost function, or to a given sample size, then sereral approaches can be adopted to solve this, see Sukhatme et al. (1984). However, the above-mentioned approaches do not solve the over-sampling problem, i.e., when the sample size in one or more strata is larger than the stratum size; furthermore, the sample sizes obtained are not integers, and must be approximated. Moreover, as we shall see, all the previously published approaches in this area are particular cases of the multi-objective optimisation technique. If these problems could be overcome, then we would have a formal overview and a unified theory for resolving the problem of optimum allocation in multivariate stratified sampling, and would be able to consider all the literature (both theory and practice) on multi-objective optimisation and related questions.

In this paper we study optimum allocation in multivariate stratified sampling as a nonlinear problem of matrix optimisation of integers constrained by a cost function or by a given sample size. Making certain assumptions, we propose a way to solve the problem, through several particular techniques, see subsection 3.1. The second aim of the paper is related to the following fact: if we define a particular vectorial function of the objective function of the matrix optimisation problem, then in subsection 3.2 we show that the optimum allocation in multivariate stratified sampling also can be studied as a non-linear problem of the multiobjective optimisation of integers. In subsections 3.2.1 and 3.2.2 we propose different techniques for solving these problems. Finally, in section 4, some of the techniques described are applied to a numerical example from forestry.

[^9]
## 2. Multivariate stratified sampling

Consider a population of size $N$, divided into $H$ subpopulations (strata). We wish to find a representative sample of size $n$ and an optimum allocation in the strata meeting the following requirements: i) to minimise the variance of the estimated mean subject, to a budgetary constraint; or ii) to minimise the cost subject to a constraint on the variances; this is the classical problem in optimum allocation in univariate stratified sampling, see Cochran (1977), Sukhatme et al. (1984) and Thompson (1997). However, if we consider more than one characteristic (variable) then the problem is known as optimum allocation in multivariate stratified sampling. For a formal expression of the problem of optimum allocation in stratified sampling, consider the following notation.

### 2.1 Notation

The subindex $h=1,2, \ldots, H$ denotes the stratum, $i=1,2, \ldots, N_{h}$ the unit within stratum $h$ and $j=1,2, \ldots, G$ denotes the characteristic (variable). Moreover:

| $N_{h}$ | Total number of units within <br> stratum $h$ |
| :--- | :--- |
| $n_{h}$ | Number of units from the sample <br> in stratum $h$ |
| $y_{h i}^{j}$ | Value obtained for the $i^{\text {th }}$ unit in <br> stratum $h$ of the $j^{\text {th }}$ characteristic <br> Vector of the number of units in <br> the sample |
| $W_{h}=\frac{N_{h}}{N}$ | Relative size of stratum $h$ |
| $\bar{Y}_{h}^{j}=\frac{\sum_{i=1}^{N_{h}} y_{h i}^{j}}{N_{h}} \quad$Population mean in stratum $h$ of <br> the $j^{\text {th }}$ characteristic |  |
| $\bar{y}_{h}^{j}=\frac{\sum_{i=1}^{n_{h}} y_{h i}^{j}}{n_{h}} \quad$Sample mean in stratum $h$ of the <br> $j^{\text {th }}$ characteristic |  |
| $\overline{\mathbf{y}}_{h}=\left(\bar{y}_{h}^{1}, \ldots, \bar{y}_{h}^{G}\right)^{\prime}$ | Sample mean vector in stratum $h$ |
| $\bar{y}_{\mathrm{ST}}^{j}=\sum_{h=1}^{H} W_{h} \bar{y}_{h}^{j}$ | Estimator of the population mean <br> in multivariate stratified sampling <br> for the $j^{\text {th }}$ characteristic |
| Estimator of the population mean |  |
| vector in multivariate stratified |  |
| sampling |  |

Sample covariance in stratum $h$ of

$$
\begin{aligned}
& s_{h_{j k}} \\
& \text { the } j^{\text {th }} \text { and } k^{\text {th }} \text { characteristics, } \\
& s_{h_{j k}}=\frac{\sum_{i=1}^{n_{h}}\left(y_{h i}^{j}-\bar{y}_{h}^{j}\right)\left(y_{h i}^{k}-\bar{y}_{h}^{k}\right)}{n_{h}-1} \text {, and } \\
& s_{h_{j j}}=s_{h j}^{2} \frac{\sum_{i=1}^{n_{h}}\left(y_{h i}^{j}-\bar{y}_{h}^{j}\right)^{2}}{n_{h}-1}
\end{aligned}
$$

$\operatorname{Cov}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \quad$ Variance-covariance matrix of $\overline{\mathbf{y}}_{\mathrm{ST}}$

$$
\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \quad \begin{aligned}
& \text { Estimator of the variance-covari- } \\
& \text { ance matrix of } \overline{\mathbf{y}}_{\mathrm{ST}}, \text { where } \\
& \widehat{\operatorname{Cov}( }\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \equiv \widehat{\operatorname{Cov}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \text { and }}
\end{aligned}
$$

$$
\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)=\left(\begin{array}{cccc}
\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{1}\right) & \widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{1}, \bar{y}_{\mathrm{ST}}^{2}\right) & \cdots & \widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{1}, \bar{y}_{\mathrm{ST}}^{G}\right) \\
\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{2}, \bar{y}_{\mathrm{ST}}^{1}\right) & \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{2}\right) & \cdots & \widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{2}, \bar{y}_{\mathrm{ST}}^{G}\right) \\
\vdots & \vdots & \ddots & \vdots \\
\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{G}, \bar{y}_{\mathrm{ST}}^{1}\right) & \widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{G}, \bar{y}_{\mathrm{ST}}^{2}\right) & \cdots & \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{G}\right)
\end{array}\right)
$$

$$
\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{j}, \bar{y}_{\mathrm{ST}}^{k}\right) \quad \begin{aligned}
& \text { Estimated covariance of } \bar{y}_{\mathrm{ST}}^{j} \text { and } \\
& \bar{y}_{\mathrm{ST}}^{k} \text { where }
\end{aligned}
$$

$$
\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{i}, \bar{y}_{\mathrm{ST}}^{j}\right) \equiv \widehat{\operatorname{Cov}\left(\bar{y}_{\mathrm{ST}}^{j}, \bar{y}_{\mathrm{ST}}^{k}\right)}, \quad \text { with }
$$

$$
\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{j}, \bar{y}_{\mathrm{ST}}^{k}\right)=\sum_{h=1}^{H} \frac{W_{h}^{2} s_{h_{j k}}}{n_{h}}-\sum_{h=1}^{H} \frac{W_{h} s_{h_{j k}}}{N}, \quad \text { and }
$$

$$
\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{j}, \bar{y}_{\mathrm{ST}}^{j}\right) \equiv \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right)=\sum_{h=1}^{H} \frac{W_{h}^{2} s_{h j}^{2}}{n_{h}}-\sum_{h=1}^{H} \frac{W_{h} s_{h j}^{2}}{N}
$$

$c_{h} \quad$ Cost per sampling unit in stratum $h$

Finally, let $\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \in \mathfrak{R}^{G}$, such that $\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)=\left(\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{1}\right)\right.$, $\left.\ldots, \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{G}\right)\right)^{\prime}$, where if $\mathbf{a} \in \mathfrak{R}^{G}, \mathbf{a}^{\prime}$ denotes the transpose of $\mathbf{a}$.

## 3. A new approach for the problem of optimum allocation in multivariate stratified sampling

In this section we propose optimum allocation in multivariate stratified sampling as a matrix optimisation problem, for which a number of possible solutions are
studied. We observe that the multi-objective optimisation problem is a particular case of a matrix optimisation. In the same sense, we note that optimum allocation in multivariate stratified sampling can be seen as a multi-objective optimisation problem. In each case, the respective solutions are straightforwardly derived.

### 3.1 Matrix optimisation

Formally, optimum allocation in stratified sampling can be studied by performing the following nonlinear matrix optimisation problem:

$$
\begin{align*}
& \min _{\mathbf{n}} \widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \\
& \text { subject to }  \tag{1}\\
& \mathbf{c}^{\prime} \mathbf{n}+c_{0}=C
\end{align*}
$$

where $C$ is the total cost, $c_{0}$ is a fixed cost and $\mathbf{c}^{\prime}=\left(c_{1}, \ldots, c_{H}\right)$.

Note that the solutions proposed for problem (1) take real values, and thus the sample sizes $n_{h}$ must be integers. We must also address the problem of over-sampling, that is, when $n_{h} \geq N_{h}$ for at least some $h$, see Arvanitis and Afonja (1971). In order to overcome these two complications, we propose the following alternative approach to (1).

$$
\begin{align*}
& \min _{\mathbf{n}} \widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \\
& \text { subject } \text { to } \\
& \mathbf{c}^{\prime} \mathbf{n}+c_{0}=C  \tag{2}\\
& 2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H \\
& n_{n} \in \mathbb{N}
\end{align*}
$$

where $\mathbb{N}$ denotes the set of natural numbers.
Obviously, the difficulty of expressing the problem in this way lies in defining the meaning of the minimum of a matrix function. The idea of minimising a matrix function, and in particular the matrix of variance-covariance, has been studied with respect to various areas of statistical theory. For example, when the regression estimators are determined for a multivariate general linear model, this is done by minimising the determinant or the trace of sums of squares and sums of products matrix of the erro, see Giri (1977). Similarly, the choice or comparison of some experimental design models is done by minimising a function of the variance-covariance matrix of treatment estimators, see Khuri and Cornell (1987) and Azaïs and Druilhet (1997).

Fortunately, it is possible to reduce the nonlinear matrix minimisation problem (2) to a univariate nonlinear minimisation problem by taking into account the following considerations (note that the prodecure described here is just
one of various possible options, see Ríos, Ríos Insua and Ríos Insua (1989) and Miettinen (1999)). Observe that $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)$ is an explicit function of $\mathbf{n}$, and so it must be denoted as as $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \equiv \widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$. Also, assume that $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$ is a positive definite matrix for all $\mathbf{n}$, $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)>\mathbf{0}$. Now, let $\mathbf{n}_{1}$ and $\mathbf{n}_{2}$ be two possible values of the vector $\mathbf{n}$ and let $\mathbf{B}=\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)-$ $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right)$. We say that

$$
\begin{equation*}
\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)<\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right) \Leftrightarrow \mathbf{B}<\mathbf{0}, \tag{3}
\end{equation*}
$$

i.e., if the matrix $\mathbf{B}$ is a negative definite matrix. Moreover, note that $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)$ and $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right)$, are diagonalizable. Then, let $D_{\mathbf{n}_{1}}$ and $D_{\mathbf{n}_{2}}$ be the diagonal matrixes associated with $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)$ and $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right)$, respectively, with $D_{\mathbf{n}_{1}}=\operatorname{diag}\left(\alpha_{1}, \ldots, \alpha_{G}\right), \alpha_{1}>\cdots>\alpha_{G}>0$ and $D_{\mathbf{n}_{2}}=\operatorname{diag}\left(\tau_{1}, \ldots, \tau_{G}\right), \tau_{1}>\cdots>\tau_{G}>0$, where $\alpha_{j}$ and $\tau_{j}$ denote the eigenvalues of $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)$ and $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right)$, respectively. Thus, expression (3) can alternatively be presented as:

$$
\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)<\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right) \Leftrightarrow D_{\mathbf{n}_{1}}-D_{\mathbf{n}_{2}}<\mathbf{0}
$$

i.e.,

$$
\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)<\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right) \Leftrightarrow \underset{\substack{j=1, \ldots, G \\ j=1}}{\alpha_{j}-\tau_{j}}<0
$$

and

$$
\begin{equation*}
\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right) \neq \widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right) \tag{4}
\end{equation*}
$$

which defines a weak Pareto order, see Steuer (1986), Ríos et al. (1989) and Miettinen (1999). Then from Steuer (1986), Ríos et al. (1989) and Miettinen (1999), there exist a function $\mathbf{f}: S \rightarrow \mathfrak{R}$, such that

$$
\begin{align*}
\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right) & <\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right) \\
& \Leftrightarrow \mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)\right)<\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{2}\right)\right)\right) . \tag{5}
\end{align*}
$$

where $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right) \in S \subset \mathfrak{R}^{G(G+1) / 2}$ and $S$ is the set of positive definite matrixes. From (5), Steuer (1986), Ríos et al. (1989) and Miettinen (1999) proof that the non-linear matrix minimisation problem (2) is reduced in the following univariate non-linear minimisation problem
$\min _{\mathbf{n}} \mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)\right)$
subject to
$\sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C$
$2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H$
$n_{h} \in \mathbb{N} ;$

Unfortunately or fortunately the function $\mathbf{f}(\cdot)$ is not unique. For example, in other statistical contexts we see the following commonly used functions $\mathbf{f}(\cdot)$, see Giri (1977):

1. The trace of the matrix $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$; $\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)=\operatorname{tr}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)$.
2. The determinant of the matrix $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$; $\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)=\left|\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right|$.
3. The sum of all the elements of the matrix $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$;

$$
\widehat{\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)=\sum_{j, k=1}^{G} \widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{j}, \bar{y}_{\mathrm{ST}}^{k}\right) . . . . . . . .}
$$

4. $\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)=\lambda_{\text {max }}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)$, where $\lambda_{\text {max }}$ is the maximum eigenvalue of the matrix $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$.
5. $\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)=\lambda_{\text {min }}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)$, where $\lambda_{\text {min }}$ is the minimum eigenvalue of the matrix $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$.
6. $\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)=\lambda_{j}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)$, where $\lambda_{j}$ is the $j^{\text {th }}$ eigenvalue of the matrix $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$, among others.

In particular Dalenius (1957), studied the problem (6) when $\mathbf{f}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right)=\left|\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right|$, in other words, the minimisation of the generalised variance $\left|\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)\right|$, see also Arvanitis and Afonja (1971).

### 3.2 Multi-objectibve optimisation

Let us now, consider the vectorial function $\mathbf{F}: S \rightarrow \mathfrak{R}^{G}$, such that $\mathbf{F}\left(\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)\right)=\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)$. An alternative way of establishing problem (2) is

$$
\begin{align*}
& \min _{\mathbf{n}} \mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)=\min _{\mathbf{n}}\left(\begin{array}{c}
\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{1}\right) \\
\vdots \\
\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{G}\right)
\end{array}\right) \\
& \text { subject to } \\
& \mathbf{c}^{\prime} \mathbf{n}+c_{0}=C  \tag{7}\\
& 2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H \\
& n_{h} \in \mathbb{N},
\end{align*}
$$

which is a nonlinear problem of the multi-objective optimisation of integers, see Steuer (1986), Ríos et al. (1989) and Miettinen (1999).

In the sampling context, observe that in multi-objective optimisation problems, there rarely exists a point $\mathbf{n}^{*}$ which is considered as an optimum, i.e., few cases satisfy the requirement that $\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\left(\mathbf{n}^{*}\right)\right)$ is minimum for all $j=1, \ldots, G$. This justifies the following notion of the

Pareto point (which is more weakly defined than an optimum point):

$$
\begin{aligned}
& \text { We say that } \mathbf{V}_{u}^{*}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \text { is a Pareto point of } \\
& \mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \text {, if there is no other point } \mathbf{V}_{u}^{1}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \\
& \text { such that } \mathbf{V}_{u}^{1}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \leq \mathbf{V}_{u}^{*}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \text {, i.e., for all } j \text {, } \\
& \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j_{1}}\right) \leq \operatorname{Var}\left(\bar{y}_{\mathrm{ST}}^{j_{*}}\right) \quad \text { and } \quad \mathbf{V}_{u}^{1}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \neq \\
& \mathbf{V}_{u}^{*}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \text {. }
\end{aligned}
$$

Existence criteria for Pareto points of a multi-objective optimisation problem are established in Ríos et al. (1989) and Miettinen (1999). In particular we have:

> Given $\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right): \mathfrak{R}^{H} \rightarrow \mathfrak{R}^{G}$ and let us consider a non empty compact $\mathcal{N} \subset \mathbb{N}^{H}$ such that $\mathcal{N}$ is the set of all possible values of $\mathbf{n}$ determined by the restrictions in $(7)$. If $\widehat{\operatorname{Var}(~}\left(\bar{y}_{\mathrm{ST}}^{j}\right)$ is an upper semicontinuous for each $j=1, \ldots, G$, then the problem $(7)$ has a Pareto optimal solution.

On the other hand, Steuer (1986), Ríos et al. (1989) and Miettinen (1999) studied the extension of scalar optimisation (Kuhn-Tucker's conditions) to the vectorial case. In particular, they proposed necessary conditions for Pareto solutions, which become sufficient conditions if: $N$ is convex; the functions $\widehat{\operatorname{Var}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}^{j}\right), j=1, \ldots, G$ are convex; and the Lagrange generalised multipliers $\delta_{j}$, associated with each function $\widehat{\operatorname{Var}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}^{j}\right)$, are positive, $\delta_{j}>0$ for all $j$. Note that the above results for the existence of a Pareto solution and for Kuhn-Tucker's conditions are valid when the optimisation problem is continuous, i.e., when the variables $n_{h}$ are continuous ones, for all $h=1, \ldots, H$. However, it should be recalled that in order to obtain the solution to an integer optimisation problem, it is normal to make the initial assumption that such a problem is one of continuous optimisation. First we derive the solution to the problem of continuous optimisation, and then, by means of heuristic or branch-and-bound methods, progress to solving that of integer optimisation. In this context, in the case of optimising integers, in a practical case, it is sufficient to see that the corresponding problem of continuous optimisation has an optimum Pareto solution, and to confirm that the set $\mathcal{N}$ of all the possible values of $\mathbf{n}$ contains at least one $\mathbf{n} \in \mathcal{N}$ for which all the coordinates are integers.

Methods for solving a multi-objective optimisation problem are based on the information possessed about a particular problem. There are three possible scenarios; when the investigator possesses: complete, partial or null information, see Ríos et al. (1989), Miettinen (1999) and Steuer (1986) In a stratified sampling context, complete information means that, the investigator knows the population in such a way that it is possible to propose a value function (Value function: This is a function $\phi: \mathfrak{R}^{H} \rightarrow \mathfrak{R}$
such that denoting $\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right) \equiv \mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})\right)$ we have that $\min \mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}^{*}\right)\right)<\min \mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right) \Leftrightarrow \phi\left(\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}^{*}\right)\right)\right)<$ $\left.\phi\left(\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\left(\mathbf{n}_{1}\right)\right)\right), \mathbf{n}^{*} \neq \mathbf{n}_{1}.\right)$ reflecting the importance of each variance of the studied characteristics, this possibility, today, is very rarely encountered. In partial information, the investigator knows the main characteristic of the study very well and this is sufficient support for the research. Finally, under null information, which is the most common situation, the researcher only possesses information about the estimators of the parameter of the experiment, and with this material an appropriate solution can be found.

For reasons of space it is impossible to give an exhaustive explanation of all the techniques proposed for solving multi-objective optimisation problems (7), see Ríos et al. (1989), Miettinen (1999) and Steuer (1986) for a detailed description. Moreover, there are heuristic methods, instead of the classical methods, by which the problem may be addressed in an alternative way, see Jones et al. (2002). As an illustration, we present below a survey of two commonly used techniques; the first one studies the complete information stage (the value function, also termed the utility function) and the second one, the null information scenario (a method based on distances).

### 3.2.1 Value function

As mentioned above, this method belongs to the complete information case, in which the investigator is able to summarise the importance of all the studied characteristics in a real function, see the next paragraph (see also Ríos et al. (1989), Miettinen (1999) and Steuer (1986), among others).

Under the value function technique, problem (7) is expressed as follows:

$$
\begin{align*}
& \min _{\mathbf{n}} \phi\left(\mathbf{V}_{u}\left(\bar{y}_{\mathrm{ST}}\right)\right), \\
& \text { subject to } \\
& \sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C  \tag{8}\\
& 2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H \\
& n_{h} \in \mathbb{N},
\end{align*}
$$

where $\phi(\cdot)$ is a scalar function that summarises the importance of each of the variances of the $G$ characteristics.

Clearly, many of the approaches described in the literature on the question of optimum allocation in multivariate stratified sampling, such as compromise assignation, compromise assignation minimising total relative loss, and compromise assignation taking the mean of the optimum
values, see Sakhatme et al. (1984), are particular cases of the above-mentioned method.

Note that the value function $\phi(\cdot)$ may take an infinite number of forms, which represents a fundamental obstacle to defining it. However, some simple functions have given excellent results in the applications and they can be considered as promising approaches. One of these particular forms is the weighting method. Under this approach, problem (8) can be expressed as:

$$
\begin{aligned}
& \min _{\mathbf{n}} \sum_{j=1}^{G} w_{j} \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right), \\
& \text { subject to } \\
& \sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C \\
& 2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H \\
& n_{h} \in \mathbb{N},
\end{aligned}
$$

such that $\sum_{j=1}^{G} w_{j}=1, w_{j} \geq 0 \forall j=1,2, \ldots, G$; where $w_{j}$ weights the importance of each characteristic.

Among the multi-objective techniques we find that the value function method is, in general, the most commonly applied, because its properties have been studied with most detail, see Ríos et al. (1989), Miettinen (1999), Steuer (1986), and the references therein.

### 3.2.2 Distance-based method

Sometimes, the researcher does not have sufficient previous information about the variables, or it is difficult to decide which are the most important characteristics of the experiment. In such cases, the method of this section is very useful, because it does not need many antecedents; moreover, it only requires a vector of ideal goals, which is determined with the null information expressed in the problem, see Ríos et al. (1989) and Steuer (1986).

Then, problem (7) is solved by obtaining the optimum values via the minimisation of the distance between the optimum and the vector of targets.

