Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Survey Methodology
45-3

Release date: December 17, 2019

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

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 "Contact us" > "Standards of service to the public."

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 co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version **is also available.**

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

December 2019

•

Volume 45

•

Number 3

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/surveymethodology). To communicate with the Editor, please use the following email: (statcan.smj-rte.statcan@canada.ca).

Survey Methodology

A Journal Published by Statistics Canada

Volume 45, Number 3, December 2019

Contents

Regular Papers

Short note

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 395-425
Statistics Canada, Catalogue No. 12-001-X

Estimation of level and change for unemployment using structural time series models

Harm Jan Boonstra and Jan A. van den Brakel[1]

Abstract

Monthly estimates of provincial unemployment based on the Dutch Labour Force Survey (LFS) are obtained using time series models. The models account for rotation group bias and serial correlation due to the rotating panel design of the LFS. This paper compares two approaches of estimating structural time series models (STM). In the first approach STMs are expressed as state space models, fitted using a Kalman filter and smoother in a frequentist framework. As an alternative, these STMs are expressed as time series multilevel models in an hierarchical Bayesian framework, and estimated using a Gibbs sampler. Monthly unemployment estimates and standard errors based on these models are compared for the twelve provinces of the Netherlands. Pros and cons of the multilevel approach and state space approach are discussed.

Multivariate STMs are appropriate to borrow strength over time and space. Modeling the full correlation matrix between time series components rapidly increases the numbers of hyperparameters to be estimated. Modeling common factors is one possibility to obtain more parsimonious models that still account for cross-sectional correlation. In this paper an even more parsimonious approach is proposed, where domains share one overall trend, and have their own independent trends for the domain-specific deviations from this overall trend. The time series modeling approach is particularly appropriate to estimate month-to-month change of unemployment.

Key Words: Small area estimation; Structural time series models; Time series multilevel models; Unemployment estimation.

1 Introduction

Statistics Netherlands uses data from the Dutch Labour Force Survey (LFS) to estimate labour status at various aggregation levels. National estimates are produced monthly, provincial estimates quarterly, and municipal estimates annually. Traditionally monthly publications about the labour force were based on rolling quarterly figures compiled by means of direct generalized regression estimation (GREG), see e.g., Särndal, Swensson and Wretman (1992). The continuous nature of the LFS allows to borrow strength not only from other areas, but also over time. A structural time series model (STM) to estimate national monthly labour status for 6 gender by age classes is in use since 2010 (van den Brakel and Krieg, 2009, 2015).

Until now, provincial estimates are produced quarterly using the GREG. In order to produce figures on a monthly basis, a model-based estimation strategy is necessary to overcome the problem of too small monthly provincial sample sizes. In this paper a model is proposed that combines a time series modeling approach to borrow strength over time with cross-sectional small area models to borrow strength over space with the purpose to produce reliable monthly estimates of provincial unemployment. As a consequence of the LFS panel design, the monthly GREG estimates are autocorrelated and estimates based on follow-up waves are biased relative to the first wave estimates. The latter phenomena is often referred to as rotation group bias (Bailar, 1975). Both features need to be accounted for in the model (Pfeffermann, 1991). Previous accounts of regional small area estimation of unemployment, where strength is borrowed over both time

1. Harm Jan Boonstra, Statistics Netherlands, Department of Statistical Methods. E-mail: hjh.boonstra@cbs.nl; Jan A. van den Brakel, Statistics Netherlands, Department of Statistical Methods and Maastricht University, Department of Quantitative Economics.

and space, include Rao and Yu (1994); Datta, Lahiri, Maiti and Lu (1999); You, Rao and Gambino (2003); You (2008); Pfeffermann and Burck (1990); Pfeffermann and Tiller (2006); van den Brakel and Krieg (2016), see also Rao and Molina (2015), Section 4.4 for an overview.