Let $\theta_{j}$ be the ideal point or goal for the objective $\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right), j=1, \ldots, G$, i.e., the vector of targets $\Theta$ is given by

$$
\boldsymbol{\Theta}=\left(\begin{array}{c}
\theta_{1} \\
\vdots \\
\theta_{G}
\end{array}\right)
$$

Note that the vector of targets $\Theta$ can be calculated without additional information, which is a great advantage of this method. In fact, it is computed by minimising each
objective $\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right), j=1, \ldots, G$ separately, such that the vector $\Theta$ is defined as the vector of its individual minima, and this is achieved by solving the following $G$ non-linear minimisation problems of integers, see Rao (1979):

$$
\begin{aligned}
& \min _{\mathbf{n}} \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right), \\
& \text { subject to } \\
& \sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C \\
& 2 \leq n_{h} \leq N_{n} \\
& h=1,2, \ldots, H \\
& n_{h} \in \mathbb{N},
\end{aligned}
$$

for $j=1, \ldots, G$.
Once the vector $\Theta$ has been computed, we study the optimisation problem with the new objective function, namely

$$
\begin{align*}
& \underset{\mathbf{n}}{\min _{n}} d\left(\mathbf{V}_{u}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right), \boldsymbol{\Theta}\right) \\
& \text { subject to } \\
& \sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C  \tag{9}\\
& 2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H \\
& n_{h} \in \mathbb{N},
\end{align*}
$$

where $d(\cdot, \cdot)$ denotes a general distance function. In particular, when the program (9) is applied to the Euclidean distance, we have

$$
\begin{aligned}
& \min _{\mathbf{n}} \sum_{j=1}^{G}\left[\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right)-\theta_{j}\right]^{2} \\
& \text { subject to } \\
& \sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C \\
& 2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H \\
& n_{h} \in \mathbb{N} .
\end{aligned}
$$

Alternatively, another distance has been proposed by Khuri and Cornell (1987):

$$
\min _{\mathrm{n}} \sum_{j=1}^{G} \frac{\left(\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right)-\theta_{j}\right)^{2}}{\theta_{j}^{2}}
$$

subject to

$$
\begin{aligned}
& \sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C \\
& 2 \leq n_{h} \leq N_{h}, h=1,2, \ldots, H \\
& n_{h} \in \mathbb{N} .
\end{aligned}
$$

## Remarks:

1. Note that we have used the cost restriction $\sum_{h=1}^{H} c_{h} n_{h}+c_{0}=C$ in every optimisation method. However, in some situations, we do not restrict the costs but we have restrictions for the availability of man-hours for carrying out a survey, or restrictions on the total available time for performing the survey, etc. These limitations can be described by using the following expression, see Arthanari and Dodge (1981):

$$
\sum_{h=1}^{H} n_{h}=n .
$$

2. Note that the multi-objective optimisation methods proposed here are general and they need to be adjusted in some particular problems; for instance, we do not consider the unit (magnitude) of each variance in the respective sums for the value function. We suggest a solution, namely to replace the variance of each characteristic by its corresponding coefficient of variation

$$
\sqrt{\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{j}\right)} / \bar{y}_{\mathrm{ST}}^{j}, j=1 \ldots, G ;
$$

then, the use of Khuri and Cornell's distance is more recommendable than is the use of the Euclidean distance.
3. It is desirable to consider estimators other than a mean estimator, for example the national mean estimator, or the comparison of regional means, etc. In particular, the associated editor recommended estimators of the following type:

$$
\begin{equation*}
\widehat{\mathbf{T}}=\sum_{h=1}^{H} w_{h} \overline{\mathbf{y}}_{h} \in \mathfrak{R}^{G} \tag{10}
\end{equation*}
$$

where several weights $w_{h}$ could even be used for the same variable. For instance, if one of the weights $w_{h}$
is 1 , another is -1 , and the others are 0 , then we can compute the difference between two means of two different strata. In general, we can optimise problem (2) substituting the objective function $\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)$, by any function of interest. For example, we could use the estimated variance-covariance matrix $\widehat{\operatorname{Cov}}(\hat{\mathbf{T}})$ of the estimator (10), among many other options.

## 4. A numerical example

The input information was taken from Arvanitis and Afonja (1971) which is a forest survey conducted in Humbolt County, California. The population was subdivided into nine strata on the basis of the timber volume per unit area, as determined from aerial photographs. The two variables included in this example are the basal area (BA) (In forestry terminology, 'Basal area' is the area of a plant perpendicular to the longitudinal axis of a tree at 4.5 feet above ground) in square feet, and the net volume in cubic feet (Vol.), both expressed on a per acre basis. The variances, covariances and the number of units within stratum $h$ are listed in Table 1.

Table 1
Variances, covariances and the number of units within each stratum

|  | Variance |  |  |  |
| :---: | ---: | ---: | ---: | ---: |
| Stratum | $\boldsymbol{N}_{\boldsymbol{h}}$ | BA | Vol. | Covariance |
| 1 | 11,131 | 1,557 | 554,830 | 28,980 |
| 2 | 65,857 | 3,575 | $1,430,600$ | 61,591 |
| 3 | 106,936 | 3,163 | $1,997,100$ | 72,369 |
| 4 | 72,872 | 6,095 | $5,587,900$ | 166,120 |
| 5 | 78,260 | 10,470 | $10,603,000$ | 293,960 |
| 6 | 51,401 | 8,406 | $15,828,000$ | 357,300 |
| 7 | 24,050 | 20,115 | $26,643,000$ | 663,300 |
| 8 | 46,113 | 9,718 | $13,603,000$ | 346,810 |
| 9 | 102,985 | 2,478 | $1,061,800$ | 39,872 |

For this example, the matrix optimisation problem under approach (2) is

$$
\min _{\mathbf{n}}\left(\begin{array}{cc}
\widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{1}\right) & \widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{1}, \bar{y}_{\mathrm{ST}}^{2}\right) \\
\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}^{2}, \bar{y}_{\mathrm{ST}}^{1}\right) & \widehat{\operatorname{Var}}\left(\bar{y}_{\mathrm{ST}}^{2}\right)
\end{array}\right)
$$

subject to

$$
\begin{equation*}
\sum_{h=1}^{9} n_{h}=1,000 \tag{11}
\end{equation*}
$$

$$
\begin{aligned}
& 2 \leq n_{h} \leq N_{h}, h=1, \ldots, 9 \\
& n_{h} \in \mathbb{N} .
\end{aligned}
$$

Table 2 shows the optimisation solutions obtained by some of the methods described in Sections 2 and 3; specifically, we present the solutions via the trace, the determinant, the value function, the Euclidean distance and the Khuri and Cornell distance. We also include the optimum allocation for each characteristic, BA and Vol. (the first two rows in Table 2). The last two columns show the minimum values of the individual variances for the respective optimum allocations identified by each method. The results were computed using the commercial software Hyper LINGO/PC, release 6.0, see Winston (1995). The default optimisation methods used by LINGO to solve the nonlinear integer optimisation programs are Generalised Reduced Gradient (GRG) and branch-and-bound methods, see Bazaraa et al. (2006). Some technical details of the computations are the following: the maximum number of iterations of the methods presented in Table 2 was 1,193 (determinant problem) and the mean execution time for all the programs was 1 second. Finally, note that the greatest discrepancy found by the different methods among the sizes of the strata occurred when minimising the generalised variance $\left|\widehat{\operatorname{Cov}}\left(\overline{\mathbf{y}}_{\mathrm{ST}}\right)\right|$. Beyond doubt, this is because it is the only method presented in Table 2 that takes into account the covariance between the two characteristics studied.

Table 2
Sample sizes and estimator of variances for the different allocations calculated

| Allocation | $n_{1}$ | $\boldsymbol{n}_{2}$ | $n_{3}$ | $n_{4}$ | $n_{5}$ | $n_{6}$ | $\boldsymbol{n}_{7}$ | $n_{8}$ | $n_{9}$ | $\widehat{\operatorname{Var}}\left(\bar{y}_{\text {ST }}^{1}\right)$ | $\widehat{\operatorname{Var}}\left(\bar{y}_{\text {ST }}^{2}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| BA | 10 | 94 | 144 | 136 | 191 | 113 | 81 | 109 | 122 | 5.591 | 5,441.105 |
| Vol. | 7 | 62 | 119 | 136 | 200 | 161 | 98 | 134 | 83 | 5.953 | 5,139.531 |
| $\operatorname{tr}\left(\widehat{\operatorname{Cov}}\left(\bar{y}_{\mathrm{ST}}\right)\right)$ | 7 | 62 | 119 | 135 | 200 | 161 | 98 | 134 | 84 | 5.591 | 5,139.531 |
| $\left\|\widehat{\operatorname{Cov}}\left(\bar{y}_{\text {ST }}\right)\right\|$ | 9 | 93 | 128 | 129 | 193 | 123 | 86 | 106 | 133 | 5.616 | 5,403.876 |
| Value Function ${ }^{a}$ | 7 | 62 | 119 | 135 | 200 | 161 | 98 | 134 | 84 | 5.591 | 5,139.531 |
| $d_{E}{ }^{b}$ | 7 | 62 | 119 | 136 | 200 | 160 | 98 | 134 | 84 | 5.944 | 5,139.557 |
| $d_{K C}{ }^{\text {c }}$ | 10 | 86 | 137 | 135 | 192 | 126 | 86 | 115 | 113 | 5.613 | 5,308.11 |
| $w_{1}=w_{2}=0.5$ <br> Euclidean dist Khuri and Cor | dista |  |  |  |  |  |  |  |  |  |  |

## 5. Conclusions

It is difficult to suggest rules for choosing a method in matrix optimisation (2) when there are important numerical differences between two of them. For example, Table 2 shows opposing results in the optimum allocations and the minimum variances for the trace and the determinant techniques. A similar situation occurs in the criterion selection for testing hypotheses in the MANOVA problem, see Giri (1977). In fact, the existence of general criteria based on power tests is not sufficient for an objective decision to be made and the final choice depends on the skill of the investigator.

However, when the problem of optimum allocation in multivariate stratified sampling is considered as a nonlinear problem of the multi-objective optimisation of integers, we can give some general suggestions to reduce the number of appropriate methods in accordance with each situation. First we need to recognise the research context of the problem (i.e., total information, partial information or null information). Then, we can decide the technique according to the available information. It is important to note that the solution for an allocation problem should be achieved by the implementation of a single method. For this reason, the results obtained for any example are comparable only within the context in which the example was established.

## Acknowledgement

This research work was partially supported by CONACYT-Mexico, Research Grant No. 45974-F, 81512 and IDI-Spain, Grant No. MTM2005-09209. The authors also wish to thank the referees and Editor who handled this article for their very helpful suggestions.

## References

Arthanari, T.S., and Dodge, Y. (1981). Mathematical Programming in Statistics. A Wiley-Interscience Publication John Wiley \& Sons, Inc.

Arvanitis, L.G., and Afonja, B. (1971). Use of the generalized variance and the gradient projection method in multivariate stratified sampling. Biometrics, 27, 119-127.
Azaïs, J.-M., and Druilhet, P. (1997). Optimality of neighbour balanced designs when neighbour effects are neglected. Journal of Statistical Planning and Inference, 64, 353-367.

Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. (2006). Nonlinear Programming: Theory and Algorithms, $3^{\text {rd }}$ Edition. WileyInterscience, 2006.
Cochran, W.G. (1977). Técnicas de muestreo. C.E.C.S.A. México (in Spanish).
Dalenius, T. (1957). Sampling in Sweden. Contribution to the methods and theories of sample survey practice. Almqvist and Wicksell, Stockholm.

Díaz-García, J.A., and Garay Tapia, M.M. (2007). Optimum allocation in stratified surveys: Stochastic programming. Computational Statistics and Data Analysis, 51, 3016-3026.
Giri, N.C. (1977). Multivariate statistical inference. New York: Academic Press.

Jones, D.F., Mirrazavi, S.K. and Tamiz, M. (2002). Multi-objective metaheuristics: An overview of the current state-of-the art. European Journal of Operational Research, 137, 1-9.
Khuri, A.I., and Cornell, J.A. (1987) Response Surface: Designs Analysis. New York: Marcel Dekker, Inc.
Miettinen, K.M. (1999). Non linear multiobjective optimization. Kluwer Academic Publishers, Boston.

Prékopa, A. (1978). The use of stochastic programming for the solution of some problems in statistics and probability. Technical Summary report 1983. University of Wisconsin-Madison.
Rao, S.S. (1979). Optimization theory and applications. Wiley Eastern Limited.

Ríos, S., Ríos Insua, S., and Ríos Insua, M.J. (1989). Procesos de decisión Multicriterio. EUDEMA, Madrid (in Spanish).

Steuer, R.E. (1986). Multiple criteria optimization: Theory, computation and applications. New York: John Wiley \& Sons, Inc.
Stuart, A. (1954). A simple presentation of optimum sampling results. Journal of the Royal Statistical Society, B, 16, 239-241.
Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). Sampling theory of surveys with application. State University Press.

Thompson, M.E. (1997). Theory of sample surveys. Chapman \& Hall.
Winston, W.L. (1995). Introduction to mathematical programming: Applications and algorithms. Duxbury Press.

# A balanced sampling approach for multi-way stratification designs for small area estimation 

Piero Demetrio Falorsi and Paolo Righi ${ }^{1}$


#### Abstract

The present work illustrates a sampling strategy useful for obtaining planned sample size for domains belonging to different partitions of the population and in order to guarantee the sampling errors of domain estimates be lower than given thresholds. The sampling strategy that covers the multivariate multi-domain case is useful when the overall sample size is bounded and consequently the standard solution of using a stratified sample with the strata given by cross-classification of variables defining the different partitions is not feasible since the number of strata is larger than the overall sample size. The proposed sampling strategy is based on the use of balanced sampling selection technique and on a GREG-type estimation. The main advantages of the solution is the computational feasibility which allows one to easily implement an overall small area strategy considering jointly the sampling design and the estimator and improving the efficiency of the direct domain estimators. An empirical simulation on real population data and different domain estimators shows the empirical properties of the examined sample strategy.


Key Words: Planning sampling size of small domains; Controlled selection; Balanced sampling.

## 1. Introduction

The small area problem is usually considered to be treated via estimation. However, if the domain indicator variables are available for each unit in the population there are opportunities to be exploited at the survey design stage. This condition is usually met in the business survey context where the domain indicator variables are available in the business register. As noted by Singh, Gambino and Mantel (1994), there is a need to develop an overall strategy that deals with small area problems, involving both planning sample design and estimation aspects. In this framework, it is crucial to control the sample size for each domain of interest, so that the domain is treated as a planned domain, at design stage, for which it is possible to produce direct estimates with a prefixed level of precision. In general, with a design-based approach to the inference, the presence of sample units in each domain allows one to compute domain estimates although not always reliably. Furthermore, in the model-based or model-assisted approach, the presence of sample units in each estimation domain allows one to use models with specific small area effects, giving more accurate estimates of the parameters of interest at small area level (Lehtonen, Särndal and Veijanen 2003). Marker $(1999,2001)$ deals with the problems of sampling design issues in small area context suggesting sample strategies, based on stratification and over-sampling, increasing the number of small areas for which accurate direct estimation is possible. These strategies are feasible in case of nested domains, but they may be unfeasible when the aim of the survey is to produce estimates for two or more partitions of
the population. A standard solution to obtain planned sample sizes for the domains of two or more partitions is to use a stratified sample in which strata are identified by crossclassification of variables defining the different partitions. In the following, this design will be denoted as crossclassification design. In many practical situations, however the cross-classification design is unsuitable since it needs the selection of at least a number of sampling units as large as the product of the number of categories of the stratification variables. Cochran well illustrates (1977, page 124) this problem giving a clear example in which the crossclassification design is unfeasible.

The above background is typical of the business survey context. The European Council Regulation on Structural Business Statistics establishes that the parameters of interest refer to estimation domains defined by three different partition subsets of the population of enterprises. For instance, as we may note by table 1.1, in Italy the total number of estimation domains is 1,821 ; while the number of non-empty strata of the cross-classification design is larger than 37,000 .

In order to overcome some problems of crossclassification designs, an easy strategy is to drop one or more stratifying variables or to group some of the categories. Nevertheless, some planned domains become unplanned and some of them can have small or null sample size.

Many methods have been proposed in the literature to keep under control the sample size in all the categories of the stratifying variables without using cross-classification design. These methods are generally referred to as multi-way stratification techniques, and have been developed under two

[^10]main approaches: (i) Latin Squares or Latin Lattices schemes (Jessen 1970); (ii) controlled rounding problems via linear programming (Lu and Sitter 2002). Both approaches have some drawbacks which have limited the use of multi-way stratification techniques as a standard solution for planning the survey sampling designs in real survey contexts. Indeed, as described in Falorsi, Orsini and Righi (2006), it is not possible to implement the Latin Lattices schemes in many real survey contexts; as for example if there are no population units in one or more cross-classification strata. The main weakness of the linear programming approach is the computational complexity. The sampling strategy considered in this paper does not suffer from the disadvantages of the above mentioned methods and allows one the control of the sample sizes for domains of interest, which are defined by different partitions of the reference population. Furthermore it guarantees that the sampling errors of domain estimates are lower than the given thresholds.

The proposed sampling strategy is based on the use of both a balanced sampling selection technique (Deville and Tillé 2004) and a GREG-type estimation (Lehtonen et al. 2003). As shown in the study on empirical data herein illustrated and in Falorsi and Righi (2008), the main advantages of this solution is the computational feasibility and the efficiency, that is the sampling errors for multi-domain-multivariate case are reasonably close to those defined by the optimal univariate solutions. This allows one to fairly implement an overall small-area strategy considering jointly the sampling design and the estimator and improving the efficiency of the direct domain estimators.

In some survey context, the proposed sampling strategy might define a too large overall sample size for assuring the prefixed bound of the direct domain estimates sampling errors. This may happen due to a too large number of domains of a given population partition. If the overall sample size is bounded by budget constraints, then the proposed sampling strategy with direct estimators may be not feasible. Therefore, it could be necessary to adopt an indirect smallarea estimator in order to control the mean square errors of partition domain estimates. However, the proposed approach may be easily extended to a strategy using the direct estimator and the indirect small area estimators for the
partitions requiring a too large overall sample size for bounding the sampling errors.

The paper is organised as follows. Section 2 states the problem, introduces the essential notation and describes the overall sampling strategy. Section 3 shows the algorithms for finding the inclusion probabilities and the corresponding planned domain sample sizes. Sections 4 and 5 illustrate two extensions of the sampling strategy. In section 4 the case in which the variance criterion is represented by the anticipated variance is studied. An extension to the case of a simple small area indirect estimator is presented in section 5. The main results of an empirical study on a real population of Italian enterprises are shown in section 6. Some brief conclusions are finally underlined in section 7 .

## 2. The sampling strategy

### 2.1 Parameters of interest

In order to define formally the problem, let us denote with $U$ a population of $N$ elements and with $b$ a specific partition of $U(b=1, \ldots, B)$ in which $b^{\text {th }}$ partition defines $M_{b}$ different non overlapping domains, $U_{b d}(d=1, \ldots$, $M_{b}$ ), of size $N_{b d}$ being $\sum_{d=1}^{M_{b}} N_{b d}=N$ and, finally let $\sum_{b=1}^{B} M_{b}=Q$ be the overall number of domains.

Let $y_{r, k}$ and ${ }_{b d} \delta_{k}$ denote respectively the value of the $r^{\text {th }}(r=1, \ldots, R)$ variable of interest in the $k^{\text {th }}$ population unit and the domain membership indicator, being ${ }_{b d} \delta_{k}=1$ if $k \in U_{b d}$ and ${ }_{b d} \delta_{k}=0$, otherwise. Let us suppose that the ${ }_{b d} \delta_{k}$ values are known for each unit in the population.

The parameters of interest are the $M=Q \times R$ domains totals

$$
\begin{align*}
&{ }_{b d} t_{r}=\sum_{k \in U} y_{r, k d d} \delta_{k}=\sum_{k \in U_{b d}} y_{r, k} \\
& \quad\left(r=1, \ldots, R ; b=1, \ldots, B ; d=1, \ldots, M_{b}\right) . \tag{2.1.1}
\end{align*}
$$

The expression (2.1.1) defines a multivariate-multidomain problem since there are $R$ variables of interest (multivariate aspect) and $Q>1$ domains (multi-domain aspect).

Table 1.1
Number of domains of the Italian Structural Business Statistics Survey by partition

| Partitions | Number of domains |
| :--- | :---: |
| Economic activity class (4-digits of the NACE rev.1 classification) | 465 |
| Economic activity group (3-digits of the NACE rev.1 classification) by Size class ${ }^{1}$ | 395 |
| Economic activity division (2-digits of the NACE rev.1 classification) by Region $^{1}$ | 961 |
| Total number of estimation domains | 1,821 |

[^11]
### 2.2 A concise description of the sampling strategy

Let us suppose that, in order to estimate the ${ }_{b d} t_{r}$ parameters, a sample $s$ of fixed size $n$ is selected from population $U$, with inclusion probabilities $\pi_{k}(k \in U)$. Let $s_{b d}=s \cap U_{b d}$ be the sample of $n_{b d}$ units belonging to the $U_{b d}$ domain (with $\sum_{d=1}^{M_{b}} n_{b d}=n$ ), being

$$
\begin{equation*}
n_{b d}=\sum_{k \in U_{b d}} \lambda_{k}=\sum_{k \in U_{b d}} \pi_{k}, \tag{2.2.1}
\end{equation*}
$$

with $\lambda_{k}=1$ if $k \in s$ and $\lambda_{k}=0$ otherwise.
The sample is selected by a multi-way stratification technique developed under the balanced sampling framework guaranteeing that the selected sample respects the following balancing equations

$$
\begin{equation*}
\hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}} \tag{2.2.2}
\end{equation*}
$$

where $\hat{t}_{\mathbf{z}, h t}=\sum_{k \in U} \mathbf{z}_{k} \lambda_{k} a_{k}$ denote the Horvitz-Thompson estimates of $t_{\mathbf{z}}=\sum_{k \in U} \mathbf{z}_{k}$, being $\mathbf{z}_{k}$ a value vector of auxiliary variables known for each population unit at the design stage and $a_{k}=1 / \pi_{k}$. A suitable specification of the $\mathbf{z}_{k}$ vectors can assure that the realized sample sizes, $n_{b d}$, are equal to fixed quantities known in advance, as described in section 2.3.