In this paper, multivariate STMs for provincial monthly labour force data are developed as a form of small area estimation to borrow strength over time and space, to account for rotation group bias and serial correlation induced by the rotating panel design. In a STM, an observed series is decomposed in several unobserved components like a trend, a seasonal component, regression components, other cyclic components and a white noise term for remaining unexplained variation. These components are based on stochastic models, to allow them to vary over time. The classical way to fit STMs is to express them as a state space model and apply a Kalman filter and smoother to obtain optimal estimates for state variables and signals. The unknown hyperparameters of the models for the state variables are estimated by means of maximum likelihood (ML) (Harvey, Chapter 3). Alternatively, state space models can be fitted in a Bayesian framework using a particle filter (Andrieu, Poucet and Holenstein (2010); Durbin and Koopman (2012), Chapter 9). STMs can also be expressed as time series multilevel models and can be seen as an extension of the classical Fay-Herriot model (Fay and Herriot, 1979). Connections between structural time series models and multilevel models have been explored before from several points of view in Knorr-Held and Rue (2002); Chan and Jeliazkov (2009); McCausland, Miller and Pelletier (2011); Ruiz-Cárdenas, Krainski and Rue (2012); Piepho and Ogutu (2014); Bollineni-Balabay, van den Brakel, Palm and Boonstra (2016). In these papers the equivalence between state space model components and multilevel components is made more explicit. Multilevel models can both be fitted in a frequentist and hierarchical Bayesian framework, see Rao and Molina (2015), Section 8.3 and 10.9, respectively.

This paper contributes to the small area estimation literature by comparing differences between STMs for rotating panel designs that are expressed as state space models and as time series multilevel models. State space models are fitted using a Kalman filter and smoother in a frequentist framework where hyperparameters are estimated with ML. In this case models are compared using AIC and BIC. Time series multilevel models are fitted in an hierarchical Bayesian framework, using a Gibbs sampler. Models with different combinations of fixed and random effects are compared based on the Deviance Information Criterion (DIC). The estimates based on multilevel and state space models and their standard errors are compared graphically and contrasted with the initial survey regression estimates. Modeling cross-sectional correlation in multivariate time series models rapidly increases the number of hyperparameters to be estimated. One way to obtain more parsimonious models is to use common factor models. In this paper an alternative approach to model correlations between time series components indirectly is proposed, based on a global common trend and local trends for the domain-specific deviations.

The paper is structured as follows. In Section 2 the LFS data used in this study are described. Section 3 describes how the survey regression estimator (Battese, Harter and Fuller, 1988) is used to compute initial estimates. These initial estimates are the input for the STM models, which are discussed in Section 4. In Section 5 the results based on several state space and multilevel models are compared, including estimates

for period-to-period change for monthly data. Section 6 contains a discussion of the results as well as some ideas on further work. Throughout the paper we refer to the technical report by Boonstra and van den Brakel (2016) for additional details and results.

2 The Dutch Labour Force Survey

The Dutch LFS is a household survey conducted according to a rotating panel design in which the respondents are interviewed five times at quarterly intervals. Each month a stratified two-stage sample of addresses is selected. All households residing on an address are included in the sample. In this study 72 months of LFS data from 2003 to 2008 are used. During this period the sample design was self-weighted. The first wave of the panel consists of data collected by means of computer assisted personal interviewing (CAPI), whereas the four follow-up waves contain data collected by means of computer assisted telephone interviewing (CATI).

The Netherlands is divided into twelve provinces which serve as the domains for which monthly unemployment figures are to be estimated. Monthly national sample sizes vary between 5 and 7 thousand persons in the first wave and between 3 and 5 thousand in the fifth wave. Provincial sample sizes are diverse, ranging from 31 to 1,949 persons for single wave monthly samples.

LFS data are available at the level of units, i.e., persons. A wealth of auxiliary data from several registrations is also available at the unit level. Among these auxiliary variables is registered unemployment, a strong predictor for the unemployment variable of interest. These predictors are used to compute initial estimates, which are input to the time series models.

The target variable considered in this study is the fraction of unemployed in a domain, and is defined as $\overline{Y}_{it} = \sum_{j \in i} y_{ijt} / N_{it}$, with y_{ijt} equal to one if person j from province i in period t is unemployed and zero otherwise and N_{it} the population size in province i and period t.

3 Initial estimates

Let $\hat{\overline{Y}}_{itp}$ denote the initial estimate for Y_{it} based on data from wave p. The initial estimates used as input for the time series small area models are survey regression estimates (Woodruff, 1966; Battese et al., 1988; Särndal et al., 1992)

$$\hat{\overline{Y}}_{itp} = \overline{y}_{itp} + \hat{\beta}'_{tp} \left(\overline{X}_{it} - \overline{x}_{itp} \right), \tag{3.1}$$

where $\overline{y}_{itp}$, $\overline{x}_{itp}$ denote sample means, $\overline{X}_{it}$ is the vector of population means of the covariates x, and $\hat{\beta}_{tp}$ are estimated regression coefficients. The coefficients are estimated separately for each period and each wave, but they are based on the national samples combining data from all areas. The survey regression estimator is an approximately design-unbiased estimator for the population parameters that, like the GREG estimator, uses auxiliary information to reduce nonresponse bias. See Boonstra and van den Brakel (2016) for more details on the model selected to compute the survey regression estimates. Even though the

regression coefficient estimates in (3.1) are not area-specific, the survey regression estimator is a direct domain estimator in the sense that it is primarily based on the data obtained in that particular domain and month, and therefore it has uncacceptably large standard errors due to the small monthly domain sample sizes.

The initial estimates for the different waves give rise to systematic differences in unemployment estimates, generally termed rotation group bias (RGB) (Bailar, 1975). The initial estimates for unemployement for waves 2 to 5 are systematically smaller compared to the first wave. This RGB has many possible causes, including selection, mode and panel effects (van den Brakel and Krieg, 2009). See Boonstra and van den Brakel (2016) for details and graphical illustrations.

The time series models also require variance estimates corresponding to the initial estimates. We use the following cross-sectionally smoothed estimates of the design variances of the survey regression estimates,

$$v\left(\hat{\bar{Y}}_{itp}\right) = \frac{1}{n_{itp}} \frac{1}{\left(n_{tp} - m_A\right)} \sum_{i=1}^{m_A} \left(n_{itp} - 1\right) \hat{\sigma}_{itp}^2 \equiv \hat{\sigma}_{tp}^2 \,/\, n_{itp}, \quad \text{with} \quad \hat{\sigma}_{itp}^2 = \frac{1}{\left(n_{itp} - 1\right)} \sum_{j=1}^{n_{itp}} \hat{e}_{ijtp}^2. \tag{3.2}$$

Here m_A denotes the number of areas, n_{itp} is the number of respondents in area i, period t and wave p, $n_{tp} = \sum_{i=1}^{m_A} n_{itp}$, and $\hat{e}_{ijtp}$ are residuals of the survey regression estimator. The within-area variances $\hat{\sigma}_{itp}^2$ are pooled over the domains to obtain more stable variance approximations. The use of (3.2) can be further motivated as follows. Recall that the sample design is self-weighted. Calculating within-area variances $\hat{\sigma}_{itp}^2$ therefore approximately accounts for the stratification, which is a slightly more detailed regional variable than province. The variance approximation also accounts for calibration and nonresponse correction, since the within-area variances are calculated over the residuals of the survey regression estimator. The variance approximation does not explicitly account for the clustering of persons within households. However, the intra-cluster correlation for unemployment is small. In addition, registered unemployment is used as a covariate in the survey regression estimator. Since this covariate explains a large part of the variation of unemployment, the intra-cluster correlation between the residuals is further reduced.

The panel design induces several non-zero correlations among initial estimates for the same province and different time periods and waves. These correlations are due to partial overlap of the sets of sample units on which the estimates are based. Such correlations exist between estimates for the same province in months t_1, t_2 and based on waves p_1, p_2 whenever $t_2 - t_1 = 3(p_2 - p_1) \leq 12$. The covariances between $\hat{\bar{Y}}_{it_1p_1}$ and $\hat{\bar{Y}}_{it_2p_2}$ are estimated as (see e.g., Kish (1965))

$$v\left(\hat{\bar{Y}}_{it_1p_1}, \hat{\bar{Y}}_{it_2p_2}\right) = \frac{n_{it_1p_1t_2p_2}}{\sqrt{n_{it_1p_1} n_{it_2p_2}}} \hat{\rho}_{t_1p_1t_2p_2} \sqrt{v\left(\hat{\bar{Y}}_{it_1p_1}\right) v\left(\hat{\bar{Y}}_{it_2p_2}\right)}, \tag{3.3}$$

with

$$\hat{\rho}_{t_1p_1t_2p_2} = \frac{1}{\left(n_{t_1p_1t_2p_2} - m_A\right)} \sum_{i=1}^{m_A} \sum_{j=1}^{n_{it_1p_1t_2p_2}} \hat{e}_{ijt_1p_1} \hat{e}_{ijt_2p_2},$$

where $n_{it_1p_1t_2p_2}$ is the number of units in the overlap, i.e., the number of observations on the same units in area i between period and wave combinations $(t_1,\ p_1)$ and $(t_2,\ p_2)$, and $n_{t_1p_1t_2p_2} = \sum_{i=1}^{m_A} n_{it_1p_1t_2p_2}$. The estimated (auto)correlation coefficient $\hat{\rho}_{t_1p_1t_2p_2}$ is computed as the correlation between the residuals of the linear regression models underlying the survey regression estimators at $(t_1,\ p_1)$ and $(t_2,\ p_2)$, based on the overlap of both samples over all areas. This way they are pooled over areas in the same way as are the variances $\hat{\sigma}^2_{tp}$. Together, (3.2) and (3.3) estimate (an approximation of) the design-based covariance matrix for the initial survey regression estimates. See Boonstra and van den Brakel (2016) for more details.