The estimates of ${ }_{b d} t_{r}$, denoted with ${ }_{b d} \hat{t}_{r \text {, greg }}$, are obtained with the modified GREG estimator (Rao 2003, page 20), given by:

$$
\begin{equation*}
{ }_{b d} \hat{t}_{r, \text { greg }}=\sum_{k \in s}{ }_{b d} w_{k} y_{r, k} \tag{2.2.3}
\end{equation*}
$$

where

$$
\begin{aligned}
b_{d d} w_{k} & =a_{k b d} \delta_{k} \\
& +\left({ }_{b d} t_{\mathbf{x}}-{ }_{b d} \hat{t}_{\mathbf{x}, h t}\right)^{\prime}\left[\sum_{k \in s}\left(a_{k} \mathbf{x}_{k} \mathbf{x}_{k}^{\prime} / c_{k}\right)\right]^{-1} a_{k} \mathbf{x}_{k} / c_{k}
\end{aligned}
$$

denote the sampling weights, $\mathbf{x}_{k}$ indicates a value vector of the auxiliary variables, $c_{k}$ is a known constant, being ${ }_{b d} t_{\mathbf{x}}=\sum_{k \in U_{b d}} \mathbf{x}_{k}$ and ${ }_{b d} \hat{t}_{\mathbf{x}, h t}=\sum_{k \in s_{b d}} \mathbf{x}_{k} a_{k}$. The estimator (2.2.3), may be derived under the following working super population model

$$
\begin{equation*}
y_{r, k}=\mathbf{x}_{k}^{\prime} \boldsymbol{\beta}_{r}+\varepsilon_{r, k} \tag{2.2.4}
\end{equation*}
$$

where $\boldsymbol{\beta}_{r}$ is an unknown vector of fixed regression parameters and $\varepsilon_{r, k}$ is the random residual. The model expectation, $E_{m}$, and model variances, $V_{m}$, are respectively given by $\quad E_{m}\left({ }_{r} \varepsilon_{k}\right)=0 ; V_{m}\left(\varepsilon_{r, k}\right)=c_{k} \sigma_{r}^{2} ; E_{m}\left(\varepsilon_{r, k}, \varepsilon_{r, i}\right)=0 \quad$ if $k \neq i$.

The approximated sampling variance of the modified GREG estimator under balanced sampling is:

$$
\begin{equation*}
V_{p}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathrm{z}}\right)=\frac{N}{N-Q} \sum_{k \in U}\left(\frac{1}{\pi_{k}}-1\right)_{b d} \eta_{r, k}^{2}, \tag{2.2.5}
\end{equation*}
$$

being

$$
{ }_{b d} \eta_{r, k}= \begin{cases}\varepsilon_{r, k}-\mathbf{z}_{k b d}^{\prime} \mathbf{B}_{\mathbf{z}, \varepsilon} & \text { for } k \in U_{b d}  \tag{2.2.6}\\ -\mathbf{z}_{k b d}^{\prime} \mathbf{B}_{\mathbf{z}, \varepsilon} & \text { for } k \in U_{b \bar{d}}\end{cases}
$$

where

$$
{ }_{b d} \mathbf{B}_{\mathbf{z}, \varepsilon}=\left[\sum_{k \in U} \mathbf{z}_{k} \mathbf{z}_{k}^{\prime}\left(1 / \pi_{k}-1\right)\right]^{-1} \sum_{k \in U} \mathbf{z}_{k} \varepsilon_{r, k b d} \delta_{k}\left(1 / \pi_{k}-1\right)
$$

being $U_{b \bar{d}}$ the subset of $U$ complementary to $U_{b d}$. A proof of (2.2.5) is given in section 2.5.

The inclusion probabilities, $\pi_{k}$, and the domain sample sizes, $n_{b d}$, are determined with a procedure which attempts to minimize the overall sample size, $n$, guaranteeing that the sampling variances are lower than prefixed level of precision thresholds, ${ }_{b d} \bar{V}_{r}: V_{p}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) \leq_{b d} \bar{V}_{r}(b=1, \ldots$, $\left.B ; d=1, \ldots, M_{b} ; r=1, \ldots, R\right)$. The technical details are described in section 3.

Let us note that two different sets of covariates have been introduced in order to underline that the set of covariates available at the design stage ( $\mathbf{z}$ variables) could be different from the set available at the estimation stage ( $\mathbf{x}$ variables) even if in many practical situations they could be the same. As for example the covariates at estimation stage could be updated with respect to those available at the design stage. In our context (see section 2.3) the $\mathbf{z}_{k}$ vectors are characterized as specified by the expression (2.3.2) being defined only by the domain membership indicator variables and by the inclusion probabilities, while the $\mathbf{x}_{k}$ vectors could contain the values of some other variables more explicative of the phenomena of interest. For instance, in the business survey context the $\mathbf{x}$ variables could include, among others, the number of employees or the turnover.

### 2.3 The balanced sampling for marginal stratification

Multi-way stratification designs can be treated in the context of the balanced sampling.

The definition of a balanced sample depends on the assumed inferential framework. In the model based approach, a sample is defined as balanced on a set of auxiliary variables if there is the equality between the sample and the known population means of the auxiliary variables (Valliant, Dorfman and Royall 2000). Following the design based (or model assisted approach) considered in this paper, a sample is balanced when the HorvitzThompson estimates of the auxiliary variables totals are equal to their known population totals (Deville and Tillé 2004).

For defining the balanced sampling in the design or model assisted approach, let us introduce the general definition of sampling design as a probability distribution $p(\cdot)$ on the set $S$ of all the subset $s$ of the population $U$
such that $\sum_{s \in \mathrm{~S}} p(s)=1$, where $p(s)$ is the probability of the sample $s$ to be drawn. Each set $s$ may be represented by the outcome $\lambda^{\prime}=\left(\lambda_{1}, \ldots, \lambda_{k}, \ldots, \lambda_{N}\right)$ of a vector of $N$ random variables. Let $\boldsymbol{\pi}^{\prime}=\left(\pi_{1}, \ldots, \pi_{k}, \ldots, \pi_{N}\right)$ be the vector of inclusion probabilities, where $\pi=E_{p}(\lambda)=\sum_{s \in S} p(s) \lambda$, being $E_{p}(\cdot)$ the expected value over repeated sampling. Let $\mathbf{z}_{k}^{\prime}=\left(z_{l k}, \ldots, z_{h k}, \ldots, z_{Q k}\right)$ be a vector of $Q$ auxiliary variables available for each population unit. The sampling design $p(s)$ with inclusion probabilities $\pi$ is said to be balanced with respect to the $Q$ auxiliary variables if and only if it satisfies the balancing equations given by (2.2.2) for all $s \in S$ such that $p(s)>0$.

Let us suppose that a vector of inclusion probabilities $\pi$, consistent with the marginal sampling distributions $n_{b d}$ $\left(b=1, \ldots, B ; d=1, \ldots, M_{b}\right)$, is available, that is

$$
\begin{equation*}
n_{b d}=\sum_{k \in U_{b d}} \pi_{k}\left(b=1, \ldots, B ; d=1, \ldots, M_{b}\right) . \tag{2.3.1}
\end{equation*}
$$

Multi-way stratification design represents a special case of balanced design where for unit $k$ the auxiliary variable vector is given by

$$
\begin{align*}
\mathbf{z}_{k}^{\prime} & =(\underbrace{(\overbrace{0, \ldots, \pi_{k}, \ldots, 0}^{b=1}, \ldots, \overbrace{0, \ldots, \pi_{k}, \ldots, 0}^{b=B}}_{Q}) \\
& =\pi_{k}\left({ }_{11} \delta_{k}, \ldots,{ }_{b d} \delta_{k}, \ldots,{ }_{B M_{B}} \delta_{k}\right) . \tag{2.3.2}
\end{align*}
$$

The expression (2.3.2) defines the $\mathbf{z}_{k}$ as vectors of $(Q-B)$ zeros and with $B$ entries equal to $\pi_{k}$ in the places indicating the domains which the unit $k$ belongs to. When defining the $\mathbf{z}_{k}$ vector as (2.3.2), if condition (2.3.1) holds, the selection of sample satisfying the system of balancing equations (2.2.2), $\sum_{k \in U}\left(\mathbf{z}_{k} \lambda_{k}\right) / \pi_{k}=\sum_{k \in s} \mathbf{z}_{k}$, guarantees that the $n_{b d}$ values are non random quantities. The left hand-side of the balancing equation (2.2.2) is $\sum_{k \in U}\left(\pi_{k b d} \delta_{k} \lambda_{k}\right) / \pi_{k}=\sum_{k \in U_{b d}} \lambda_{k}=n_{b d}$, while the right hand-side is $\sum_{k \in U} \pi_{k b d} \delta_{k}=\sum_{k \in U_{b d}} \pi_{k}=n_{b d}$.

Deville and Tillé (2004) proposed the cube method that allows one the selection of balanced (or approximately balanced) samples for a large set of auxiliary variables and with respect to different vectors of inclusion probabilities. In particular, Deville and Tillé (2000) show that with specification (2.3.2) of the $\mathbf{z}_{k}$ vectors, the balancing equations (2.2.2) can be exactly satisfied. The cube method is implemented by an enhanced algorithm for large data sets (Chauvet and Tillé 2006) available in a free software code that may be downloaded in the website http://www.insee.fr/ fr/nom_df_met/outils_stat/cube/accueil_cube.htm.

### 2.4 The modified direct GREG estimator

Following Lehtonen et al. (2003), the estimator (2.2.3), may be expressed under the general form

$$
b \hat{d}_{r, \operatorname{greg}}=\sum_{k \in U_{b d}} \tilde{y}_{r, k}+\sum_{k \in s_{s d}} a_{k}\left(y_{r, k}-\tilde{y}_{r, k}\right)
$$

where $\tilde{y}_{r, k}$ denotes the prediction of $y_{r, k}$ under the assumed super population model. The predictions $\left\{\tilde{y}_{r, k}\right.$; $k \in U\}$ differ from one model specification to another, depending on the functional form and from the choice of the auxiliary variables. The estimator (2.2.3) is derived under the working super-population model (2.2.4). The predictions $\tilde{y}_{r, k}$ are then obtained by

$$
\begin{equation*}
\tilde{y}_{r, k}=\mathbf{x}_{k}^{\prime} \hat{\boldsymbol{\beta}}_{r}, \tag{2.4.2}
\end{equation*}
$$

being

$$
\begin{equation*}
\hat{\boldsymbol{\beta}}_{r}=\left(\sum_{k \in s} \mathbf{x}_{k} \mathbf{x}_{k}^{\prime} a_{k} / c_{k}\right)^{-1} \sum_{k \in s} \mathbf{x}_{k} y_{r, k} a_{k} / c_{k} . \tag{2.4.3}
\end{equation*}
$$

Let us observe that the linear model (2.2.4) allows one to define the estimator only knowing the domain totals of the auxiliary information and the $\mathbf{x}_{k}$ values for the sampling units. However, knowing the $\mathbf{x}_{k}$ values for every $k \in U$, it is possible to build an estimators with more efficient predictions $\tilde{y}_{r, k}$ obtained by generalized linear models (Lehtonen and Veijanen 1998) or non parametric regression techniques (Montanari and Ranalli 2003).

As noted by Rao (2003, page 20) the estimator (2.2.3) is approximately design unbiased as the overall sample size increases, even if the domain sample size $n_{b d}$ is small. Moreover, the sum of the ${ }_{b d} \hat{t}_{r}$,greg estimates over all the domains of a partitions is benchmarked to the usual GREG estimate of the total, $\sum_{d=1}^{M_{b}} \hat{t}_{r} \hat{r}_{\text {, reg }}=\sum_{k \in s} y_{r, k} a_{k}[1+$ $\left.\left(\sum_{k \in U} \mathbf{x}_{k}-\sum_{k \in S} \mathbf{x}_{k} a_{k}\right)^{\prime}\left(\sum_{k \in s} \mathbf{x}_{k} \mathbf{x}_{k}^{\prime} a_{k} / c_{k}\right)^{-1} \mathbf{x}_{k} / c_{k}\right]$.

### 2.5 Sampling variances

In order to derive the expression of the variance (2.2.5), consider the results given by Deville and Tillé (2005). They have proposed approximating the variance of the HorvitzThompson estimator $\hat{r}_{r, h t}=\sum_{k \in s} y_{r, k} a_{k}$ of the total $t_{r}=$ $\sum_{k \in U} y_{r, k}$, by supposing that the balanced sampling can be viewed as a conditional Poisson sampling and assuming that, at least for large sample sizes, the inclusion probabilities $\pi_{k}$ well approximate the inclusion probabilities of the Poisson design. Assuming that, through Poisson sampling, the vector $\left(\hat{t}_{r, h t}, \hat{t}_{\mathbf{z}, h t}^{\prime}\right)^{\prime}$ has approximately a multinormal distribution, the authors suggest a good approximation of the sampling variance given by

$$
\begin{align*}
V_{p}\left(\hat{t}_{r, h t} \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) & =V_{p}\left(\hat{t}_{r, h t}+\left(t_{\mathbf{z}}-\hat{t}_{\mathbf{z}, h t}\right)^{\prime} \mathbf{B}_{\mathbf{z}, y}\right) \\
& =V_{p}\left(\hat{t_{r, h t}}-\hat{t}_{\mathbf{z}, h t}^{\prime} \mathbf{B}_{\mathbf{z}, y}\right) \\
& =V_{p}\left(\sum_{k \in s} a_{k}\left(y_{r, k}-\mathbf{z}_{k}^{\prime} \mathbf{B}_{\mathbf{z}, y}\right)\right) \\
& \cong \frac{N}{N-Q} \sum_{k \in U}\left(\frac{1}{\pi_{k}}-1\right)\left(y_{r, k}-\mathbf{z}_{k}^{\prime} \mathbf{B}_{\mathbf{z}, y}\right)^{2} \tag{2.5.1}
\end{align*}
$$

where $\mathbf{B}_{\mathbf{z}, y}=\left[\sum_{k \in U} \mathbf{z}_{k} \mathbf{z}_{k}^{\prime}\left(1 / \pi_{k}-1\right)\right]^{-1} \sum_{k \in U} \mathbf{z}_{k} y_{r, k}\left(1 / \pi_{k}-1\right)$. The expression (2.5.1) has been validated by a set of simulations.

Let us consider, now, the linear approximation, ${ }_{b d} \hat{t}_{r, \text { greg }}^{*}$, of the GREG estimator, the derivation of which may be obtained according to Särndal, Swensson and Wretman (1992, pages 450-451)

$$
\begin{align*}
{ }_{b d} \hat{t}_{r, \text { greg }} \simeq_{b d} \hat{t}_{r, \text { greg }}^{*} & =\sum_{k \in U_{b d}} \mathbf{x}_{k}^{\prime} \boldsymbol{\beta}_{r}+\sum_{k \in s_{b d}} a_{k} \varepsilon_{r, k} \\
& =\sum_{k \in U_{b d}} \mathbf{x}_{k}^{\prime} \boldsymbol{\beta}_{r}+\sum_{k \in s} a_{k} \varepsilon_{r, k b d} \delta_{k} . \tag{2.5.2}
\end{align*}
$$

On the basis of expressions (2.5.1) and (2.5.2), it is possible to derive the following result

$$
\begin{aligned}
V_{p}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) & \cong V_{p}\left({ }_{b d} \hat{t}_{r, \text { greg }}^{*} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) \\
& =V_{p}\left(\sum_{k \in s} a_{k} \varepsilon_{r, k b d} \delta_{k} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) \\
& =V_{p}\left(\sum_{k \in s} a_{k} \varepsilon_{r, k b d} \delta_{k}+\left(t_{\mathbf{z}}-\hat{t}_{\mathbf{z}, h t}\right)^{\prime}{ }_{b d} \mathbf{B}_{\mathbf{z}, \varepsilon}\right) \\
& =V_{p}\left(\sum_{k \in s} a_{k}\left(\varepsilon_{r, k b d} \delta_{k}-\mathbf{z}_{k b d}^{\prime} \mathbf{B}_{\mathbf{z}, \varepsilon}\right)\right. \\
& =V_{p}\left(\sum_{k \in s} a_{k b d} \eta_{r, k}\right) \\
& \cong \frac{N}{N-Q} \sum_{k \in U}\left(\frac{1}{\pi_{k}}-1\right)_{b d} \eta_{r, k}^{2}
\end{aligned}
$$

where ${ }_{b d} \eta_{r, k}^{2}$ is defined in (2.2.6).
The approximated sampling variance of ${ }_{b d} \hat{t}_{r, \text { greg }}$ depends on the residuals of the whole set of units, because of balanced selection. Therefore, the units not belonging to $U_{b d}$ have an influence on the sampling variance of the estimator.

Let us examine now the univariate unidomain case and assume that the survey has an unique target parameter, ${ }_{b d} t_{r}$. Furthermore, let us suppose that the selected sample respects the balancing equations, $\hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}$, being fixed the overall sample size $n$.

Following the arguments proposed by Särndal et al. (1992; Result 12.2.1, page 452), it is trivial to prove that, in this sampling context, each unit $k$ could be selected with $(Q \times R)$ different optimal inclusion probabilities, ${ }_{b d} \ddot{\pi}_{r, k}$ $\left(b=1, \ldots, B ; d=1, \ldots, M_{b} ; r=1, \ldots, R\right)$

$$
{ }_{b d} \ddot{r}_{r, k}=\left.n\right|_{b d} \eta_{r, k}\left|/ \sum_{i \in U}\right|_{b d} \eta_{r, i} \mid,
$$

which allow one to attain the $(Q \times R)$ different lower bounds, ${ }_{b d} V_{r \mid n}^{*}$, of the approximated variances:

$$
\begin{aligned}
V_{p}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}\right. & \left.=t_{\mathbf{z}}\right) \geq{ }_{b d} V_{r \mid n}^{*}= \\
& \frac{N}{N-Q}\left[\frac{1}{n}\left(\sum_{k \in U}\left|{ }_{b d} \eta_{r, k}\right|\right)^{2}-\sum_{k \in U}{ }_{b d} \eta_{r, k}^{2}\right] .
\end{aligned}
$$

Let us finally underline that in Tillé and Favre (2005) is given a criterion for obtaining a prediction ${ }_{b d} \hat{\eta}_{r, k}$ of the ${ }_{b d} \eta_{r, k}$ values, that may be used in repeated sampling contexts.

## 3. Sampling algorithms for the determination of the domain sample sizes

The inclusion probabilities $\pi_{k}$ and the derived domain sample sizes, $n_{b d}=\sum_{k \in U_{b d}} \pi_{k}$, are obtained with a two steps procedure: (i) in the first step, denoted as optimization, the preliminary inclusion probabilities, $\pi_{k}^{\prime}$, are determined solving a minimum constrained problem; (ii) in the second step, denoted as calibration, the inclusion probabilities, $\pi_{k}$, are obtained as a slight modification of the $\pi_{k}^{\prime}$; the calibration problem is implemented for assuring that the domain sample sizes $n_{b d}$ are integers.

As illustrated in the following, the $\pi_{k}$ values may be expressed as implicit functions of the unknown residuals ${ }_{b d} \eta_{r, k}^{2}$. But, in real survey context, the determination of the inclusion probabilities $\pi_{k}$ may be done using the predicttions ${ }_{b d} \hat{\eta}_{r, k}^{2}$ instead of ${ }_{b d} \eta_{r, k}^{2}$. This is a general problem concerning the planning the sampling designs, because the variances are generally unknown quantities that may be suitably estimated. In repeated survey contexts the effect of using the estimates ${ }_{b d} \hat{\eta}_{r, k}^{2}$ as a replacement for ${ }_{b d} \eta_{r, k}^{2}$ may be tested by computing the sampling variances after the data collection. The empirical results may then be used for introducing proper adjustments in planning the next survey design. However, as illustrated in the empirical analysis and in Falorsi and Righi (2008), the proposed strategy seems to be efficient and sufficiently robust with respect to small departures of ideal conditions.

The sections 3.1 and 3.2 respectively describe the two steps of the algorithm for the determination of the domain sample sizes. A simplified allocation rule, which seems to be worthwhile in many real survey contexts, is described in section 3.3.

### 3.1 First step: Optimization

The inclusion probabilities $\pi_{k}^{\prime}$ can be defined as solution of the following non linear programming problem with $N$ unknowns, $\pi_{k}^{\prime}$, and $(N+Q \times R)$ constraints

$$
\left\{\begin{array}{l}
\operatorname{Min}\left(\sum_{k \in U} \pi_{k}^{\prime}\right) \\
\frac{N}{N-Q} \sum_{k \in U}\left(\frac{1}{\pi_{k}^{\prime}}-1\right){ }_{b d} \eta_{r, k}^{2} \leq_{b d} \bar{V}_{r}  \tag{3.1.1}\\
\quad\left(b=1, \ldots, B ; d=1, \ldots, M_{b} ; r=1, \ldots, R\right) \\
0<\pi_{k}^{\prime} \leq 1 \quad(k=1, \ldots, N) .
\end{array}\right.
$$

A numerical solution to (3.1.1) may be derived considering the algorithms developed for the multivariate allocation in stratified surveys. Such algorithms allow one to find the unknown values $v_{h}>0(h=1,2, \ldots)$ which represent the solution of the following non linear problem $\operatorname{Min}\left(\sum_{h} v_{h}\right)$ under the constraints $\sum_{h} A_{r h} / v_{h} \leq \bar{A}_{r}$, where $A_{r h}$ and $\bar{A}_{r}(r=1,2, \ldots)$ are known positive quantities.