Time series model estimates for monthly provincial unemployment figures will be compared with direct estimates. The procedure for calculating monthly direct estimates is based on the approach that was used before 2010 to calculate official rolling quarterly figures for the labour force. Let $\hat{\bar{Y}}_{it.}$ denote the monthly direct estimate for provinces, which is calculated as the weighted mean over the five panel survey regression estimates where the weights are based on the variance estimates. To correct for RGB, these direct estimates are multiplied by a ratio, say f_{it}, where the numerator is the mean of the survey regression estimates (3.1) for the first wave over the last three years and the denominator is the mean of monthly direct estimates $\hat{\bar{Y}}_{it.}$ also over the last three years, i.e., $\tilde{\bar{Y}}_{it.} = f_{it}\hat{\bar{Y}}_{it.}$. See Boonstra and van den Brakel (2016) for details on calculating $\hat{\bar{Y}}_{it.}$ and $\tilde{\bar{Y}}_{it.}$ including a variance approximation.

4 Time series small area estimation

The initial monthly domain estimates for the separate waves, accompanied by variance and covariance estimates, are the input for the time series models. In the next step STM models are applied to smooth the initial estimates and correct for RGB. The estimated models are used to make predictions for provincial unemployment fractions, provincial unemployment trends, and month-to-month changes in the trends. In Subsection 4.1 the STMs are defined and subsequently expressed as state space models fitted in a frequentist framework. Subsection 4.2 explains how these STMs can be expressed as time series multilevel models fitted in an hierarchical Bayesian framework.

4.1 State space model

This section develops a structural time series model for the monthly data at provincial level for twelve provinces simultaneously to take advantage of temporal and cross-sectional sample information. Let $\hat{\bar{Y}}_{it} = \left(\hat{\bar{Y}}_{it1}, \ldots, \hat{\bar{Y}}_{it5}\right)^t$ denote the five-dimensional vector containing the survey regression estimates $\hat{\bar{Y}}_{itp}$ defined by (3.1) in period t and domain i. This vector can be modeled with the folowing structural time series model (Pfeffermann, 1991; van den Brakel and Krieg, 2009, 2015):

$$\hat{\bar{Y}}_{it} = \iota_5\theta_{it} + \lambda_{it} + e_{it}, \tag{4.1}$$

where $\iota_5 = (1, 1, 1, 1, 1)^t$, θ_{it} a scalar denoting the true population parameter for period t in domain i, λ_{it} a five-dimensional vector that models the RGB and e_{it} a five-dimensional vector with sampling errors. The population parameter θ_{it} in (4.1) is modeled as

$$\theta_{it} = L_{it} + S_{it} + \epsilon_{it}, \tag{4.2}$$

where L_{it} denotes a stochastic trend model to capture low frequency variation (trend plus business cycle), S_{it} a stochastic seasonal component to model monthly fluctuations and ϵ_{it} a white noise for the unexplained variation in θ_{it}. For the stochastic trend component, the so-called smooth trend model is used, which is defined by the following set of equations:

$$L_{it} = L_{it-1} + R_{it-1},\ R_{it} = R_{it-1} + \eta_{R,it},\ \eta_{R,it} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma^2_{Ri}\right). \tag{4.3}$$

For the stochastic seasonal component the trigonometric form is used, see Boonstra and van den Brakel (2016) for details. The white noise in (4.2) is defined as $\epsilon_{it} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma^2_{\epsilon_i}\right)$.

The RGB between the series of the survey regression estimates, is modeled in (4.1) with $\lambda_{it} = (\lambda_{it1}, \lambda_{it2}, \lambda_{it3}, \lambda_{it4}, \lambda_{it5})^t$. The model is identified by taking $\lambda_{it1} = 0$. This implies that the relative bias in the follow-up waves with respect to the first wave is estimated and it assumes that the survey regression estimates of the first wave are the most reliable approximations for θ_{it}, see van den Brakel and Krieg (2009) for a motivation. The remaining components model the systematic difference between wave p with respect to the first wave and are modeled as random walks to allow for time dependent patterns in the RGB,

$$\lambda_{itp} = \lambda_{it-1;p} + \eta_{\lambda,itp},\ \ \eta_{\lambda,itp} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma^2_{\lambda_i}\right),\ \ p = 2, 3, 4, 5. \tag{4.4}$$

Finally, a time series model for the survey errors is developed. Let $e_{it} = (e_{it1}, e_{it2}, e_{it3}, e_{it4}, e_{it5})^t$ denote the five-dimensional vector containing the survey errors of the five waves. The variance estimates of the survey regression estimates are used as prior information in the time series model to account for heteroscedasticity due to varying sample sizes over time using the following survey error model:

$$e_{itp} = \sqrt{v\left(\hat{\bar{Y}}_{itp}\right)}\ \tilde{e}_{itp}, \tag{4.5}$$

and $v\left(\hat{\bar{Y}}_{itp}\right)$ defined by (3.2). Since the first wave is observed for the first time there is no autocorrelation with samples observed in the past. To model the autocorrelation between survey errors of the follow-up waves, appropriate AR models for $\tilde{e}_{itp}$, are derived by applying the Yule-Walker equations to the correlation coefficients

$$\frac{n_{it_1p_1t_2p_2}}{\sqrt{n_{it_1p_1}n_{it_2p_2}}}\hat{\rho}_{t_1p_1t_2p_2}, \tag{4.6}$$

which are derived from the micro data as described in Section 3. Based on this analysis an AR(1) model is assumed for wave 2 through 5 where the autocorrelation coefficients depend on wave and month. These considerations result in the following model for the survey errors:

$$\begin{aligned}
\tilde{e}_{it1} &= \nu_{it1},\ \ \nu_{it1} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma^2_{\nu_{i1}}\right), \\
\tilde{e}_{itp} &= \varrho_{it(p-1)p}\tilde{e}_{i(t-3)(p-1)} + \nu_{itp},\ \nu_{itp} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma^2_{\nu_{ip}}\right),\ \ p = 2, \ldots, 5,
\end{aligned} \tag{4.7}$$

with $\varrho_{it(p-1)p}$ the time-dependent partial autocorrelation coefficients between wave p and $p-1$ derived from (4.6). As a result, $\text{Var}(e_{it1}) = v\left(\hat{\bar{Y}}_{it1}\right)\sigma^2_{\nu_{i1}}$, and $\text{Var}(e_{itp}) = v\left(\hat{\bar{Y}}_{itp}\right)\sigma^2_{\nu_{ip}} \Big/ \left(1-\varrho^2_{it(p-1)p}\right)$ for $p = 2, \ldots, 5$. The variances $\sigma^2_{\nu_{ip}}$ are scaling parameters with values close to one for the first wave and close to $\frac{1}{T}\sum_{t=1}^{T}\left(1-\varrho^2_{it(p-1)p}\right)$ for the other waves, where T denotes the length of the observed series.

Model (4.1) uses sample information observed in preceding periods within each domain to improve the precision of the survey regression estimator and accounts for RGB and serial correlation induced by the rotating panel design. To take advantage of sample information across domains, model (4.1) for the separate domains can be combined in one multivariate model:

$$\begin{pmatrix} \hat{\bar{Y}}_{1t} \\ \vdots \\ \hat{\bar{Y}}_{m_A t} \end{pmatrix} = \begin{pmatrix} \iota_5\theta_{1t} \\ \vdots \\ \iota_5\theta_{m_A t} \end{pmatrix} + \begin{pmatrix} \lambda_{1t} \\ \vdots \\ \lambda_{m_A t} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ \vdots \\ e_{m_A t} \end{pmatrix}, \tag{4.8}$$

where m_A denotes the number of domains, which is equal to twelve in this application. This multivariate setting allows to use sample information across domains by modeling the correlation between the disturbance terms of the different structural time series components (trend, seasonal, RGB) or by defining the hyperparameters or the state variables of these components equal over the domains. In this paper models with cross-sectional correlation between the slope disturbance terms of the trend (4.3) are considered, i.e.,

$$\text{Cov}\left(\eta_{R,it}, \eta_{R,i't'}\right) = \begin{cases} \sigma^2_{Ri} & \text{if } i = i' \text{ and } t = t' \\ \varsigma_{Rii'} & \text{if } i \neq i' \text{ and } t = t' \\ 0 & \text{if } t \neq t' \end{cases}. \tag{4.9}$$