Bethel (1989) invokes the Kuhn-Tucker theorem to show that there exists a solution to the above problem. He describes a simple algorithm and discusses its convergence properties. Chromy (1987) develops an algorithm, suitable for automated spreadsheets but without an explicit proof that always converges. A slight modification of the Chromy's algorithm - able to solve the problem (3.1.1) guaranteeing the inequalities $0<\pi_{k}^{\prime} \leq 1 \quad(k=1, \ldots, N)$ are respected is described herein in the following. After the Initialization, the algorithm finds the $\pi_{k}^{\prime}$ values by iterating the two actions of Calculus and Check. As far as the convergence issue is concerned, it is worthwhile to note that the Chromy's algorithm have been mostly used for stratified sampling design, and indeed, the documentation refers to stratified samples. In the applied sampling literature, there is a lot of empirical proofs of the successful use of the algorithm in this sampling context. Let us note that the modification of the Chromy's algorithm, herein proposed, treats the sampling units as strata and the resulting allocation, being fractional, defines the inclusion probabilities. Also in this case there is no formal proof that the proposed modified algorithm converges. Nevertheless, in all the different empirical experiments developed by the authors the algorithm has always converged and no critical conditions have been encountered.

Initialization: at initial iteration $(\tau=0)$, set ${ }^{\tau} \gamma_{k}=1$ $(k=1, \ldots, N)$.
Calculus: the generic iteration $(\tau=1,2, \ldots)$ consists of a sequence of steps denoted with $u=(0,1,2, \ldots)$.

- At initial step $(u=0)$, set ${ }_{b d}^{\tau, u} \phi_{r}=1$ and calculate

$$
{ }_{b d}^{\tau} V_{0 r}=\frac{N}{N-Q} \sum_{k \in U} \quad{ }_{b d} \eta_{r, k}^{2}{ }^{\tau} \gamma_{k} .
$$

- At subsequent steps $(u=1,2, \ldots)$, calculate the values of the following equations

$$
\begin{align*}
& { }_{\tau, u} \pi_{k}= \\
& {\left[\left(1-{ }^{\tau} \gamma_{k}\right)+{ }^{\tau} \gamma_{k} \frac{N}{N-Q} \sum_{b=1}^{B} \sum_{d=1}^{M_{b}} \sum_{r=1}^{R}{ }_{b d}^{\tau, u} \phi_{r b d} \eta_{r, k}^{2}\right]^{1 / 2} .}  \tag{3.1.2}\\
& \quad{ }_{b d}^{\tau, u} V_{r}=\frac{N}{N-Q} \sum_{k \in U} \frac{1}{\tau, u} \pi_{k}{ }_{b d} \eta_{r, k}^{2}{ }^{\tau} \gamma_{k},
\end{align*}
$$

and

$$
\begin{equation*}
{ }_{b d}^{\tau, u} V_{r}^{\prime}={ }_{b d}^{\tau, u} V_{r}+{ }_{b d}^{\tau} V_{0 r} \tag{3.1.3}
\end{equation*}
$$

- If the following two conditions:

$$
\begin{equation*}
{ }_{b d}^{\tau, u} V_{r}^{\prime} \leq_{b d} \bar{V}_{r} \text { and }{ }_{b d}^{\tau, u} \phi_{r}\left({ }_{b d}^{\tau, u} V_{r}^{\prime}-{ }_{b d} \bar{V}_{r}\right)=0, \tag{3.1.4}
\end{equation*}
$$

are respected (for all $b=1, \ldots, B ; d=1, \ldots, M_{b} ; r=$ $1, \ldots, R)$ then the action of Calculus stops and the inclusion probabilities ${ }^{\tau} \pi_{k}$ are those calculated in equation (3.1.2). Otherwise, the updated quantities ${ }_{b d}^{\tau, u+1} \phi_{r}$ are computed

$$
\begin{equation*}
\left.{ }_{b d}^{\tau, u+1} \phi_{r}={ }_{b d}^{\tau, u} \phi_{r}\left[{ }_{b d}^{\tau, u} V_{r} / c_{b d}^{\tau, u} V_{r}^{\prime}-{ }_{b d}^{\tau} \bar{V}_{r}\right)\right]^{2} \tag{3.1.5}
\end{equation*}
$$

and the equations (3.1.2) and (3.1.3) are calculated at $u+1$, over and over again with ${ }^{\tau}, u+1$ b $\phi_{r}$ replacing ${ }_{b d}^{\tau, u} \phi_{r}$ until conditions (3.1.4) are respected.

Check: if the condition ${ }^{\tau} \pi_{k} \leq 1$ is true for all $k$, then the algorithm stops and the $\pi_{k}^{\prime}$ values are set equal to $\pi_{k}^{\prime}=$ ${ }^{\tau} \pi_{k}$. Otherwise the ${ }^{\tau} \gamma_{k}$ values are updated as ${ }^{\tau+1} \gamma_{k}=1$ if ${ }^{\tau} \pi_{k} \leq 1$ and ${ }^{\tau+1} \gamma_{k}=0$ if ${ }^{\tau} \pi_{k}>1$. The calculus is iterated at $\tau+1$ with ${ }^{\tau+1} \gamma_{k}$ replacing ${ }^{\tau} \gamma_{k}$. A SAS macro that allows one to solve the problem (3.1.1) has been developed by the authors of this paper and may be released on demand.

### 3.2 Second step: Calibration

The quantities $n_{b d}$ are defined, first, by rounding the results of the $Q$ sums, $\sum_{k \in U_{b d}} \pi_{k}^{\prime}\left(b=1, \ldots, B ; d=1, \ldots, M_{b}\right)$. Sometimes a further data manipulation could be necessary in order to assure the condition $\sum_{d=1}^{M_{b}} n_{b d}=n$ for each $b$. The probabilities $\pi_{k}$ are then obtained as solution of calibration problem

$$
\left\{\begin{array}{l}
\operatorname{Min}\left(\sum_{k \in U} G\left(\pi_{k} ; \pi_{k}^{\prime}\right)\right)  \tag{3.2.1}\\
\sum_{k \in U} \pi_{k}=n, \sum_{k \in U_{b d}} \pi_{k}=n_{b d} \\
\quad\left(b=1, \ldots, B ; d=1, \ldots, M_{b}-1\right),
\end{array}\right.
$$

where, $G\left(\pi_{k} ; \pi_{k}^{\prime}\right)$ is a distance function between $\pi_{k}$ and $\pi_{k}^{\prime}$. Note that (3.2.1) may be solved by the well known Iterative Proportional Fitting algorithm (Bishop, Fienberg and Holland 1975) or the Generalized Iterative Proportional Fitting algorithm (GIPF; Dykstra and Wollan 1987)
procedures. The logarithmic distance function $G\left(\pi_{k} ; \pi_{k}^{\prime}\right)=$ $\pi_{k} \ln \left(\pi_{k} / \pi_{k}^{\prime}\right)-\left(\pi_{k}+\pi_{k}^{\prime}\right)$ avoids to define the $\pi_{k}$ probabilities lower than 0 , while GIPF prevents to obtain $\pi_{k}$ values larger than 1 .

### 3.3 A simplified allocation rule

In many real survey contexts in which the overall sample size $n$ is fixed and there is not enough information to obtain good predictions ${ }_{b d} \hat{\eta}_{r, k}^{2}$ of the ${ }_{b d} \eta_{r, k}^{2}$ values, the following procedure may be implemented. Firstly the marginal sample sizes $n_{b d}$ are determined by a quite simple rule

$$
\begin{equation*}
n_{b d}=\alpha_{b} n\left(N_{b d} / N\right)+\left(1-\alpha_{b}\right) n / M_{b} \tag{3.3.1}
\end{equation*}
$$

being $\alpha_{b}\left(0 \leq \alpha_{b} \leq 1\right)$ a fixed constant which have to be properly defined. The (3.3.1) turns out to be a compromise between the allocation proportional to population size ( $\alpha_{b}=1$ ) and the allocation uniform for each domain of a given partition ( $\alpha_{b}=0$ ).

The probabilities $\pi_{k}$ are then obtained as solution of the calibration problem (3.2.1) where the marginal sample sizes are computed as above indicated and the initial probabilities $\pi_{k}^{\prime}$ are set uniformly equal to $\pi_{k}^{\prime}=n / N$. The resulting inclusion probabilities are no more optimal, in the sense above described and do not guarantee that the sampling variances are lower than prefixed level of precision thresholds. However they are computed with a reasonable procedure, which may be fairly implemented and thus representing an interesting point of reference with respect to any real survey context.

## 4. The anticipated variance

A frequently used criterion for planning the sampling strategies is that of controlling the anticipated variance, which may be defined as:

$$
\begin{equation*}
\operatorname{AV}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\boldsymbol{z}, h t}=t_{\mathbf{z}}\right)=E_{m} E_{p}\left({ }_{b d} \hat{t}_{r, \text { greg }}-{ }_{b d} t_{r} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right)^{2} . \tag{4.1}
\end{equation*}
$$

The following result may be derived under the assumptions of the model (2.2.4) and using the results given in section (2.5):

$$
\begin{align*}
& \operatorname{AV}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) \\
& \begin{aligned}
\cong & E_{m} V_{p}\left({ }_{b d} \hat{t}_{r, \mathrm{greg}}^{*} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) \\
= & E_{m}\left[\frac{N}{N-Q} \sum_{k \in U_{b d}}\left(\frac{1}{\pi_{k}}-1\right) \varepsilon_{r, k}^{2}\right. \\
& \frac{N}{N-Q} \sum_{k \in U}\left(\frac{1}{\pi_{k}}-1\right)\left(\mathbf{z}_{k b d}^{\prime} \mathbf{B}_{\mathbf{z}, \mathrm{\varepsilon}}\right)^{2} \\
\quad & \left.\quad 2 \frac{N}{N-Q} \sum_{k \in U_{b d}}\left(\frac{1}{\pi_{k}}-1\right) \varepsilon_{r, k} \mathbf{z}_{k b d}^{\prime} \mathbf{B}_{\mathbf{z}, \varepsilon}\right] \\
= & \frac{N}{N-Q} \sum_{k \in U}\left(\frac{1}{\pi_{k}}-1\right)_{b d}^{a} \eta_{r, k}^{2}
\end{aligned}
\end{align*}
$$

being

$$
{ }_{b d}^{a} \eta_{r, k}^{2}= \begin{cases}\sigma_{r}^{2} c_{k}\left(1-g_{k k}\right)^{2}+\sigma_{r}^{2} \sum_{j(\neq k) \in U_{b d}} g_{k j}^{2} c_{j} & \text { if } k \in U_{b d} \\ \sigma_{r}^{2} \sum_{j \in U_{b d}} g_{k j}^{2} c_{k} & \text { otherwise }\end{cases}
$$

where: $\left(g_{k 1}, \ldots, g_{k j}, \ldots, g_{k N}\right)=\mathbf{g}_{k}^{\prime}=\mathbf{z}_{k}^{\prime}\left(\mathbf{Z}_{U}^{\prime} \mathbf{\Omega}_{U}^{-1} \mathbf{Z}_{U}\right)^{-1} \mathbf{Z}_{U}^{\prime} \mathbf{\Omega}_{U}^{-1}$, $\mathbf{Z}_{U}=\operatorname{col}\left\{\mathbf{Z}_{k}^{\prime}\right\}_{k=1}^{N}, \mathbf{\Omega}_{U}^{-1}=\operatorname{diag}\left\{1 / \pi_{k}-1\right\}_{k=1}^{N}$. The expression (4.2) has been derived using the following two results

$$
\begin{aligned}
& E_{m}\left(\mathbf{z}_{k b d}^{\prime} \mathbf{B}_{\mathbf{z}, \varepsilon}\right)^{2} \\
& \quad=\sigma_{r}^{2} \mathbf{z}_{k}^{\prime}\left(\mathbf{Z}_{U}^{\prime} \mathbf{\Omega}_{U}^{-1} \mathbf{Z}_{U}\right)^{-1} \mathbf{Z}_{U}^{\prime} \mathbf{\Omega}_{U b d}^{-1} \mathbf{V}_{r} \mathbf{\Omega}_{U}^{-1} \mathbf{Z}_{U}\left(\mathbf{Z}_{U}^{\prime} \mathbf{\Omega}_{U}^{-1} \mathbf{Z}_{U}\right)^{-1} \mathbf{z}_{k} \\
& \quad= \\
& =\sigma_{r}^{2} \mathbf{g}_{k b d}^{\prime} V_{r} \mathbf{g}_{k}=\sigma_{r}^{2} \sum_{j \in U_{b d}} g_{k j}^{2} c_{k}, \\
& E_{m}\left(\varepsilon_{r, k} \mathbf{z}_{k b d}^{\prime} \mathbf{B}_{\mathbf{z}, \varepsilon}\right) \\
& =\mathbf{z}_{k}^{\prime}\left(\mathbf{Z}_{U}^{\prime} \mathbf{\Omega}_{U}^{-1} \mathbf{Z}_{U}\right)^{-1} \mathbf{Z}_{U}^{\prime} \mathbf{\Omega}_{U}^{-1} E_{m} \\
& \\
& \\
& \left(\varepsilon_{r, k} \mathbf{I}_{N}\left(\varepsilon_{r, 1 b d} \delta_{1}, \ldots, \varepsilon_{r, k b d} \delta_{k}, \ldots, \varepsilon_{r, N} \delta_{d d} \delta_{N}\right)^{\prime}\right) \\
& \quad=\sigma_{r}^{2} g_{k k} c_{k b d} \delta_{k},
\end{aligned}
$$

where ${ }_{b d} \mathbf{V}_{r}=\operatorname{diag}\left\{c_{k b d} \delta_{k}\right\}_{k=1}^{N}$, and $\mathbf{I}_{N}=\operatorname{diag}\{1\}_{k=1}^{N}$.
The result (4.2) shows that it is possible to define a sampling strategy which aims at controlling the anticipated variances. Indeed, if the quantities ${ }_{b d}^{a} \eta_{r, k}^{2}$ (or their proper predictions ${ }_{b d}^{a} \hat{\eta}_{r, k}^{2}$ ) are used as a replacement for the residuals ${ }_{b d} \eta_{r, k}^{2}$, the problem (3.1.1) defines a sampling design which allows one to guarantee the following conditions $\operatorname{AV}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) \leq{ }_{b d} \bar{V}_{r}(b=1, \ldots, B ; d=1, \ldots$, $\left.M_{b} ; r=1, \ldots, R\right)$.

An interesting result is the following. In the special case of a single partition, if the inclusion probabilities, $\pi_{k}$, and the heteroschedastic factors, $c_{k}$, are quite constant in each domain, then the selection of a balanced sample decreases the anticipated variance. This result is demonstrated in Falorsi and Righi (2008).

## 5. Brief extension to the case of a simple small area indirect estimator

If a given population partition defines a too large number of domains, it could happen that the budget constraints oblige to define a too large prefixed sampling errors of the direct estimators of the domains of the partition; in this situation, it could be necessary to adopt an indirect smallarea estimator, in order to control the mean square errors of partition domain estimates. Herein in the following we will show as the sampling strategy, described in sections 2 and 3, may be extended to the case of a simple small area indirect estimator. Let us consider the enough general case in which
the vector $\mathbf{x}_{k}$ of the auxiliary covariates has an intercept, such as $N_{b d}=\sum_{k \in s b d} w_{k}$.

Let $\ddot{b}$ denote the partition for which it is necessary to adopt a small area indirect estimator and let us consider the model (7.1.1) described in Rao (2005, page 116). In the herein studied context, the model for direct estimator, $\hat{b i d}^{\dot{b}} \hat{t}_{r \text {,greg }}={ }_{\dot{b} d} \hat{t}_{r, \text { greg }} / N_{\ddot{b} d}$, of the $\ddot{b} d$ domain may be defined as

$$
\begin{align*}
& \ddot{b d} d^{\hat{t}_{r, g r e g}}={ }_{\ddot{b} d} \mathbf{a}^{\prime} \boldsymbol{\varphi}_{r}+{ }_{\ddot{b} d} h_{\ddot{b} d} v_{r}+{ }_{\ddot{b} d} u_{r} \\
&\left(d=1, \ldots, M_{\ddot{b}} ; r=1, \ldots, R\right) \tag{5.1}
\end{align*}
$$

where ${ }_{\ddot{b} d} \mathbf{a}$ is a $p \times 1$ vector of area level covariates, $\boldsymbol{\varphi}_{r}$ is an unknown $p \times 1$ vector of regression coefficients, $\ddot{b i d} h$ is a known quantity related to the $\ddot{b} d^{\text {th }}$ domain, $\ddot{b} d v_{r} \sim$ $\operatorname{iid}\left(0,{ }_{\dot{b}} \sigma_{r v}^{2}\right)$ independent of the sampling error ${ }_{\dot{b} d} u_{r} \sim$ approximately $\operatorname{ind}\left(0,{ }_{\ddot{b} d} \sigma_{r \bar{t}}^{2}\right)$, being ${ }_{\ddot{b} d} \sigma_{r \bar{t}}^{2}=V_{p}\left({ }_{\ddot{b} d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}=\right.$ $t_{\mathbf{z}}$ )/ $N_{\dot{b} d}^{2}$. For known $\ddot{b} \sigma_{r v}^{2}$ and ${ }_{\ddot{b} d} \sigma_{r \bar{t}}^{2}$ values, the BLUP estimator of ${ }_{b i d} t_{r}$ is

$$
\begin{equation*}
\ddot{b} d d, \hat{t}_{r, \text { blup }}=N_{\ddot{b} d}\left(\ddot{b} d \gamma_{r} \quad \hat{b}_{d} \hat{\bar{t}}_{r} \text { greg }+\left(1-{ }_{\ddot{b} d} \gamma_{r}\right)_{\ddot{b} d} \mathbf{a}^{\prime} \hat{\boldsymbol{\varphi}}_{r}\right) \tag{5.2}
\end{equation*}
$$

being

$$
\begin{equation*}
\ddot{b} d \gamma_{r}={ }_{\ddot{b}} \sigma_{r v \ddot{b} d}^{2} h^{2} /\left({ }_{(\ddot{b} d} \sigma_{r \bar{t}}^{2}+{ }_{\ddot{b}} \sigma_{r v \ddot{b} d}^{2} h^{2}\right) \tag{5.3}
\end{equation*}
$$

and

$$
\begin{align*}
\hat{\boldsymbol{\varphi}}= & {\left[\sum_{d=1}^{M_{\dot{b}}}{ }_{\ddot{b} d} \mathbf{a}_{\ddot{b} d} \mathbf{a}^{\prime} /\left({ }_{\ddot{b} d} \sigma_{r \bar{t}}^{2}+{ }_{\dot{b}} \sigma_{r v}^{2} \ddot{b}_{d} h^{2}\right)\right]^{-1} } \\
& {\left[\sum_{l=1}^{M_{\dot{b}}} \mathbf{a}_{\dot{b} l} \hat{t}_{r, \text { greg }} /\left({ }_{\ddot{b} d} \sigma_{r \bar{t}}^{2}+{ }_{\ddot{b}} \sigma_{r v \ddot{b} d}^{2} h^{2}\right)\right] } \tag{5.4}
\end{align*}
$$

The MSE of the BLUP estimator is

$$
\begin{align*}
& \operatorname{MSE}(\ddot{b} d \\
&\left.\hat{t}_{r, \text { blup }}\right)=N_{\ddot{b} d}^{2}\left[\ddot{b} d \gamma_{r \ddot{b} d} \sigma_{r \bar{t}}^{2}+\left(1-\ddot{b}_{\dot{b} d} \gamma_{r}\right)^{2}\right.  \tag{5.5}\\
& \quad \mathbf{a}^{\prime}\left(\sum_{d=1}^{M_{\ddot{b}}}{ }_{\ddot{b} d} \mathbf{a}_{\ddot{b} d} \mathbf{a}^{\prime} /\left(\left(\ddot{b} d \sigma_{r \bar{t}}^{2}+{ }_{\ddot{b}} \sigma_{r v \ddot{b} d}^{2} h^{2}\right)\right)^{-1} \mathbf{\ddot { b } d} \mathbf{a}\right] .
\end{align*}
$$

Looking at expressions (5.3) and (5.5), it is possible to note that for a given values of the variance $\ddot{b}_{b} \sigma_{r v}^{2}$, it is possible to control the $\operatorname{MSE}\left({ }_{\dot{b} d} \hat{t}_{r \text {, blup }}\right)$ in planning the sampling design, by defining a proper value of the variance $\ddot{b}_{d} \sigma_{r \bar{t}}^{2}$. The following iterative procedure finds the $\pi_{k}^{\prime}$ inclusion probabilities which guarantee the minimum sample size and assure the respects of the following constraints: $V_{p}\left({ }_{b d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathrm{z}, h t}=t_{\mathrm{z}}\right) \leq{ }_{b d} \bar{V}_{r} \quad$ (for $\quad b \neq \ddot{b} ; d=1, \ldots, M_{b}$; $r=1, \ldots, R) \quad$ and $\quad \operatorname{MSE}\left(\ddot{\ddot{b} d}\left(\hat{t}_{r, \text { blup }}\right) \leq_{\ddot{b} d} \bar{V}_{r}\left(d=1, \ldots, M_{\ddot{b}} ;\right.\right.$ $r=1, \ldots, R)$.