The most parsimonious covariance structure is a diagonal matrix where all the domains share the same variance component, i.e., $\sigma^2_{Ri} = \sigma^2_R$ for all i and $\varsigma_{Rii'} = 0$ for all i and i'. These are so-called seemingly unrelated structural time series models and are a synthetic approach to use sample information across domains. A slightly more complex and realistic covariance structure is a diagonal matrix where each domain has a separate variance component, i.e., $\varsigma_{Rii'} = 0$ for all i and i'. In this case the model only borrows strength over time and does not take advantage of cross-sectional information. The most complex covariance structure allows for a full covariance matrix. Strong correlation between the slope disturbances across the domains can result in cointegrated trends. This implies that $q < m_A$ common trends are required to model the dynamics of the trends for the m_A domains and allows the specification of so-called common trend models (Koopman, Harvey, Doornik and Shephard, 1999; Krieg and van den Brakel, 2012). Initial STM analyses showed that the seasonal and RGB component turned out to be time independent. It is therefore not sensible to model correlations between seasonal and RGB disturbance terms. Since the hyperparameters of the white noise population domain parameters tend to zero, it turned out to be better to remove this component completely from the model implying that modeling correlations between population noise is not considered. Correlations between survey errors for different domains is also not considered, since the domains are geographical regions from which samples are drawn independently.

As an alternative to a model with a full covariance matrix for the slope disturbances, a trend model is considered that has one common smooth trend model for all provinces plus $m_A - 1$ trend components that describe the deviation of each domain from this overall trend. In this case (4.2) is given by

$$\begin{aligned} \theta_{1t} &= L_t + S_{1t} + \epsilon_{1t}, \\ \theta_{it} &= L_t + L_{it}^* + S_{it} + \epsilon_{it}, \quad i = 2, \ldots, m_A. \end{aligned} \tag{4.10}$$

Here L_t is the overal smooth trend component, defined by (4.3), and L_{it}^* the deviation from the overall trend for the separate domains, defined as local levels

$$L_{it}^* = L_{it-1}^* + \eta_{L,it}, \ \eta_{L,it} \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma_{Li}^2\right), \tag{4.11}$$

or as smooth trends as in (4.3). These trend models implicitly allow for (positive) correlations between the trends of the different domains.

The parameters to be estimated with the time series modeling approach are the trend and the signal. The latter is defined as the trend plus the seasonal component. The time series approach is particularly suitable for estimating month-to-month changes. Seasonal patterns hamper a straightforward interpretation of month-to-month changes of direct estimates and smoothed signals. Therefore month-to-month changes are calculated for the trends only. Due to the strong positive correlation between the levels of consecutive periods, the standard errors of month-to-month changes in the level of the trends are much smaller than those of e.g., month-to-month changes of the direct estimates. The month-to-month change of the trend is defined as $\Delta_{it}(1) = L_{it} - L_{it-1}$ for models with separate trends for the domains or $\Delta_{it}(1) = L_t - L_{t-1} + L_{it}^* - L_{it-1}^*$ for models with an overall trend and $m_A - 1$ trends for the deviation from the overall trend for the separate domains. This modeling approach is also useful to estimate year-to-year developments for trend defined as $\Delta_{it}(12) = L_{it} - L_{it-12}$ or $\Delta_{it}(12) = L_t - L_{t-12} + L_{it}^* - L_{it-12}^*$. Year-to-year differences are also sensible for signals, since the main part of the seasonal component cancels out. These developments are defined equivalently to the year-to-year developments of the trend.

The aforementioned structural time series models are analyzed by putting them in the so-called state space form. Subsequently the Kalman filter is used to fit the models, where the unknown hyperparameters are replaced by their ML estimates. The analysis is conducted with software developed in OxMetrics in combination with the subroutines of SsfPack 3.0, (Doornik, 2009; Koopman, Shephard and Doornik, 1999, 2008). ML estimates for the hyperparameters are obtained using the numerical optimization procedure maxBFGS in OxMetrics. More details about the state space representation, initialization of the Kalman filter and software used to fit these models is included in Boonstra and van den Brakel (2016).

4.2 Time series multilevel model

For the description of the multilevel time series representation of the STMs, the initial estimates $\hat{\bar{Y}}_{itp}$ are combined into a vector $\hat{\bar{Y}} = \left(\hat{\bar{Y}}_{111}, \hat{\bar{Y}}_{112}, \ldots, \hat{\bar{Y}}_{115}, \hat{\bar{Y}}_{121}, \ldots\right)'$, i.e., wave index runs faster than time index

which runs faster than area index. The numbers of areas, periods and waves are denoted by m_A, m_T and m_P, respectively. The total length of $\hat{\bar{Y}}$ is therefore $m = m_A m_T m_P$ = 12(areas) * 72(months) * 5(waves) = 4,320. Similarly, the variance estimates $v\left(\hat{\bar{Y}}_{itp}\right)$ are put in the same order along the diagonal of a $m \times m$ covariance matrix Φ.

The covariance matrix Φ is not diagonal because of the correlations induced by the panel design. It is a sparse band matrix, and the ordering of the vector $\hat{\bar{Y}}$ is such that it achieves minimum possible bandwidth, which is advantageous from a computational point of view.

The multilevel models considered for modeling the vector of direct estimates $\hat{\bar{Y}}$, take the general linear additive form

$$\hat{\bar{Y}} = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)} + e, \tag{4.12}$$

where X is a $m \times p$ design matrix for the fixed effects β, and the $Z^{(\alpha)}$ are $m \times q^{(\alpha)}$ design matrices for random effect vectors $v^{(\alpha)}$. Here the sum over α runs over several possible random effect terms at different levels, such as a national level smooth trend, provincial local level trends, white noise, etc. This is explained in more detail below. The sampling errors $e = (e_{111}, e_{112}, \ldots, e_{115}, e_{121}, \ldots)'$ are taken to be normally distributed as

$$e \sim \mathcal{N}(0, \Sigma) \tag{4.13}$$

where $\Sigma = \oplus_{i=1}^{m_A} \lambda_i \Phi_i$ with Φ_i the covariance matrix for the initial estimates for province i, and λ_i a province-specific variance scale parameter to be estimated. As described in Section 3 the design variances in $\Phi = \oplus_i \Phi_i$ are pooled over provinces and because of the discrete nature of the unemployment data they thereby lose some of their dependence on the unemployment level. It was found that incorporating the variance scale factors λ_i allows the model to rescale the estimated design variances to a level that better fits the data.

To describe the general model for each vector $v^{(\alpha)}$ of random effects, we suppress the superscript α. Each vector v has $q = dl$ components corresponding to d effects allowed to vary over l levels of a factor variable. In particular,

$$v \sim \mathcal{N}(0, A \otimes V), \tag{4.14}$$

where V and A are $d \times d$ and $l \times l$ covariance matrices, respectively. As in Section 4.1 the covariance matrix V is allowed to be parameterised in three different ways. Most generally, it is an unstructured, i.e., fully parameterised covariance matrix. More parsimonious forms are $V = \text{diag}\left(\sigma^2_{v;1}, \ldots, \sigma^2_{v;d}\right)$ or $V = \sigma^2_v I_d$. If $d = 1$ the three parameterisations are equivalent. The covariance matrix A describes the covariance structure between the levels of the factor variable, and is assumed to be known. It is typically more convenient to use the precision matrix $Q_A = A^{-1}$ as it is sparse for many common temporal and spatial correlation structures (Rue and Held, 2005).

4.2.1 Relations between state space and time series multilevel representations

A single smooth trend can be represented as a random intercept $(d = 1)$ varying over time $(l = m_T)$, with temporal correlation determined by a $m_T \times m_T$ band sparse precision matrix Q_A associated with a second order random walk (Rue and Held, 2005). In this case $V = \sigma_v^2$ and the design matrix Z is the $m \times m_T$ indicator matrix for month, i.e., the matrix with a single 1 in each row for the corresponding month and 0s elsewhere. The sparsity of both Q_A and Z can be exploited in computations. The precision matrix for the smooth trend component has two singular vectors, $\iota_{m_T} = (1, 1, \ldots, 1)$ and $(1, 2, \ldots, m_T)'$. This means that the corresponding specification (4.14) is completely uninformative about the overall level and linear trend. In order to prevent unidentifiability among various terms in the model, the overall level and trend can be removed from v by imposing the constraints $Rv = 0$, where R is the $2 \times m_T$ matrix with the two singular vectors as its rows. The overall level and trend are then included in the vector β of fixed effects. In the state space representation, this model is obtained by defining one trend model (4.3) for all domains, i.e., $L_{it} = L_t$ and $R_{it} = R_t$ for all i. Defining the state variables for the trend equal over the domains is a very synthetic approach to use sample information from other domains and is based on assumptions that are not met in most cases.