Initialization: at initial iteration $(j=0)$ find the ${ }^{j} \pi_{k}^{\prime}$ inclusion probabilities, solution of the problem (3.1.1),
using the constraints $V_{p}\left({ }_{\ddot{b} d} \hat{t}_{r, \text { greg }} \mid \hat{t}_{\mathbf{z}, h t}=t_{\mathbf{z}}\right) \leq{ }_{b d} \bar{V}_{r} \quad$ (for $\left.b=1, \ldots, B ; d=1, \ldots, M_{b} ; r=1, \ldots, R\right)$.

Iteration: the generic iteration $(j=1,2, \ldots)$ is articulated as follows.

- Calculate ${ }_{\ddot{b} d}^{j} \sigma_{r \bar{t}}^{2}=\left[N /\left(N_{\dot{b} d}^{2}(N-Q)\right)\right] \sum_{k \in U}\left[\left(1 /{ }^{j-1} \pi_{k}\right)-\right.$ $1]_{\ddot{b} d} \eta_{r, k}^{2}\left(d=1, \ldots, M_{\ddot{b}} ; r=1, \ldots, R\right)$.
- Calculate $\stackrel{j}{b} d_{j} \gamma_{r}$ and ${ }^{j} \operatorname{MSE}\left({ }_{\ddot{b} d} \hat{t}_{r \text {, blup }}\right)\left(d=1, \ldots, M_{\ddot{b}}\right.$; $r=1, \ldots, R$ ) respectively by means of equation (5.3) and (5.5) by using the sampling variances $\frac{j}{b \bar{b} d} \sigma_{r \bar{t}}^{2}$ instead of ${ }_{\dot{b} d} \sigma_{r \bar{t}}^{2}$.
- Calculate ${ }_{\stackrel{b}{b} d}^{j}$ eff $_{r}={ }^{j} \operatorname{MSE}\left({ }_{\ddot{b} d} \hat{t}_{r, \text { blup }}\right) /\left({ }_{\ddot{b} d}^{j} \sigma_{r \bar{t}}^{2} N_{\dot{b} d}^{2}\right)$.
- Find the ${ }^{j} \pi_{k}^{\prime}$ inclusion probabilities, solution of the problem (3.1.1), using the ${ }_{\ddot{b} d}^{j} \eta_{r, k}^{2}={ }_{\ddot{b} d} \eta_{r, k}^{2} \stackrel{j}{\dot{b} d} \mathrm{eff}_{r}(d=$ $\left.1, \ldots, M_{\ddot{b}} ; r=1, \ldots, R ; k=1, \ldots, N\right)$ as replacement for the ${ }_{\ddot{b} d} \eta_{r, k}^{2}$ values.
Check: if the following condition is satisfied, for a small quantity $v, \sum_{k \in U}\left|{ }^{j-1} \pi_{k}-{ }^{j} \pi_{k}\right| \leq v$, then the algorithm stops and the inclusion probabilities $\pi_{k}^{\prime}$ are those calculated at iteration $j$. Otherwise, the iteration is calculated over and over again until the above condition is respected.


## 6. Empirical analysis

In order to verify the empirical properties of the proposed sampling strategy, two experiments have been implemented. Both experiments have showed good performances of the proposed strategy. The first experiment, on artificial data, is described in Falorsi and Righi (2008); the whole sampling strategy proposed in section 2 is implemented including the sampling allocation described in sections 3.1 and 3.2. The second experiment, based on a simulation on real enterprise data, is described herein in the following.

The analysis has been carried out on the 1999 population of the enterprises from 1 to 99 employees belonging to the Computer and related economic activities (2-digits of the NACE rev. 1 classification. The data base used for the simulation study has $N=10,392$ enterprises. The value added and labour cost are the variables of interest chosen in the simulation. The variable values are available for each unit in the population by an administrative data source. We consider two partitions: (DOM1) geographical region with 20 marginal domains; (DOM2) Economic activity group (3digits of the NACE rev. 1 classification with 6 different groups) by Size class (defined in terms of number of persons employed: $1=1-4 ; \quad 2=5-9 ; \quad 3=10-19 ; \quad 4=20-99)$ with 24 marginal domains. Therefore, the overall number of marginal domains is 44 , while the number of the crossclassification strata is 480 but only 360 strata have one or more population units.

In this study $n$ is set equal to to 360 . Five sampling designs have been considered, as reported in table 6.1. The first two benchmarking designs are two simple one-way stratification designs with simple random sampling without replacement in each stratum. The first design herein referred as STDOM1 is stratified by partition 1 and the second one, STDOM2, is stratified by partition 2 . The marginal sample sizes for STDOM1 have been defined by (3.3.1). The parameter $\alpha_{1}$ and the related marginal sample sizes $n_{1 d}$ $(d=1, \ldots, 20)$ guarantee the percent Coefficient of Variation (CV) of the Horvitz-Thompson estimates of totals of the auxiliary variable number of employers be lower than than $34.5 \%$ for all domain of the partition 1 . Analougsly, the parameter value $\alpha_{2}$ has been defined by means of (3.3.1), assuring that, with the STDOM2 sample design, the percent CV of the Horvitz-Thompson estimates of totals of the auxiliary variable are lower than $8.7 \%$ for all the domains of the partition 2. In the following we refer to the domains with the planned sample size greater than the sample size deriving from an allocation rule with $\alpha_{b}=1(b=1,2)$ as small domains. These domains need to be oversampled to bound the sampling errors (Marker 2001).

We note that the above allocation rules are straightforward to implement in any real survey contexts. Two balanced sample designs are examined respecting the marginal sample sizes defined by STDOM1 for the first partition and by STDOM2 for the second one: the BAL design consider the balancing equations (2.2.2) with the specification (2.3.2) of the $\mathbf{z}_{k}$ vector; the BALPOP samples satisfy (or approximately satisfy) the following balancing equations $\quad \sum_{k \in s} \pi_{k b d} \delta_{k} / \pi_{k}=n_{b d} \quad$ and $\quad \sum_{k \in s b d} \delta_{k} / \pi_{k}=$ $N_{b d}\left(b=1, \ldots, B ; d=1, \ldots, M_{b}\right)$. The probabilities $\pi_{k}$ of both designs have been obtained with the simplified procedure described in section 3.3. Furthermore, the comparison has been completed considering a coordinated design (referred as CPAR) selecting a single sample for each marginal population with Pareto Sampling (Särndal and Lundström 2005) and assuring the maximum overlap of the two samples. The marginal sample sizes, respectively defined by the STDOM1 and STDOM2 designs, are satisfied only as expectation over repeated sampling in the CPAR design; the inclusion probabilities are computed with the iterative procedure described in Falorsi et al. (2006). Five hundred Monte Carlo samples have been selected for each sampling design.

For each sample, the estimates of the domain totals have been computed by the Horvitz-Thompson (HT) estimator, modified GREG (greg) estimator and synthetic (syn) estimator, expressed as ${ }_{b d} \hat{t}_{r, \text { syn }}=\sum_{k \in U_{b d}} \tilde{y}_{r, k}$. As far as the estimators using auxiliary information are concerned, two simple homoschedastic linear models have been implemented: the model (6.1) uses 10 auxiliary variables, six of
them are the economic activity group membership indicators, and the remaining four are the size class membership indicators; the model (6.2) uses the 44 domain membership indicator variables. The linear model (6.1) is expressed by

$$
\begin{equation*}
E_{m}\left(y_{k}\right)=\beta_{h}+\beta_{j} \text { for } k \in U_{h} \cap U_{j}, \tag{6.1}
\end{equation*}
$$

where $U_{h}$ is the population of enterprises of $h^{\text {th }}(h=$ $1, \ldots, 6$ ) economic activity group and $U_{j}$ is the population of enterprises of $j^{\text {th }}(j=1, \ldots, 4)$ size class of the number of employers and $\beta_{h}$ and $\beta_{j}$ are the fixed effects of the $h^{\text {th }}$ economic activity group and of the $j^{\text {th }}$ size class.
The linear model (6.2) is

$$
\begin{equation*}
E_{m}\left(y_{k}\right)=\beta_{1 d}+\beta_{2 d} \text { for } k \in U_{1 d} \cap U_{2 d}, \tag{6.2}
\end{equation*}
$$

where $\beta_{1 d}$ and $\beta_{2 d}$ are the separate domain-specific effects.

Table 6.1
Sampling design used in the simulation study

| Sampling Design | Abbreviation |
| :--- | :---: |
| Stratified by Partition 1 with SRSWOR* in <br> each stratum | STDOM1 |
| Stratified by Partition 2 with SRSWOR* in <br> each stratum | STDOM2 |
| Balanced sampling on the marginal sample <br> sizes and on population sizes | BALPOP |
| Balanced sampling on the marginal sample <br> sizes | BAL |
| Coordinated Pareto sampling | CPAR |

*SRSWOR: Simple Random Sampling Without Replacement

We point out that the main aim of the experiment is to compare different sampling designs using the same estimator. In this context, the choice of the best model does not represent a central issue; hence, we have considered two quite general feasible models that can be implemented in all situations of planned domains. The model (6.1) is somewhat more reliable, since the estimates of the regression parameters are based on large sample sizes; while in model (6.2) it is possible to evaluate the effect of planning the domain sample sizes, although the estimates of each regression parameter are based on small sample sizes. Using the model (6.2) the syn and the greg estimators give identical results. In the following each sampling strategy is indicated in short by the couple (dis, est), where dis indicates one of the 5 sample designs referred in table 6.1 and est assumes the categories HT, syn, and greg above indicated.

In the following the analysis is based on the set of small domains. Two quality measures have been computed: the average Absolute mean Relative Bias ( $\overline{\mathrm{ARB}}$ ) and the average Relative Mean Square Error (RMSE) expressed by

$$
\begin{aligned}
& \overline{\operatorname{ARB}}_{F}(\text { dis,est })= \\
& \left.\frac{1}{\operatorname{card}(F)} \sum_{b d \in F} \left\lvert\, \frac{1}{500} \sum_{i=1}^{500}\left[{ }_{b d} \hat{t}_{r, \text { est }}^{i}(\text { dis })-{ }_{b d} t_{r}\right] /{ }_{b d} t_{r}\right. \right\rvert\, \times 100,
\end{aligned}
$$

$\overline{\operatorname{RMSE}}_{F}($ dis,est $)=$

$$
\frac{1}{\operatorname{card}(F)} \sum_{b d \in F}\left\{\frac{1}{500} \sum_{i=1}^{500}\left[{ }_{b d} \hat{t}_{r, \text { est }}^{i}(\text { dis })-{ }_{b d} t_{r}\right]^{2} /{ }_{b d} t_{r}^{2}\right\} \times 100
$$

denoting with: $F$ a specific subset of the marginal domains; $\operatorname{card}(F)$ the cardinality of $F ;{ }_{b d} \hat{t}_{r, \text { est }}^{i}($ dis $)$ the $i^{\text {th }}$ Monte Carlo sample estimate $(i=1, \ldots, 500)$ of the total ${ }_{b d} t_{r}$ in the strategy (dis, est). In particular, $F$ represents alternatively the subset of small domains of DOM1, DOM2 or the overall set of small domains (of both DOM1 and DOM2).

The Monte Carlo simulation study highlights that the multi-way stratification techniques proposed in this paper are able to take bias and variability under control with respect to two benchmark strategies (STDOM1 and STDOM2) collapsing one of the two stratification variables.

The main results of the experiment referred to the small domains set are shown in table 6.2. The table is organised in four blocks: the first one illustrates the quality measures of the HT estimator; the second and third block are dedicated respectively to the syn and greg estimators based on 10 auxiliary variables (model (6.1)); the forth block presents the results of syn or greg estimators based on the 44 domain membership indicator variables (model (6.2)). We restrict the comments only on the value added variable, but similar consideration could be expressed for the labour cost variable. In general, the comments are referred to the overall set of small domains.

Examining firstly the HT estimator, we observe the following.

- The two benchmark designs (STDOM1 and STDOM2) have an $\overline{\mathrm{RMSE}}$ value for the unplanned domains equal to $148.28 \%$ and $107.49 \%$ respectively. These values cause the large $\overline{\mathrm{RMSE}}$ values computed for the overall set of small domains and respectively equal to $102.74 \%$ and 55.23\%.
- The STDOM2 shows better results than those attained by STDOM1. This finding is explained by the fact that the STDOM2 stratification criterion is correlated with the variables of interest and takes under control a larger number of small domains than the STDOM1 stratification.
- As far as the overall set of small domains, the BALPOP is the more efficient design, both in terms of $\overline{\text { ARB }}$ ( $1.06 \%$ ) and $\overline{\mathrm{RMSE}}$ ( $32.58 \%$ ), even if BAL is only slightly worse.
- The strategy adopting the coordinated sampling shows worse values with respect to balanced sampling but it performs better in terms of $\overline{\text { RMSE }}$ than benchmark strategies.
Considering the synthetic estimator based on 10 auxiliary variables, some issues may be pointed out.
- All designs are characterized by a large bias. The STDOM1 has an $\overline{\mathrm{ARB}}$ equal to $13.99 \%$ (although it has an unacceptable $\overline{\mathrm{RMSE}}$ that is equal to $65.16 \%$ ). The rest of the designs have the $\overline{\mathrm{ARB}}$ values higher than $18 \%$. This evidence gives a warning against the use of synthetic estimator.
. The STDOM2 design has the lowest $\overline{\operatorname{RMSE}}$ (26.16\%), because of a strong reduction of the DOM1 variance. However, the $\overline{\mathrm{ARB}}$ value (20.34\%) is the largest than all designs.
- The behaviour of balanced and coordinated designs in terms of bias and variance are more or less equal. The BAL has the lowest $\overline{\mathrm{ARB}}(18.33 \%)$ and $\overline{\mathrm{RMSE}}$ (31.61\%) values.

The experimental results of the greg estimator suggest some considerations.

- All the designs show strong improvements of the quality measures. In general, the $\overline{\mathrm{ARB}}$ measure has a remarkable reduction with respect to the same indicator computed on the synthetic estimator. Only the STDOM1 still presents a high $\overline{\text { ARB }}$ value (7.40\%).
. In the STDOM2, the reduction of the bias is more than compensated from the increase of the variability. This produces an $\overline{\mathrm{RMSE}}$ equal to $34.05 \%$.
- Both the balanced and the coordinated designs have good performances, though the balanced designs are slightly better being the $\overline{\mathrm{RMSE}}$ roughly equal to the $23 \%$.

Finally in the fourth block we note that the syn or greg estimator based on 44 auxiliary variables show analogous results to those of the greg estimator based on 10 auxiliary variables. The balanced designs are the best with slight preference for the BALPOP sampling.

As general findings, the balanced designs seem to guarantee a good strategy to take under control bias and variance of the overall set of the small domains.

The conclusion is that for all blocks, BALPOP generally shows the best overall performance with respect to bias and accuracy. The strategy based on the BALPOP sample design coupled with the greg estimator with the ten auxiliary variables (block 3) is a safe choice for both value added and labour cost. The BAL design performs well too. Moreover, the results show that the synthetic estimator of block 2 must be considered carefully because the bias can be unexpectedly large and the squared bias would be the dominating part of the RMSE.

Table 6.2
Average Absolute Relative Bias ( $\overline{\mathrm{ARB}}$ ) and Relative Mean Square Error ( $\overline{\mathrm{RMSE}}$ ) of small domain sampling strategies

| Sampling Design | Value Added |  |  |  |  |  | Labour Cost |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\overline{\text { ARB }}$ | $\overline{\text { RMSE }}$ | $\overline{\text { ARB }}$ | $\overline{\text { RMSE }}$ | $\overline{\text { ARB }}$ | $\overline{\text { RMSE }}$ | $\overline{\text { ARB }}$ | $\overline{\text { RMSE }}$ | $\overline{\text { ARB }}$ | $\overline{\text { RMSE }}$ | $\overline{\text { ARB }}$ | $\overline{\text { RMSE }}$ |
| Horvitz-Thompson estimator (block 1) |  |  |  |  |  |  |  |  |  |  |  |  |
| STDOM1 | 1.79 | 43.19 | 8.18 | 148.28 | 5.41 | 102.74 | 1.72 | 42.82 | 6.86 | 155.87 | 4.63 | 106.88 |
| STDOM2 | 3.42 | 107.49 | 0.47 | 15.26 | 1.75 | 55.23 | 3.32 | 105.66 | 0.46 | 12.66 | 1.70 | 52.96 |
| BALPOP | 0.77 | 24.86 | 1.29 | 38.49 | 1.06 | 32.58 | 0.74 | 23.60 | 1.20 | 34.26 | 1.00 | 29.64 |
| BAL | 0.84 | 25.43 | 1.45 | 40.61 | 1.19 | 34.03 | 0.79 | 24.22 | 1.57 | 35.80 | 1.23 | 30.78 |
| CPAR | 1.35 | 32.52 | 2.18 | 53.85 | 2.18 | 44.60 | 1.44 | 31.68 | 2.62 | 51.44 | 2.11 | 42.88 |
| Synthetic estimator with 10 auxiliary variables (block 2) |  |  |  |  |  |  |  |  |  |  |  |  |
| STDOM1 | 14.22 | 18.88 | 13.81 | 100.55 | 13.99 | 65.16 | 12.29 | 18.40 | 9.25 | 95.03 | 10.57 | 61.83 |
| STDOM2 | 24.82 | 33.96 | 14.48 | 15.96 | 20.34 | 26.16 | 13.13 | 14.79 | 12.46 | 23.11 | 12.75 | 19.51 |
| BALPOP | 13.68 | 17.51 | 24.98 | 43.98 | 20.09 | 32.51 | 11.89 | 15.60 | 12.35 | 33.08 | 12.15 | 25.50 |
| BAL | 14.92 | 18.46 | 21.82 | 41.66 | 18.83 | 31.61 | 13.37 | 16.91 | 10.41 | 32.64 | 11.69 | 25.82 |
| CPAR | 13.68 | 17.83 | 23.45 | 44.63 | 19.22 | 33.02 | 11.82 | 16.13 | 11.69 | 34.93 | 11.75 | 26.78 |
| Modified GREG estimator with 10 auxiliary variables (block 3) |  |  |  |  |  |  |  |  |  |  |  |  |
| STDOM1 | 2.35 | 30.13 | 11.26 | 119.95 | 7.40 | 81.03 | 1.86 | 29.28 | 11.79 | 119.23 | 7.49 | 80.25 |
| STDOM2 | 3.98 | 58.62 | 0.95 | 15.26 | 2.26 | 34.05 | 2.90 | 52.66 | 0.93 | 12.66 | 1.78 | 29.99 |
| BALPOP | 1.11 | 19.41 | 2.20 | 25.80 | 1.73 | 23.03 | 1.01 | 16.42 | 1.99 | 21.73 | 1.57 | 19.43 |
| BAL | 1.63 | 19.41 | 1.76 | 26.11 | 1.70 | 23.21 | 1.21 | 16.72 | 2.08 | 21.96 | 1.70 | 19.69 |
| CPAR | 1.04 | 21.27 | 1.63 | 29.30 | 1.37 | 25.82 | 1.03 | 18.27 | 1.11 | 24.60 | 1.08 | 21.86 |
| Synthetic or Modified GREG estimator with 44 auxiliary variables (block 4) |  |  |  |  |  |  |  |  |  |  |  |  |
| STDOM1 | 3.39 | 31.30 | 27.48 | 63.22 | 17.04 | 49.39 | 2.76 | 30.80 | 28.67 | 63.05 | 17.44 | 49.08 |
| STDOM2 | 17.24 | 102.24 | 1.37 | 20.65 | 8.25 | 56.00 | 23.00 | 102.64 | 1.42 | 19.10 | 10.77 | 55.30 |
| BALPOP | 1.07 | 20.71 | 1.97 | 26.98 | 1.58 | 24.26 | 1.08 | 17.62 | 1.93 | 24.07 | 1.56 | 21.27 |
| BAL | 1.47 | 20.36 | 2.13 | 28.46 | 1.84 | 24.95 | 1.41 | 17.66 | 2.02 | 25.10 | 1.75 | 21.88 |
| CPAR | 1.79 | 23.38 | 2.22 | 32.39 | 2.03 | 28.48 | 1.65 | 20.73 | 2.08 | 30.39 | 1.90 | 26.21 |

## 7. Conclusions

This work illustrates an efficient sampling strategy useful for obtaining planned sample size for domains belonging to different partitions of the population and in order to guarantee that sampling errors of domain estimates are lower than given thresholds. The sampling strategy, that covers the multivariate-multi-domain case, is useful when the overall sample size is bounded. In these instances the standard solution, using a stratified sample with the strata given by the cross-classification of variables defining the different partitions, is not feasible since the number of strata is larger than the overall sample size.

The sampling strategy which is proposed is based on the use of the balanced sampling selection technique and on a GREG-type estimator. The proposal may be easily extended to a strategy employing the use of both direct and indirect small area estimators.

The easy feasibility is one of the main advantages of the proposed solution since it is implemented by algorithms that
are either based on free software tools or suitable for automated spreadsheets. But some other interesting aspects seem to appear.

The empirical analysis of real enterprise data shows good performances of the proposed strategy, which seems to be robust even when departing from ideal conditions (i.e., the estimates appear to be of high quality even when the inclusion probabilities of the sample differ from the optimal ones). These results encourage additional work to give a systematic account of conditions under which the proposed method will have good performance.

Furthermore, the proposed strategy does seems to work well for large datasets, in terms of computer time, and therefore it seems to be suitable for large scale surveys.

Finally, the approach represents an original overall small area sampling strategy, which jointly considers the sampling design and the estimator. The paper deeply analyzes the design issues, but more research is needed to study more carefully the estimation issues. In particular, future research should be focused on the improvement of the model-based
or model-assisted estimators due to the presence of sample units in each estimation domain, allowing the use of models with specific small area effects and giving more accurate estimates of the parameters of interest at small area level. These aspects seem to be an appealing strength to be investigated.

## References

Bethel, J. (1989). Sample allocation in multivariate surveys. Survey Methodology, 15, 47-57.

Bishop, Y., Fienberg, S. and Holland, P. (1975). Discrete Multivariate Analysis. MIT Press, Cambridge, MA.

Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. Computational Statistics, 21, 53-62.

Chromy, J. (1987). Design optimization with multiple objectives. Proceedings of the Survey Research Methods Section, American Statistical Association, 194-199.

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley \& Sons, Inc.

Deville, J.-C., and Tillé, Y. (2000). Selection a several unequal probability samples from the same population. Journal of Statistical Planning and Inference, 86, 89-101.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. Biometrika, 91, 893-912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. Journal of Statistical Planning and Inference, 128, 569-591.

Dykstra, R., and Wollan, P. (1987). Finding I-projections subject to a finite set of linear inequality constraints. Applied Statistics, 36, 377-383.

Falorsi, P.D., Orsini, D. and Righi, P. (2006). Balanced and coordinated sampling designs for small domain estimation. Statistics in Transition, 7, 1173-1198.

Falorsi, P.D., and Righi, P. (2008). An efficient multi-way design strategy for domain estimation. Contributi ISTAT (downloadable on http://www.istat.it/dati/pubbsci/contributi/).

Jessen, R.J. (1970). Probability sampling with marginal constraints. Journal American Statistical Society, 65, 776-795.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. Survey Methodology, 24, 51-55.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for Domains, including small domains. Survey Methodology, 1, 33-44.

Lu, W., and Sitter, R.R. (2002). Multi-way stratification by linear programming made practical. Survey Methodology, 2, 199-207.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. Journal of Official Statistics, 15, 1-24.

Marker, D.A. (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. Survey Methodology, 27, 183-188.

Montanari, G.E., and Ranalli, M.G. (2003). Nonparametric methods in survey sampling. In New Developments in Classification and Data Analysis, (Eds., M. Vinci, P. Monari, S. Mignani, A. Montanari), Springer, Berlin.

Rao, J.N.K. (2003). Small Area Estimation. New York: John Wiley \& Sons, Inc.

Särndal, C.-E., and Lundström, S. (2005). Estimation in Surveys with Nonresponse. New York: Springer-Verlag.

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

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. Survey Methodology, 20, 3-22.

Tillé, Y., and Favre, A.C. (2005). Optimal allocation in balanced sampling. Statistics and Probability Letters, 74, 31-37.
Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: John Wiley \& Sons, Inc.

# Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries 

Danny Pfeffermann, Bénédicte Terryn and Fernando A.S. Moura ${ }^{1}$


#### Abstract

This paper considers situations where the target response value is either zero or an observation from a continuous distribution. A typical example analyzed in the paper is the assessment of literacy proficiency with the possible outcome being either zero, indicating illiteracy, or a positive score measuring the level of literacy. Our interest is in how to obtain valid estimates of the average response, or the proportion of positive responses in small areas, for which only small samples or no samples are available. As in other small area estimation problems, the small sample sizes in at least some of the sampled areas and/or the existence of nonsampled areas requires the use of model based methods. Available methods, however, are not suitable for this kind of data because of the mixed distribution of the responses, having a large peak at zero, juxtaposed to a continuous distribution for the rest of the responses. We develop, therefore, a suitable two-part random effects model and show how to fit the model and assess its goodness of fit, and how to compute the small area estimators of interest and measure their precision. The proposed method is illustrated using simulated data and data obtained from a literacy survey conducted in Cambodia.


Key Words: Credibility intervals; Generalized linear mixed model; Goodness of fit; Linear mixed model; MCMC; Prediction bias; Prediction MSE.

## 1. Introduction

In this paper we consider situations where the target response value is either zero or an observation from a continuous distribution. A typical example analyzed in the paper is the assessment of literacy proficiency based on a written test with the possible outcome being either zero, indicating illiteracy, or a positive score in a given range measuring the level of literacy. Another example is the consumption of illicit drugs (or certain food items), where a zero value indicates "no consumption", whereas a positive outcome measures the amount consumed. Our interest lies in how to obtain valid estimates of the average response (average literacy level in our example), or the proportion of positive responses (proportion of literate people), in small areas for which only small samples or no samples are available. As in other small area estimation problems, the small sample sizes within the sampled areas and the existence of nonsampled areas requires the use of model based methods.

We propose the use of a two-part random effects model and show how to fit the model and assess its goodness of fit, and how to obtain the small area estimates of interest and measure their precision. The first part of the model specifies the probability of a zero score. The second part specifies the distribution of the positive scores. Although the model is not new and is used in other applications, (see, e.g., Olsen and Schafer 2001 and the discussion and references in that paper), to the best of our knowledge this kind of mixed distribution has not been considered before in the small area
estimation literature. Notice that the zero scores in our application are 'structural' (true) zeroes. There exists a related body of literature that handles excess of zeros in count data, which may arise from a combination of overdispersion or true zero inflation. Zero inflated data are data that have a larger proportion of zeros than expected from pure count (Poisson) data. See, e.g., Barry and Welsh (2002).

The first part of our model is the logistic function, used to model the probability of a positive score. The second part is a linear model with normal error terms fitted to the non-zero responses. Both models include individual and area level covariates, as well as area random effects that account for variations not explained by the covariates. The model allows for correlations between the corresponding random effects of the two parts and is fitted by application of Markov Chain Monte Carlo (MCMC) simulations.

The two-part model is fitted to data collected as part of the national literacy household survey carried out in Cambodia in 1999, known as the 'Assessment of the Functional Literacy Levels of the Adult Population'. Figure 1 displays the histogram of the literacy scores observed for this survey. In this application we produce small area estimates for districts of residence and nested villages, requiring the use of a two-part three-level random effects model. We assess the goodness of fit of the model by use of simple descriptive statistics and by simulating data from the model. The use of simulations enables also to compare the results of fitting the 'full' two-part model with results obtained by fitting the two parts of the model separately,

[^12]without accounting for the correlations between the random effects in the two parts. Another comparison of interest is to results obtained when ignoring the special nature of the data and fitting the linear part to all the responses, ignoring the existence of many zero scores.

In order to facilitate the presentation and discussion in the rest of the paper, we consider literacy scores measured for individuals residing in villages nested in districts, but as noted above, the model considered in this paper can be used for many other important applications.


Figure 1 Histogram of literacy scores in the national literacy survey in Cambodia, 1999

## 2. Model and small area predictors

### 2.1 The two-part model

Let $y$ define the response (literacy test score in our application) and $R$ the covariate variables and random effects. Then,

$$
\begin{align*}
E(y \mid R=r)= & E(y \mid R=r, y=0) \operatorname{Pr}(y=0 \mid R=r) \\
& +E(y \mid R=r, y>0) \operatorname{Pr}(y>0 \mid R=r) \\
= & E(y \mid R=r, y>0) \operatorname{Pr}(y>0 \mid R=r) \tag{1}
\end{align*}
$$

For the small area estimation problem considered in this paper we apply a nested three-level model with districts of residence defining the first level, villages defining the second level and individuals defining the third level. For individual $k$ residing in village $j$ of district $i$, with covariates and random effects $R_{i j k}=r$, we have therefore the relationship,

$$
\begin{align*}
& E\left(y_{i j k} \mid R_{i j k}=r\right)= \\
& \quad E\left(y_{i j k} \mid R_{i j k}=r, y_{i j k}>0\right) \operatorname{Pr}\left(y_{i j k}>0 \mid R_{i j k}=r\right) . \tag{2}
\end{align*}
$$

In what follows we model the two parts in the right hand side of (2). For individuals with positive responses we assume the 'linear mixed model',

$$
\begin{align*}
y_{i j k} & =x_{i j k}^{\prime} \beta+u_{i}+v_{i j}+\varepsilon_{i j k} \\
u_{i} & \sim N\left(0, \sigma_{u}^{2}\right) ; v_{i j} \sim N\left(0, \sigma_{v}^{2}\right) ; \varepsilon_{i j k} \sim N\left(0, \sigma_{\varepsilon}^{2}\right), \tag{3}
\end{align*}
$$

where $x_{i j k}$ represents individual and area level values of covariates, $u_{i}$ is a random district effect and $v_{i j}$ is a nested random village effect. The random effects $u_{i}$ and $v_{i j}$, and the residual terms $\varepsilon_{i j k}$ are assumed to be mutually independent between and within the districts and villages. They account for the variation of the individual scores not explained by the covariates, and define the correlations between the scores of individuals residing in the same village and the correlations between the scores of individuals residing in the same district but in different villages.

$$
\begin{align*}
& \operatorname{Corr}\left(y_{i j k}, y_{i i^{\prime} k^{\prime}}\right)= \\
& \begin{cases}\left(\sigma_{u}^{2}+\sigma_{v}^{2}\right) /\left(\sigma_{u}^{2}+\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right) & \text { if } i=i^{\prime}, j=j^{\prime}, k \neq k^{\prime} \\
\sigma_{u}^{2} /\left(\sigma_{u}^{2}+\sigma_{v}^{2}+\sigma_{\varepsilon}^{2}\right) & \text { if } i=i^{\prime}, j \neq j^{\prime} \\
0 & \text { if } i \neq i^{\prime}\end{cases} \tag{4}
\end{align*} .
$$

For the probabilities of positive responses (the second part of Equation (2)), we assume the 'generalized linear mixed model',

$$
\begin{array}{r}
p_{i j k}=\operatorname{Pr}\left(y_{i j k}>0 \mid x_{i j k}^{*}, u_{i}^{*}, v_{i j}^{*}\right)=\frac{\exp \left(x_{i j k}^{\prime *} \gamma+u_{i}^{*}+v_{i j}^{*}\right)}{1+\exp \left(x_{i j k}^{\prime *} \gamma+u_{i}^{*}+v_{i j}^{*}\right)} \\
u_{i}^{*} \sim N\left(0, \sigma_{u^{*}}^{2}\right) ; v_{i j}^{*} \sim N\left(0, \sigma_{v^{*}}^{2}\right) \tag{5}
\end{array}
$$

implying that $\operatorname{logit}\left(p_{i j k}\right)=\log \left(p_{i j k} /\left(1-p_{i j k}\right)\right)=x_{i j k}^{\prime *} \gamma+$ $u_{i}^{*}+v_{i j}^{*}$. Here again, $u_{i}^{*}$ and $v_{i j}^{*}$ represent independent random district and village effects not accounted for by the covariates $x_{i j k}^{*}$. Notice that the covariates $x_{i j k}$ in Equation (3) and the covariates $x_{i j k}^{*}$ in Equation (5) may differ, see the empirical study in Section 4.

Remark 1. One could argue that the mixed linear model (3) with the added normality assumptions implies a corresponding probit model for the probabilities $p_{i j k}$. This, however, is not true since the model (3) is only assumed for the positive scores. It follows that the parameters of the two models can be assumed to be distinct in the sense of Rubin (1976).

We allow for nonzero correlations between the district random effects in the two parts, and similarly for the village random effects. This is a reasonable assumption since it can be expected that for given values of the covariates, an individual residing in an area characterized by high literacy scores will have a higher probability of a positive score than an individual residing in an area with low scores. The magnitude of these correlations and the importance of
accounting for them when fitting the model depends on the prediction power of the covariates available for the two parts of the model, or alternatively, on the variances of the random effects, (the higher the prediction power of the covariates, the lower are the variances). The correlations are modelled by assuming,

$$
\begin{equation*}
u_{i}^{*}\left|u_{i} \sim N\left(K_{u} u_{i}, \sigma_{u \mid u}^{2}\right) ; v_{i j}^{*}\right| v_{i j} \sim N\left(K_{v} v_{i j}, \sigma_{v \mid v}^{2}\right) . \tag{6}
\end{equation*}
$$

Figure 2 provides supporting evidence for this proposition using the sample data from the center of Cambodia that is used for the empirical study in Section 4. (The empirical correlations between the variables measured on the two axes are 0.25 for villages and 0.38 for districts.)

### 2.2 Parameters of interest and predictors

For village $(i, j)$ of size $N_{i j}$, the small area parameters of interest are the true mean of the literacy scores, $\bar{Y}_{i j}=$ $\sum_{\substack{k=1 \\ N_{i} \\ N_{i j}}}^{i j k}$ / $N_{i j}$, and the proportion of positive scores, $P_{i j}=$ $\sum_{k=1}^{N_{i j}} \mathrm{I}\left(y_{i j k}>0\right) / N_{i j}$, where $\mathrm{I}\left(y_{i j k}>0\right)=1$ if $y_{i j k}>0$ and is 0 otherwise. Notice that the means are computed over all the individuals, including individuals with zero scores.

Under the model (2), the mean is predicted as,

$$
\begin{equation*}
\hat{\bar{Y}}_{i j}=\left[\sum_{k \in S_{i j}} y_{i j k}+\sum_{k \notin S_{i j}} \hat{y}_{i j k}\right] / N_{i j} \tag{7}
\end{equation*}
$$

where $S_{i j}$ defines the sample from village $(i, j)$. By (3) and (5), the missing scores can be predicted under the frequentist approach as,

$$
\begin{equation*}
\hat{y}_{i j k}=\frac{\exp \left(x_{i j k}^{\prime *} \hat{\gamma}+\hat{u}_{i}^{*}+\hat{v}_{i j}^{*}\right)}{1+\exp \left(x_{i j k}^{\prime *} \hat{\gamma}+\hat{u}_{i}^{*}+\hat{v}_{i j}^{*}\right)} \times\left[x_{i j k}^{\prime} \hat{\beta}+\hat{u}_{i}+\hat{v}_{i j}\right] \tag{8}
\end{equation*}
$$

where $\hat{\beta}, \hat{\gamma}, \hat{u}_{i}, \hat{v}_{i}, \hat{u}_{i}^{*}, \hat{v}_{i j}^{*}$ define appropriate sample estimates, see next section. One could add an estimate $\hat{\varepsilon}_{i j k}$ to the estimated mean, $\left(x_{i j k}^{\prime} \hat{\beta}+\hat{u}_{i}+\hat{v}_{i j}\right)$, obtained either by drawing from the $N\left(0, \hat{\sigma}_{\varepsilon}^{2}\right)$ distribution, or by selecting at
random an estimated residual, $\hat{\varepsilon}_{i^{\prime}, j^{\prime} k^{\prime}}=\left(y_{i^{\prime}, j^{\prime}, k^{\prime}}-x_{i j^{\prime} k^{\prime}}^{\prime} \hat{\beta}-\right.$ $\hat{u}_{i^{\prime}}-\hat{v}_{i j^{\prime}}$ ) from the estimated residuals computed for the sampled individuals. Adding estimates $\hat{\varepsilon}_{i j k}$ to the estimated mean values reflects more closely the variability of the positive responses. Under the Bayesian approach, the missing scores are predicted by drawing at random from their predictive distribution, see next section.

By (5), the proportion $P_{i j}$ is predicted under the frequentist approach as,

$$
\begin{equation*}
\hat{P}_{i j}=\frac{1}{N_{i j}}\left[\sum_{k \in S_{i j}} \mathrm{I}\left(y_{i j k}>0\right)+\sum_{k \notin S_{i j}} \hat{\mathrm{I}}\left(y_{i j k}>0\right)\right], \tag{9}
\end{equation*}
$$

where

$$
\hat{\mathrm{I}}\left(y_{i j k}>0\right)=\frac{\exp \left(x_{i j k}^{\prime *} \hat{\gamma}+\hat{u}_{i}^{*}+\hat{v}_{i j}^{*}\right)}{1+\exp \left(x_{i j k}^{\prime *} \hat{\gamma}+\hat{u}_{i}^{*}+\hat{v}_{i j}^{*}\right)} .
$$

A Bayesian solution consists of predicting the indicators $\mathrm{I}\left(y_{i j k}>0\right)$ by drawing at random from their predictive distribution.

The district means and proportions are predicted analogously, which is the same as computing the weighted average of the corresponding village predictors, with the weights defined by the relative village sizes.
Remark 2. The computation of the predictor defined by (7) and (8) requires knowledge of the covariates $\mathrm{x}_{i j k}, \mathrm{x}_{i j k}^{*}$ for every unit in the population. Similarly, the computation of the predictor in (9) requires knowledge of the covariates $x_{i j k}^{*}$ for every unit in the population. This is generally true for all generalized linear mixed models. Information on the auxiliary covariate variables is often obtained from censuses or other administrative records. In the absence of such information, the missing covariates can be imputed by drawing at random from their estimated parametric distribution or empirical distribution.


Figure 2 Proportion of positive scores by average of positive scores for districts and villages in center of Cambodia. National literacy survey, 1999

## 3. Inference

The use of the small area predictors defined by (7)-(9) requires estimating the fixed parameters (hyperparameters) $\left(\beta, \sigma_{u}^{2}, \sigma_{v}^{2}, \sigma_{\varepsilon}^{2}\right)$ of the linear part (Equation 3), the fixed parameters $\left(\gamma, K_{u}, K_{v}, \sigma_{u^{\mu}|u|}^{2}, \sigma_{v \mid v}^{2}\right)$ of the logistic part (Equations 5, 6), and predicting the random effects $\left\{\left(u_{i}, v_{i j} ; u_{i}^{*}, v_{i j}^{*}\right)\right\}$. Methods for estimating fixed and random effects when fitting linear mixed models, or generalized linear mixed models alone, have been developed over the last two decades under both the frequentist and the Bayesian paradigms. The use of these methods permits also the computation of estimators of the mean square error (MSE) or the Bayes risk of the small area predictors that account for hyper parameter estimation to correct order. See the book by Rao (2003) and the more recent article by Jiang and Lahiri (2005) for thorough reviews and discussions. However, the two-part model defined by (2)-(6) has not been considered in the small area literature, and in what follows we consider a few possibilities of fitting this model.

### 3.1 Full likelihood based inference

Define, $\mathrm{I}_{i j k_{*}}=1(0)$ if $Y_{i j k}>0(=0)$ and denote, $r_{i j k}=$ $\left(x_{i j k}, u_{i}, v_{i j}\right), r_{i j k}^{*}=\left(x_{i j k}^{*}, u_{i}^{*}, v_{i j}^{*}\right)$. For given vectors $r_{i j k}, r_{i j k}^{*}$, the likelihood for the two-part model takes the form,

$$
\begin{equation*}
L=\prod_{i, j, k \in s}\left(p_{i j k}\right)^{\mathrm{I}_{j / k}}\left[f\left(y_{i j k} \mid r_{i j k}, y_{i j k}>0\right)\right]^{\mathrm{I}, k}\left(1-p_{i j k}\right)^{\left(1-\mathrm{I}_{i j k}\right)}, \tag{10}
\end{equation*}
$$

where $s=\cup s_{i j}$ denotes the sample from all the villages, $p_{i j k}$ is defined by (5) and $f\left(y_{i j k} \mid r_{i j k}, y_{i j k}>0\right)$ is the normal density with mean $\left(x_{i j k}^{\prime} \beta+u_{i}+v_{i j}\right)$ and variance $\sigma_{\varepsilon}^{2}$ (Equation 3). The use of this likelihood for inference is, however, problematic because the random effects $\left\{\left(u_{i}, v_{i j}\right.\right.$; $\left.\left.u_{i}^{*}, v_{i j}^{*}\right)\right\}$ are in fact unobservable. One possibility, therefore, is to integrate the likelihood over the joint (normal) distribution of the random effects as defined by (3) (5) and (6), and maximize the integrated likelihood with respect to the fixed (hyper) parameters $\left(\beta, \sigma_{u}^{2}, \sigma_{v}^{2}, \sigma_{\varepsilon}^{2}\right)$ and ( $\gamma, K_{u}$, $\left.K_{v}, \sigma_{u \mid u}^{2}, \sigma_{v v^{*} v}^{2}\right)$. Having estimated the fixed parameters, the random effects can be predicted by their expected values given the data (with the maximum likelihood estimates held fixed), which requires another set of integrations. Olsen and Schafer (2001) consider a two-part model for fitting longitudinal data and approximate the integrated likelihood by a high order multivariate Laplace approximation (Raudenbush, Yang and Yosef 2000). The authors calculate empirical Bayes predictors of the random effects by use of importance sampling (Tanner 1996), setting the fixed parameters at their maximum likelihood estimates. The application of this procedure, however, is very complicated computationally, and the mean square estimators of the errors (MSE) of the small area predictors obtained this way
fail to account for the variation induced by estimating the fixed parameters. The contribution to the total MSE from estimating the fixed parameters can not be ignored in general, unless the numbers of sampled districts and villages are very large.

### 3.2 Separate model fitting

The idea here is to fit the two parts of the model separately, and then combine the estimates for computing the small area predictor defined by (7) and (8). The predictor in (9) is obtained directly from fitting the second part only. As mentioned earlier, the fitting of the separate parts has been studied extensively in the literature and computer softwares are readily available, particularly for linear mixed models. It is important to note in this regard that under the present two-part model, the predictors (7)-(9) are nonlinear functions of the data and even when the hyper parameters are known, no explicit formulae are available for the prediction MSEs. Estimating the MSE under the frequentist approach with bias of small order requires therefore developing new appropriate approximations or resampling procedures, which in the case of the predictor $\hat{\bar{Y}}_{i j}$ defined by (7) and (8), account for the correlations between the data in the two parts. This is further complicated by the fact that by fitting the two parts separately, it is not clear how to estimate the coefficients $\left(K_{u}, K_{v}\right)$ defining the correlations between the random effects in the two parts (Equation 6). A Jackknife procedure for estimating the prediction MSE of the predictor in (9) under separate model fitting has been developed by Jiang, Lahiri and Wan (2002). Bootstrap estimators applicable to this predictor, again under separate model fitting, are studied in Hall and Maiti (2006).

### 3.3 Bayesian inference under the two-part model

The use of Bayesian methods requires specification of prior distributions for the fixed parameters underlying the two-part model (the coefficients $\beta, \gamma, K_{u}, K_{v}$ and the variances $\left.\sigma_{u}^{2}, \sigma_{v}^{2}, \sigma_{\varepsilon}^{2}, \sigma_{u|u|}^{2}, \sigma_{v_{\|}^{2} v}^{2}\right)$, but with the aid of Markov Chain Monte Carlo (MCMC) simulations, the application of this approach permits sampling from the posterior distribution of the fixed parameters and the random effects, and hence sampling from the predictive distribution of the unobserved responses. Thus, the use of this approach yields the whole posterior distribution of the small area parameters of interest, allowing thereby the computation of correct MSE (posterior variance) measures or confidence (credibility) intervals that account for all the sources of variation. As discussed above, estimation of the prediction MSE under the previous approaches is problematic, particularly with regard to the predictor $\hat{\bar{Y}}_{i j}$ defined by (7) and (8). Computer software is available to
perform all the necessary computations but it should be noted that with complex models, the computations can be intensive and time consuming.

In the empirical study of this article we followed the Bayesian approach using the WinBUGS software (Spiegelhalter, Thomas and Best 2003). This software is known to be "user friendly", and based on our past experience it operates very well. Clearly, there are many other software available for MCMC simulations, such as MLwiN (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse, Draper, Langford and Lewis 2002) or R (Development Core Team 2008).

WinBUGS implements the MCMC algorithm with the Gibbs sampler (Gelfand and Smith 1990). The Gibbs sampler samples alternately from the conditional distribution of each of the fixed and random parameters (random effects), given the data and the remaining parameters. It defines a Markov chain, which under some regularity conditions converges to a realization from the joint posterior distribution of all the model parameters. Thus, at the end of the sampling process (upon convergence), the algorithm produces a (single) realization of each of the fixed and random parameters from their joint posterior distribution given the data. The realizations are denoted below by a tilde above the symbols. Realizations $\tilde{y}_{i j k}$ from the posterior distribution of $y_{i j k}$ are obtained by randomly drawing $\tilde{I}_{i j k}=1$ (or 0 ) with probabilities $\tilde{p}_{i j k}\left(\right.$ or $\left.1-\tilde{p}_{i j k}\right) ; \tilde{p}_{i j k}=\exp \left(x_{i j k}^{* *} \tilde{\gamma}+\tilde{u}_{i}^{*}+\tilde{v}_{i j}^{*}\right) \times$ $\left[1+\exp \left(x_{i j k}^{\prime *} \tilde{\gamma}+\tilde{u}_{i}^{*}+\tilde{v}_{i j}^{*}\right)\right]^{-1}$, and defining,

$$
\begin{equation*}
\tilde{y}_{i j k}=\left(x_{i j k}^{\prime} \tilde{\beta}+\tilde{u}_{i}+\tilde{v}_{i j}+\tilde{\varepsilon}_{i j k}\right) \times \tilde{I}_{i j k} . \tag{11}
\end{equation*}
$$

Substituting $\tilde{y}_{i j k}$ for $\hat{y}_{i j k}$ in (7) and $\tilde{I}_{i j k}$ for $\hat{I}_{i j k}$ in (9) yields a single sampled value of the mean $\bar{Y}_{i j}$ and the proportion $P_{i j}$ from their respective posterior distributions, for every village ( $i, j$ ). Repeating the same process independently a large number of times (using parallel chains, see below) yields an empirical approximation to the posterior distribution of the mean and the proportion. The true village means are then predicted by averaging the corresponding sampled values in all the chains and similarly for the village proportions. The MSE (Bayes risk) is estimated by computing the empirical variance of the sampled values. Credibility (confidence) intervals with coverage rates of $(1-\alpha)$ are defined by the $\alpha / 2$ and $(1-\alpha / 2)$ level quantiles of the empirical posterior distribution. The same procedure is applied for predicting the district means and proportions, and the corresponding for computing prediction variance and credibility intervals.

In practice, the use of parallel chains for producing independent realizations from the posterior distributions is often too time consuming, in which case the samples can be generated from a single long chain or a few chains, but selecting only every $\mathrm{r}^{\text {th }}$ sampled value (after convergence),
thus reducing as much as possible the correlations between adjacent sampled values.

## 4. Empirical results

### 4.1 Data and model

We use data from the 1999 survey, 'Assessment of the Functional Literacy Levels of the Adult Population’ in Cambodia for the empirical illustrations. This is a household survey, interviewing 6,548 adults and administering a literacy test consisting of 20 tasks in the Khmer language, with scores ranging from 0 to 100 (see Figure 1 in the introduction). The survey used a stratified multi-stage sampling design with the strata defined by the 24 provinces that comprise the country. Each of the provinces is divided into districts, and about half of them were selected to the sample (a total of 96 districts out of the 184 districts in the country). Two communes were sampled from each of the selected districts and other than in a few cases, three villages were selected from each of the sampled communes. Finally, households were sampled in each village and one adult selected from each household, alternating according to age and sex. The sampling design at each stage was systematic sampling. The number of households selected in each village was the same for all the villages belonging to the same province. The total province sample sizes were allocated proportionally to the province population sizes.

The small areas of interest are the districts and villages. In the present study we restrict to the 50 rural districts sampled in provinces located in the center of the country, for which the same model is expected to hold. In these 50 districts 5 districts had samples of 20 adults or less, and the remaining 45 districts had samples of 41 to 120 adults. The number of villages in the reduced data set is 286 , with 47 villages having samples of 9 or less adults and 193 villages having samples of 10 to 19 adults. The total number of adults in the sample is $n=4,028$.

Table 1 shows the results obtained when fitting the full two part model to the sample data, using the Bayesian methodology and software described in Section 3.3. The covariate (regressor) variables in the two models have been selected by application of some standard model selection procedures. All the covariates except for age, education and household size are dummy variables, taking the value 1 when the variable definition is satisfied. We used normal prior distributions with large variances for the elements of the vector coefficients $\beta, \gamma$, and uniform priors with large (but finite) range for the standard deviations underlying the two parts of the model and the coefficients $K_{u}$ and $K_{v}$ in Equation 6. By default, WinBUGS automatically selects the method of sampling from the conditional distribution of
each of the fixed and random parameters when applying the Gibbs sampler. Notice that the conditional distributions don't have a closed-form under the present full model. The software selects an acceptance/rejection method for the logistic part, and slice sampling (Neal 2000) for most of the other parameters and random effects.

For the MCMC simulations we generated a chain of length 50,000 , discarded the first 5,000 sampled values as "burn in", and then thinned the chain by taking every $150^{\text {th }}$ sample value. Discarding the first 5,000 sampled values was found sufficient to guarantee the convergence of the chain, using some informal commonly used graphical techniques. These include comparing the histograms of the posterior distributions of the various parameters based on different sub-sequences of the chain, inspecting the traces of several chains simulated in parallel, each with different starting values to check for stabilization of the chain, and plotting the autocorrelations of the sampled values to verify independence after appropriate thinning. See Gamerman and Lopes (2006) for further discussion and illustrations, including more formal tests of convergence. Note also that the simulation results in Section 4, using the model fitted to the real data and generating a separate chain of length 50,000 for each simulation and discarding the first 5,000 values as "burn in" yield very satisfactory results, thus providing another indication for the convergence of the chain after the first 5,000 values.

The estimated $K$-coefficients and variances of the random effects imply, $\widehat{\operatorname{Corr}}\left(u_{i}, u_{i}^{*}\right)=0.45 ; \widehat{\operatorname{Corr}}\left(v_{i j}, v_{i j}^{*}\right)=$ 0.21 . Interestingly, the correlations are close to the empirical correlations reported at the end of Section 2.1, using the raw means.

The main results emerging from Table 1 can be summarized as follows. All the regressor coefficients are highly significant (based on standard t-tests) and generally have anticipated signs. Other variables considered for inclusion in the two models were found to be nonsignificant. The variances of the random effects are highly significant in both models, indicating their contribution in explaining the variation of the scores, or the probabilities of positive scores, not explained by the covariates included in the two models.

As a further diagnostic for the logistic mixed model we show in Figure 3 a scatter plot of the observed proportions of positive scores $\left(\mathrm{I}_{i j k}=1\right)$ against the average of the predicted probabilities of positive scores under the model, in groups of 50 individuals defined by the ordered values of the predicted probabilities. The plotted values are almost on a straight line, showing a good fit. Figure 4 shows a histogram of the estimated standardized residuals of the mixed linear part, $\hat{z}_{i j k}=\hat{\varepsilon}_{i j k} / \operatorname{SD}\left(\hat{\varepsilon}_{i j k}\right)=\left(y_{i j k}-x_{i j k}^{\prime} \hat{\beta}-\right.$ $\left.\hat{u}_{i}-\hat{v}_{i j}\right) / \operatorname{SD}\left(\hat{\varepsilon}_{i j k}\right)$, where $\operatorname{SD}\left(\hat{\varepsilon}_{i j k}\right)$ is the empirical standard deviation of the estimated residuals. Although not a 'perfect' bell shape, the histogram does not indicate severe divergence from a normal distribution.

Table 1
Estimated parameters and standard errors (Std Err.) when fitting the two-part model

|  | Linear part |  | Logistic part |  |
| :---: | :---: | :---: | :---: | :---: |
| Regressors | Estimate | Std Err. | Estimate | Std Err. |
| Constant | $\hat{\beta}_{0}=6.90$ | 4.00 | $\hat{\gamma}_{0}=-6.48$ | 0.58 |
| Years at school | $\hat{\beta}_{1}=7.28$ | 0.53 | $\hat{\gamma}_{1}=2.16$ | 0.12 |
| Years at school ${ }^{2}$ | $\hat{\beta}_{2}=-0.24$ | 0.05 | $\hat{\gamma}_{2}=-0.13$ | 0.01 |
| Attended literacy program | ${ }^{1}$ | - | $\hat{\gamma}_{3}=2.44$ | 0.27 |
| Helped by interviewer | - | - | $\hat{\gamma}_{4}=2.00$ | 0.17 |
| Low income | $\hat{\beta}_{5}=-2.61$ | 0.88 | $\hat{\gamma}_{5}=-0.35$ | 0.14 |
| Civil servant/professional | $\hat{\beta}_{6}=13.91$ | 1.89 | ${ }^{\text {r }}$ | - |
| Gender (1 for female) | $\hat{\beta}_{7}=-1.60$ | 0.81 | $\hat{\gamma}_{7}=-0.59$ | 0.14 |
| Household size (adults) | $\hat{\beta}_{8}=0.94$ | 0.29 | $\hat{\gamma}_{7}$ | - |
| Age | $\hat{\beta}_{9}=0.84$ | 0.16 | $\hat{\gamma}_{9}=0.14$ | 0.02 |
| Age ${ }^{2}$ | $\hat{\beta}_{10}=-0.01$ | 0.002 | $\hat{\gamma}_{10}=-0.002$ | 0.00 |
| Variances | Estimate | Std Err. | Estimate | Std Err. |
| Between Districts | $\hat{\sigma}_{u}^{2}=66.31$ | 16.72 | $\hat{\sigma}_{u^{*}}^{2}=1.28$ | 0.34 |
| Between Villages | $\hat{\sigma}_{v}^{2}=66.58$ | 10.45 | $\hat{\sigma}_{v^{* *}}^{u^{*}}=0.86$ | 0.19 |
| Residual | $\hat{\sigma}_{\varepsilon}^{2}=322.0$ | 10.12 | $v$ | - |
| $K$-Coefficients (Equation 6)* | Estimate |  | Std Err. |  |
| District random effects | $\hat{K}_{u}=0.06$ |  | 0.02 |  |
| Village random effects | $\hat{K}_{v}=0.02$ |  | 0.01 |  |



Figure 3 Observed and predicted probabilities of positive scores


Figure 4 Histogram of standardized residuals for the linear part
As a final assessment of the goodness of fit of the two part model, we generated 200 new data sets of size $n=4,028$ from the estimated two-part model of Table 1, using the same covariates as for the original sample. The test scores where generated by generating random effects and residuals $\left(u_{i}, v_{i j}, \varepsilon_{i j k}\right)$ with estimated variances $\left(\hat{\sigma}_{u}^{2}, \hat{\sigma}_{v}^{2}, \hat{\sigma}_{\varepsilon}^{2}\right)$ (Equation 3), generating random effects $\left(u_{i}^{*}, v_{i j}^{*}\right)$ using Equation (6) with estimated coefficients and variances $\left(\hat{K}_{u}, \hat{\sigma}_{u^{*} \mid u}^{2}\right.$, $\left.\hat{K}_{v}, \hat{\sigma}_{v^{*} \mid v}\right)$ drawing at random 1 or 0 with probabilities $\operatorname{Pr}\left(\mathrm{I}_{i j k}=1\right)=\exp \left(x_{i j k}^{\prime *} \gamma+u_{i}^{*}+v_{i j}^{*}\right) \times\left[1+\exp \left(x_{i j k}^{\prime *} \gamma+u_{i}^{*}+v_{i j}^{*}\right)\right]^{-1}$ (Equation 5), and in the case of 1 , generating the nonzero scores $y_{i j k}=x_{i j k}^{\prime} \hat{\beta}+u_{i}+v_{i j}+\varepsilon_{i j k} \quad$ (Equation 3). The
variance $\hat{\sigma}_{u^{*} \mid u}^{2}$ was computed as $\hat{\sigma}_{u^{*} \mid u}^{2}=\left(\hat{\sigma}_{u^{*}}^{2}-\hat{K}_{u}^{2} \hat{\sigma}_{u}^{2}\right)$, and similarly for $\hat{\sigma}_{v^{*} \mid v}^{2}$ (Equation 6). Next we calculated for each data set the score means and proportions for each village and district and used them to compute empirical confidence intervals based on the 200 means and proportions. Table 2 shows the proportions of times that the empirical confidence intervals (C.I.) contain the corresponding actual sample values in the Cambodia survey.

The results in Table 2 show very close coverage rates to the nominal values for the villages, but under-coverage of up to $10 \%$ for the districts, which is probably explained by the fact that the latter rates are based on only 50 districts.

### 4.2 Simulation study

The purpose of the simulation experiment is to study the effectiveness of the two-part model for producing small area predictors and associated measures of prediction errors. The simulation experiment enables also to compare the results obtained under this model with results obtained when fitting the two parts of the model separately, ignoring the correlations between the corresponding random effects in the two parts, and with the results obtained when fitting a linear mixed model to all the responses, ignoring the accumulation of zero scores. To this end, we generated 300 new populations of $N=4,028$ scores and 300 new samples of size $n=1,026$, similar to the generation of the data sets used for the computation of the confidence intervals in Table 2, but from a model with fewer regressors than in the model shown in Table 1. In the logistic part we included 4 regressors: 'number of years at school', 'attendance of a literacy programme', 'helped by the interviewer' and 'having low income'. In the linear part we included 5 regressors: 'number of years at school', 'gender', 'household size', 'age', and 'age ${ }^{2}$ '. In order to set parameter values, we fitted separately the linear part and the logistic part with the fewer regressors to the original sample data. The correlations between the random effects of the logistic and the linear parts were set to 0.5 at both the district and the village level.

Table 2
Proportions of times that the empirical confidence intervals contain the actual sample means and proportions

|  | Empirical 90\% C.I. |  | Empirical 95\% C.I. |  |
| :--- | :---: | :---: | :---: | :---: |
| Small areas | Districts | Villages | Districts | Villages |
| Number of areas | 50 | 286 | 50 | 286 |
| \% Coverage of proportions | $80 \%$ | $88 \%$ | $88 \%$ | $95 \%$ |
| \% Coverage of means | $88 \%$ | $89 \%$ | $90 \%$ | $94 \%$ |

The district and village means (proportions) in the simulated populations were taken as the true district and village means (proportions), thus allowing us to assess the performance of the various predictors. As noted in Remark 2 in Section 2.2, the prediction of small areas means and proportions under the two-part model requires knowledge of the covariates for all the population units. This requirement was satisfied in the simulation study since the simulated populations use the regressors of the original sample of the 4,028 individuals. In order to specify sampled values for the regressor variables, we sampled 1,026 individuals and used the sampled regressors for all the 300 samples. Half of the individuals in half of the 286 villages were included in the sample, except for villages with fewer than 5 adults in the original dataset, where all the individuals were sampled. This minimum size criterion was applied in order to avoid computational problems when running the simulations (see Section 4.3). The sample contained individuals from all the 50 rural districts, with 1 district having a sample of size 4, 4 districts having a sample of size 9,17 districts having samples of size $15 \leq n_{d} \leq 20$, and the remaining 28 districts having samples of size $21 \leq n_{d} \leq 30$. As mentioned above, the sample contained individuals from half of the 286 villages, with 29 villages having samples of size $2 \leq n_{v} \leq 5$, 109 villages having samples of size $6 \leq n_{v} \leq 10$, and 5 villages having samples of size $11 \leq n_{v} \leq 19$.

The results of the simulation study are shown in Tables 3 and 4 and in Figures 5-6. Table 3 shows the mean estimates of the model coefficients and the root mean square errors (RMSE) over the 300 simulations, as obtained when fitting the three models to the sample data; A- the full two-part model that accounts for the correlations between the district and village random effects in the two parts of the model, Bthe two part model that ignores the correlations between the district and village random effects in the two parts, that is, when fitting the two parts separately, and C- the linear mixed model defined by (3) but fitted to all the responses, including the zero scores. This model ignores the accumulation of zero scores, but in order to make it more comparable to the two part model, we included in this model all the regressors included in either the logistic or the linear part of the two-part model. The linear mixed model can practically only be used for predicting the district and village means. For comparability reasons we fitted all the three models using the WinBUGS software (thus following the Bayesian paradigm), but it is important to mention that fitting the models B and C using the MLwiN software (Rasbash et al. 2002), which is much faster, yields very similar results.

Table 3 exhibits only minor differences between the mean estimates and RMSEs when fitting the full model or when fitting the two parts separately. For the linear part the mean
estimates are very close to the corresponding true coefficients, indicating lack of bias. For the logistic part the mean estimates are again close to the true coefficients although the estimated biases are statistically significant based on the conventional $t$-statistic. The fact that the RMSEs are similar when fitting the full model and when fitting the two parts separately suggests that under the present simulation set-up, accounting for the correlations between the random effects in the two parts does not improve the estimation of the model regression coefficients. In contrast, the results in Table 4 reveal much smaller biases and RMSEs when estimating the variances of the logistic model by fitting the full model, although the estimation of the "between villages" variance is still highly biased. The estimation of the correlations between the random effects of the two parts is satisfactory. Finally, as indicated by both tables, fitting the mixed linear model, ignoring the accumulation of zeroes generally yields highly biased estimators and consequently large RMSEs, which of course is not surprising.

Figure 5 shows the bias and RMSE when predicting the true district and village means and proportions under the three models. Let $\hat{U}_{a}^{r}$ represent any of the predictors under the three models (means or proportions) for a given area $a$ as obtained in simulation $r$, and denote by $U_{a}^{r}$ the corresponding true predicted value. The bias and RMSE were calculated as,

$$
\begin{align*}
\operatorname{Bias}_{a} & =\sum_{r=1}^{300}\left(\hat{U}_{a}^{r}-U_{a}^{r}\right) / 300 ; \\
\operatorname{RMSE}_{a} & =\left[\sum_{r=1}^{300}\left(\hat{U}_{a}^{r}-U_{a}^{r}\right)^{2} / 300\right]^{1 / 2} . \tag{12}
\end{align*}
$$

The figures pertaining to villages are based on the 273 villages (out of the 286) where sampling took place. (As mentioned before, all the individuals in villages with fewer than 5 adults in the original dataset were included in the sample.)

The clear conclusion from Figures 5a, 5c, 5e and 5 g is that the use of the mixed linear model alone for predicting the district and village means yields biased predictors in both sampled and nonsampled areas, and hence large RMSEs. Note, however, that the RMSEs of the predictors produced under the linear model for villages without samples are similar to the RMSEs obtained under the twopart model. This outcome is probably explained by the fact that the mixed linear model is much simpler and depends on fewer parameters than the two part model, resulting in smaller prediction variances in villages with no samples than the prediction variances of the two-part model predictors. Figures (5a)-(5d) show that the predictors produced under the two-part model, whether fitted jointly or separately are basically unbiased, despite the bias in the estimation of some of the parameters of the logistic part noticed in Tables 3 and 4. Figures (5e)-(5h) don't show any
appreciable difference in the RMSE between the use of the full model or by fitting the two parts separately, which was noted also in Tables 3 and 4.

Figure 6 shows the percentage of times that $95 \%$ credibility intervals, produced under the three models, cover the true district or village means and proportions. See Section 3.3 for the construction of credibility interval boundaries when using MCMC simulations. The prominent conclusion emerging from Figure 6 is that ignoring the accumulation of zeroes and fitting the linear mixed model alone yields for most areas coverage rates for the true area means that are very different from the nominal $95 \%$ rate, with particularly low rates for villages with samples. The fitting of the full model yields somewhat better coverage
rates for the district means than the fitting of the two parts separately, but the coverage rates of the district proportions are similar under the two methods. There seems to be little difference in the credibility intervals for the village means when fitting the full model or the two-parts separately, but it is interesting to note that the use of the full model yields better coverage rates in 77 per cent of the villages, whereas fitting the two parts separately yields better coverage rates in only 15 per cent of the villages. In the remaining villages the use of the two methods yields the same coverage rates. In the case of the village proportions, the two methods yield similar credibility intervals, except in a few cases where the use of the full model is seen to be generally better.

Table 3
Means and RMSE of estimators of model coefficients under the three models

|  |  | Simulation mean |  |  | Simulation RMSE |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Coefficient | True value | Full model | Separate fit | Linear model | Full model | Separate fit | Linear model |
|  |  |  | Linear part |  |  |  |  |
| $\beta_{0}$ | 9.38 | 8.83 | 9.73 | 1.90 | 6.95 | 6.95 | 9.21 |
| $\beta_{1}$ | 4.97 | 4.97 | 4.87 | 12.59 | 0.32 | 0.33 | 7.63 |
| $\beta_{2}$ | -1.65 | -1.61 | -1.58 | -3.24 | 2.05 | 2.05 | 2.27 |
| $\beta_{3}$ | 1.02 | 1.05 | 1.05 | 1.75 | 0.57 | 0.57 | 0.86 |
| $\beta_{4}$ | 0.94 | 0.97 | 0.96 | 1.51 | 0.27 | 0.26 | 0.60 |
| $\beta_{5}$ | -0.01 | -0.01 | -0.01 | -0.02 | 0.00 | 0.00 | 0.01 |
|  |  |  |  | Logistic part |  |  |  |
| $\gamma_{0}$ | -4.09 | -4.38 | -4.38 | - | 0.58 | 0.59 | - |
| $\gamma_{1}$ | 1.63 | 1.73 | 1.73 | - | 0.18 | 0.19 | - |
| $\gamma_{2}$ | 1.98 | 2.13 | 2.13 | $2.55^{*}$ | 0.41 | 0.42 | $7.33^{*}$ |
| $\gamma_{3}$ | 2.06 | 2.41 | 2.41 | $2.05^{*}$ | 0.65 | 0.65 | $2.64^{*}$ |
| $\gamma_{4}$ | -0.35 | -0.34 | -0.34 | $0.3^{*}$ | 0.30 | 0.30 | $1.37^{*}$ |

*Estimates obtained when including these regressors in the linear model.

Table 4
Means and RMSE of estimators of model variances and correlations under the three models

|  | True value | Simulation mean |  |  | Simulation RMSE |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Full model | Separate fit | Linear model | Full model | Separate fit | Linear model |
| Variances - linear part |  |  |  |  |  |  |  |
| District | 60.40 | 62.23 | 60.66 | 103.46 | 24.52 | 24.87 | 46.52 |
| Village | 65.44 | 70.37 | 70.36 | 111.97 | 24.84 | 25.75 | 49.74 |
| Residual | 336.00 | 338.31 | 338.61 | 696.82 | 23.76 | 24.04 | 361.64 |
| Variances - logistic part |  |  |  |  |  |  |  |
| District | 0.92 | 1.08 | 1.50 | - | 0.61 | 0.91 | - |
| Village | 0.57 | 0.91 | 1.15 | - | 0.70 | 0.94 | - |
| $K$-factors |  |  |  |  |  |  |  |
| District | 0.071 | 0.075 | - | - | 0.016 | - | - |
| Village | 0.054 | 0.055 | - | - | 0.012 | - | - |
| Correlations between random effects of the two parts |  |  |  |  |  |  |  |
| District | 0.500 | 0.506 | - | - | 0.151 | - | - |
| Village | 0.500 | 0.459 | - | - | 0.148 | - | - |


$\square$ full model $\square$ separate model *linear model
(5a) bias, district means

$\square$ full model $\square$ separate model
(5b) bias, district proportions


$$
\square \text { full model separate model * linear model }
$$

(5c) bias, village means

$\square$ full model $\mathbf{\square}$ separate model
(5d) bias, village proportions


- full model $\square$ separate model $*$ linear model
(5e) RMSE, district means


Districts ordered by ascending sample size
$\square$ full model $\square$ separate model
(5f) RMSE, district proportions

full model separate model * linear model
(5g) RMSE, village means

$\square$ full model $\mathbf{\square}$ separate model
(5h) RMSE, village proportions
Figure 5 BIAS and RMSE of predictors of area means and proportions



Figure 6 Coverage rates of 95\% credibility intervals for area means and proportions

### 4.3 Computational issues

As already noted, fitting the full model, accounting for the correlations between the district and village random effects in the two parts is computationally intensive and not always
stable. In particular, we encountered severe computation problems when fitting the full model with very small samples from most of the villages. For example, for a sample of 750 individuals from 264 villages, such that almost half of the villages had sample sizes of 1 or 2 , the sampled values
from the posterior distributions generated by the Gibbs sampler were found to be strongly correlated even at very high lags, over 1,000 lags for the village random effects and the correlation between the village random effects in the two parts, and still over 500 lags after tightening the prior distributions, which required extremely long chains to obtain sufficient data for inference. This makes it excessively computer intensive and almost impossible to verify convergence of some of the posterior distributions. For this reason we selected samples of size 1,026 in our simulation study, with at least 2 individuals from every village.

## 5. Summary

The most important message emerging from this paper is that ignoring the accumulation of zeroes and fitting a linear mixed model to the whole data set can result in highly biased predictors and wrong coverage rates of credibility intervals. Clearly, the magnitude of the bias and the performance of the credibility intervals will depend in this case on the percentage of zero scores. Fitting a two-part model to such data generally yields unbiased predictors and credibility intervals with acceptable coverage rates. Fitting the full two-part model, accounting for the correlations between the random effects of the two parts is the best choice, but it improved the predictions in our simulation study only marginally, despite the use of correlations of 0.5 between the district and village random effects in the two parts.

In this study we used MCMC simulations for fitting the models and computing the small area predictors and their variances. The use of this approach requires specifying prior distributions, which can affect the inference, particularly with a small number of sampled areas even when specifying noninformative priors. See Pfeffermann, Moura and Silva (2006) for recent discussion and illustrations. The other problem with the use of MCMC simulations is that it is very computing intensive. Furthermore, the use of this approach can become unstable if there are only few observations in the sampled areas. An alternative approach is therefore to fit the full two part model following the frequency approach. Available software include MLwiN (Goldstein 2003) and aML (Lillard and Panis 2003), but the use of these or other softwares requires modifications to the estimation of the prediction variance that account for the errors in the estimation of the fixed model parameters. Resampling methods like the bootstrap or jackknife could be considered for this purpose, but they require new developments appropriate for this model.

## Acknowledgements

The work of Bénédicte Terryn was conducted for the most part while at UNESCO Institute for Statistics (UIS), Montreal, Canada. The work of Fernando Moura was funded in part by a research grant from the Brazilian National Council for the Development of Science and Technology. The authors thank an associate editor and two reviewers for their constructive comments and suggestions, and Pedro Silva for helping with the final computations.

## References

Barry, S.C., and Welsh, A.H. (2002). Generalized additive modelling and zero inflated count data. Ecological Modelling, 157, 179-188.

Gamerman, D., and Lopes, H.F. (2006). Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference. $2^{\text {nd }}$ Edition. Chapman \& Hall.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

Goldstein, H. (2003). Multilevel Statistical Models. $3^{\text {rd }}$ Edition. London, Edward Arnold.

Hall, P., and Maiti, T. (2006). On parametric bootstrap methods for small area predictions. Journal of the Royal Statistical Society, 68, Series B, 221-238.

Jiang, J., and Lahiri, P. (2005). Mixed model prediction and small area estimation. TEST, 15, 65-72.

Jiang, J., Lahiri, P. and Wan, S. (2002). A unified jackknife method. Annals of Statistics, 30, 1782-1810.

Lillard, L.A., and Panis, C.W.A. (2003). aML Multilevel Multiprocess Statistical Software, Version 2.0. EconWare, Los Angeles, California.

Neal, R.M. (2000). Slice Sampling. Technical Report No.2005, Department of Statistics, University of Toronto.

Olsen, M.K., and Schafer, J.L. (2001). A two-part random effects model for semicontinuous longitudinal data. Journal of the American Statistical Association, 96, 730-745.

Pfeffermann, D., Moura, F.A.D.S. and Silva, P.L.D.N. (2006). Multilevel modelling under informative sampling. Biometrika, 93, 943959.

Rao, J.N.K. (2003). Small Area Estimation. New York: John Wiley \& Sons, Inc.

Raudenbush, S.W., Yang, M. and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics, 9, 141-157.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. and Lewis, T. (2002). A user's guide to MLwiN. London: Institute of Education.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.Rproject.org.

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581592.

Spiegelhalter, D., Thomas, A. and Best, N.G. (2003). Bayesian Inference using Gibbs Sampling. WinBUGS version 1.4, User manual. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, U.K.

Tanner, M.A. (1996). Tools for Statistical Inference, $3^{\text {rd }}$ edition. New York: Springer-Verlag.

## ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2008.
B. Baffour, University of Southampton
J.-F. Beaumont, Statistics Canada
Y. Bélanger, Statistics Canada
W. Bell, U.S. Census Bureau
C. Bocci, Statistics Canada
H.J. Boonstra, Statistics Netherlands
J. Breidt, Colorado State University
P. J. Cantwell, U.S. Census Bureau
R. Chambers, University of Wollongong, Australia
A. Cyr, Statistics Canada
G. Datta, University of Georgia
T.J. DeMaio, U.S. Census Bureau
J.A. Dever, University of Maryland
D. Dolson, Statistics Canada
J. Eltinge, Bureau of Labor Statistics
M. Ferland, Statistics Canada
W.A. Fuller, Iowa State University
J. Gambino, Statistics Canada
W. Gamrot, University of Economics, Katowice, Poland
D. Garriguet, Statistics Canada
M. Ghosh, University of Florida
C. Goga, Institut de Mathématiques de Bourgogne
S. Haslett, Massey University, New Zealand
Y. He, Harvard Medical School
D. Haziza, Université de Montréal/Statistics Canada
M. Hidiroglou, Statistics Canada
N.J. Horton, Smith College
J. Jiang, University of California Davis
C. Julien, Statistics Canada
J. Kim, Iowa State University
J.-K. Kim, Yonsei University, Seoul
P.S. Kott, USDA/National Agricultural Statistics Service
R.A. Kulka, Abt Associates Inc.
P. Lahiri, JPSM, University of Maryland
H. Lee, Westat, Inc.
M. Larsen, Iowa State University
P. Lavallée, Statistics Canada
J.C. Legg, Iowa State University
R. Lehtonen, University of Helsinki
S. Lele, University of Alberta
C. Léon, Statistics Canada
B. Liu, Westat, Inc.
S. Lohr, Arizona State University
L. Mach, Statistics Canada
H. Mantel, Statistics Canada
T. Maiti, Iowa State University
B. Mandram, Worcester Polytechnic Institute
J. Maples, U.S. Census Bureau
A. Matei, Université de Neuchâtel, Suisse
Y. McNab, UBC
T. Merkouris, Athens University of Economics and Business
S. M. Miller, U.S. Bureau of Labor Statistics
I. Molina-Peralta, Universidad Carlos III de Madrid
G.E. Montanari, University of Perugia
F. Moura, Universidade do Brasil-UFRJ
C. Nadeau, Statistics Canada
L. Nordberg, Statistics Sweden
J. Oleson, University of Iowa
J. Opsomer, Colorado State University
M. Park, Korea University
M. Pratesi, Università di Pisa
J.N.K. Rao, Carleton University
L.-P. Rivest, Université de Laval
L. Rizzo, Westat, Inc.
S. Rubin-Bleuer, Statistics Canada
N. Salvati, Università di Pisa
F. Scheuren, National Opinion Research Center
D. Nóbrega da Silva, Universidade Federal do Rio Grande do Norte
S. Singh, University of Texas at Brownsville
S. Sinha, Carleton University
P. Smith, Office for National Statistics
J.-L. Tambay, Statistics Canada
Y. Tillé, Université de Neuchâtel
M. Torabi, University of Alberta
D. Toth, Bureau of Labor Statistics
N. Tzavidis, University of Manchester
R. Valliant, JPSM
F. Verret, Statistics Canada
D. Willimack, United States Census Bureau
M. Wolfson, Statistics Canada
C. Wu, University of Waterloo
Y. You, Statistics Canada
W. Yung, Statistics Canada

Acknowledgements are also due to those who assisted during the production of the 2008 issues: Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau of Household Survey Methods Division, Susie Fortier and Nick Budko of Business Survey Methods Division, Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Denis Coutu, Darquise Pellerin and Isabelle Poliquin (Dissemination Division), Sheri Buck (Systems Development Division) and Sylvie Dupont (Official Languages and Translation Division).

## 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 <br> Volume 24, No. 2, 2008

Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator
Carl-Erik Särndal and Sixten Lundström ..... 167
Model-Based Inference for Two-Stage Cluster Samples Subject to Nonignorable Item Nonresponse Ying Yuan and Roderick J.A. Little ..... 193
An Implementation Strategy for Efficient Convergence of the Lavallée and Hidiroglou Stratification Algorithm Patricia Gunning, Jane M. Horgan and Gary Keogh ..... 213
Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure Natalie Shlomo and Ton De Waal ..... 229
Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation
Mick P. Couper, Eleanor Singer, Frederick G. Conrad and Robert M. Groves ..... 255
Use of Deflators in Business Surveys: An Analysis Based on Italian Micro Data Leandro D'Aurizio and Raffaele Tartaglia-Polcini. ..... 277
An Experiment on Perceived Survey Response Burden Among Businesses
Dan Hedlin, Helen Lindkvist, Helena Bäckström and Johan Erikson ..... 301
Letter to the Editor
Prof. Krishnamurty Muralidhar and Dr. Rathindra Sarathy ..... 319
Book and Software Reviews
Safaa Rabie Amer, John Dixon, Michael D. Larsen, Michael W. Link, Stephen M. Miller, Dr. Jean Ritzen and Mr. Andreas Lindner ..... 323
In Other Journals ..... 339

## Contents Volume 24, No. 3, 2008

Special Section on Longitudinal Surveys ..... 341
Estimating Models for Panel Survey Data under Complex Sampling
Marcel D.T. Vieira and Chris J. Skinner ..... 343
The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study Ian Plewis, Sosthenes C. Ketende, Heather Joshi and Gareth Hughes ..... 365
Seam Effects in Longitudinal Surveys
Mario Callegaro ..... 387
Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection Annette Jackle. ..... 411
Explaining the Size and Nature of Response in a Survey on Health Status and Economic Standard Fredrik Johansson and Anders Klevmarken ..... 431
Finite Sample Revision Variances for ARIMA Model-Based Signal Extraction
Tucker McElroy, Richard Gagnon ..... 451
Weight Adjustments for the Grouped Jackknife Variance Estimator Richard Valliant, J. Michael Brick and Jill A. Dever ..... 469
Book and Software Reviews
Rob Burnside and Charlotte Steeh ..... 489

All inquires about submissions and subscriptions should be directed to jos@scb.se

## Volume 36, No. 1, March/mars 2008

Paul GUSTAFSON
Editor's report .....  1
Louis-Paul RIVEST
Capture-recapture models: Introduction ..... 3
Saman MUTHUKUMARANA, Carl J. SCHWARZ \& Tim B. SWARTZ
Bayesian analysis of mark-recapture data with travel time-dependent survival probabilities ..... 5
Olivier GIMENEZ
Discussion: Towards a Bayesian analysis template? .....  21
Saman MUTHUKUMARANA, Carl J. SCHWARZ \& Tim B. SWARTZ
Authors' response. ..... 24
Roger PRADEL, Lory MAURIN-BERNIER, Olivier GIMENEZ, Meritxell GENOVART, Ŕemi CHOQUET \& Daniel ORO Estimation of sex-specific survival with uncertainty in sex assessment .....  29
Rémi CHOQUET
Automatic generation of multistate capture-recapture models ..... 43
Gilles GAUTHIER \& Jean-Dominique LEBRETON
Analysis of band-recovery data in a multistate capture-recapture framework ..... 59
Louis-Paul RIVEST
Why a time effect often has a limited impact on capture-recapture estimates in closed populations ..... 75
Sophie VÉRAN \& Jean-Dominique LEBRETON
The potential of integrated modelling in conservation biology:
A case study of the black-footed albatross (Phoebastria nigripes) ..... 85
Kanti V. MARDIA, Gareth HUGHES, Charles C. TAYLOR \& Harshinder SINGH
A multivariate von Mises distribution with applications to bioinformatics .....  99
Qing PAN \& Douglas E. SCHAUBEL
Proportional hazards models based on biased samples and estimated selection probabilities ..... 111
Jiahua CHEN, Pengfei LI \& Yuejiao FU
Testing homogeneity in a mixture of von Mises distributions with a structural parameter ..... 129
Mayer ALVO
Nonparametric tests of hypotheses for umbrella alternatives ..... 143
Christophe CROUX, Gentiane HAESBROECK \& Kristel JOOSSENS
Logistic discrimination using robust estimators: An influence function approach ..... 157
Forthcoming papers/Articles à paraître. ..... 175
Acknowledgement of referees' services/Remerciements aux membres des jurys ..... 176
Volume 36 (2008): Subscription rates/Frais d'abonnement ..... 177

## Volume 36, No. 2, June/juin 2008

Haiying CHEN, Elizabeth A. STASNY \& Douglas A. WOLFE
Ranked set sampling for ordered categorical variables ..... 179
Benjamin KEDEM, Guanhua LU, Rong WEI \& Paul D. WILLIAMS
Forecasting mortality rates via density ratio modeling. ..... 193
Piotr KOKOSZKA, Inga MASLOVA, Jan SOJKA \& Lie ZHU
Testing for lack of dependence in the functional linear model ..... 207
Valeria SAMBUCINI \& Ludovico PICCINATO
Likelihood and Bayesian approaches to inference for the stationary point of a quadratic response surface ..... 223
Bruno SANSÓ, Alexandra M. SCHMIDT \& Aline A. NOBRE
Bayesian spatio-temporal models based on discrete convolutions ..... 239
Lei SHI \& Gemai CHEN
Local influence in multilevel models ..... 259
Sanjoy K. SINHA
Robust methods for generalized linear models with nonignorable missing covariates ..... 277
Zhenlin YANG, Eden Ka-Ho WU \& Anthony F. DESMOND Inference for general parametric functions in Box-Cox-type transformation models ..... 301
Lan WANG
Nonparametric test for checking lack of fit of the quantile regression model under random censoring ..... 321
Forthcoming papers/Articles à paraître. ..... 337
Online access to The Canadian Journal of Statistics/Services en ligne de La revue canadienne de statistique ..... 338
Volume 36 (2008): Subscription rates/Frais d'abonnement ..... 339


[^0]:    1. Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: methomps@uwaterloo.ca.
[^1]:    1. Joanne Pascale, US Census Bureau, 4600 Silver Hill Road, Washington DC 20233. E-mail: joanne.pascale@census.gov; Alice McGee, formerly with the National Centre for Social Research.
[^2]:    A. Demographics

    1. LIVE: Does NAME still live here?
    2. DOB: Can I just check, is NAME's date of birth [ffill date of birth (DOB)]?
    3. CHILD: Our records show that when we last interviewed you, you had a child called NAME, whose date of birth is [DOB]. Are these details correct?
    B. Health Conditions
    4. LAST-EYE/CVD/CHRON: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [fill EYE/CVD/CHRONIC CONDITION]. [press enter]
    5. STILL-EYE/CVD/CHRON: Do you still have [fill EYE/CVD/CHRONIC CONDITION]?
    C. Vehicle Ownership
    6. VEHICLE: Last time we saw you, you told us that you were the main user of a [MAKE OF VEHICLE], with a [LETTER] registration. Do you still have that vehicle?
[^3]:    1. Jing Xu, Jun Shao and Lin Wang, Department of Statistics, University of Wisconsin, Madison, WI 53706; Mari Palta, Department of Population Health Sciences and Department of Biostatistics and Medical Informatives, University of Wisconsin, Madison, WI 53706.
[^4]:    1. Robert G. Clark and Raymond L. Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australia. E-mail: Robert_Clark@uow.edu.au.
[^5]:    1. Lionel Qualité and Yves Tillé, Institut de Statistique, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland. E-mail: yves.tille@unine.ch.
[^6]:    1. Inho Park, Statistician, Economic Statistics Department, The Bank of Korea, Namdaemun-Ro 106, Jung-Gu, Seoul 100-794, Korea. E-mail: ipark@bok.or.kr.
[^7]:    1. Roberto Benedetti, Professor, Department of Business, Statistical, Technological and Environmental Sciences (DASTA), "G. d'Annunzio" University, Viale Pindaro 42, Pescara, IT-65127. E-mail: benedett@unich.it; Giuseppe Espa, Professor, Department of Economics, University of Trento, Via Inama 5, Trento, IT-38100. E-mail: giuseppe.espa@economia.unitn.it; Giovanni Lafratta, Assistant Professor, Department of Business, Statistical, Technological and Environmental Sciences (DASTA), "G. d'Annunzio" University, Viale Pindaro 42, Pescara, IT-65127. E-mail: giovanni.lafratta@unich.it.
[^8]:    1. Mohammad G. Mostafa Khan, Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A., (M.G.M. Khan was on leave from the School of Computing, Information and Mathematical Sciences, University of the South Pacific, Suva, Fiji); Niraj Nand and Nesar Ahmad, School of Computing, Information and Mathematical Sciences, University of the South Pacific, Suva, Fiji.
[^9]:    1. José A. Díaz-García and Liliana Ulloa Cortez, Department of Statistics and Computation, Universidad Autónoma Agraria Antonio Narro, 25350 Buenavista, Saltillo, Coahuila, MEXICO. E-mail: jadiaz@uaaan.mx and lullcor@uaan.mx.
[^10]:    1. Piero Demetrio Falorsi, Italian National Statistical Institute. E-mail: falorsi@istat.it; Paolo Righi, Italian National Statistical Institute. E-mail: parighi@istat.it.
[^11]:    ${ }^{1}$ Size classes are defined in terms of number of persons employed.
    ${ }^{2}$ Regions are 21 including autonomous provinces.

[^12]:    1. Danny Pfeffermann, Hebrew University of Jerusalem, Israël, and Southampton Statistical Sciences Research Institute, UK. E-mail: msdanny@huji.ac.il; Bénédicte Terryn, Department for International Development (DFID), UK; Fernando A.S. Moura, Federal University of Rio de Janeiro, Brazil.