A smooth trend for each province is obtained with $d = m_A$, $l = m_T$, and V a $m_A \times m_A$ covariance matrix, either diagonal with a single variance parameter, diagonal with m_A variance parameters, or unstructured, i.e., fully parametrised in terms of m_A variance parameters and $m_A(m_A - 1)/2$ correlation parameters. The design matrix is $I_{m_A} \otimes I_{m_T} \otimes \iota_{m_P}$ in this case. In the state space representation, these models are obtained with trend model (4.3) and covariance structure (4.9).

An alternative trend model consists of a single global smooth trend (second order random walk) supplemented by a local level trend, i.e., an ordinary (first order) random walk, for each province. The latter can be modeled as discussed in the previous paragraph, but with precision matrix associated with a first order random walk. This trend model corresponds to the models (4.10) and (4.11) in the state space context. In contrast to the state space approach, it is not necessary to remove one of the provincial random walk trends from the model for identifiability. The reason is that in the multilevel approach constraints are imposed to ensure that the smooth overall trend as well as all provincial random walk trends sum to zero over time. The constrained components correspond to global and provincial intercepts, which are separately included in the model as fixed effects with one provincial fixed effect excluded.

Seasonal effects can be expressed in terms of correlated random effects (4.14) as well. The trigonometric seasonal is equivalent to the balanced dummy variable seasonal model (Proietti, 2000; Harvey, 2006), corresponding to first order random walks over time for each month, subject to a sum-to-zero constraint over the months. In this case $d = 12$ (seasons), $V = \sigma_v^2 I_{12}$, and $l = m_T$ with Q_A the precision matrix of a first order random walk. The sum-to-zero constraints over seasons at each time, together with the sum-to-zero constraints over time of each random walk can be imposed as $Rv = 0$ with R the $(m_T + 12) \times 12m_T$ matrix

$$R = \begin{pmatrix} \iota'_{12} \otimes I_{m_T} \\ I_{12} \otimes \iota'_{m_T} \end{pmatrix}. \quad (4.15)$$

Together with fixed effects for each season (again with a sum-to-zero constraint imposed) this random effect term is equivalent to the trigonometric seasonal. It can be extended to a seasonal for each province, with a separate variance parameter for each province.

To account for the RGB, the multilevel model includes fixed effects for waves 2 to 5. These effects can optionally be modeled dynamically by adding random walks over time for each wave. Another choice to be made is whether the fixed and random effects are crossed with province.

Further fixed effects can be included in the model, for example those associated with the auxiliary variables used in the survey regression estimates. Some fixed effect interactions, for example season × province or wave × province might alternatively be modeled as random effects to reduce the risk of overfitting.

Finally, a white noise term can be added to the model, to account for unexplained variation by area and time in the signal.

Model (4.12) can be regarded as a generalization of the Fay-Herriot area-level model. The Fay-Herriot model only includes a single vector of uncorrelated random effects over the levels of a single factor variable (typically areas). The models used in this paper contain various combinations of uncorrelated and correlated random effects over areas and months. Earlier accounts of multilevel time series models extending the Fay-Herriot model are Rao and Yu (1994); Datta et al. (1999); You (2008). In Datta et al. (1999) and You (2008) time series models are used with independent area effects and first-order random walks over time for each area. In Rao and Yu (1994) a model is used with independent random area effects and a stationary autoregressive AR(1) instead of a random walk model over time. In You et al. (2003) the random walk model was found to fit the Canadian unemployment data slightly better than AR(1) models with autocorrelation parameter fixed at 0.5 or 0.75. We do not consider AR(1) models in this paper, and refer to Diallo (2014) for an approach that allows both stationary and non-stationary trends. Compared to the aforementioned references a novel feature of our model is that smooth trends are considered instead of or in addition to first-order random walks or autoregressive components. We also include independent area-by-time random effects as a white noise term accounting for unexplained variation at the aggregation level of interest.

4.2.2 Estimating time series multilevel models

A Bayesian approach is used to fit model (4.12)-(4.14). This means we need prior distributions for all (hyper)parameters in the model. The following priors are used:

- The data-level variance parameters λ_i for $i = 1, \ldots, m_A$ are assigned inverse chi-squared priors with degrees of freedom and scale parameters equal to 1.

- The fixed effects are assigned a normal prior with zero mean and fixed diagonal variance matrix with very large values (1e10).
- For a fully parameterized covariance matrix V in (4.14) we use the scaled-inverse Wishart prior as proposed in O'Malley and Zaslavsky (2008) and recommended by Gelman and Hill (2007). Conditionally on a d-dimensional vector parameter ξ,

$$V \mid \xi \sim \text{Inv} - \text{Wishart}\,(V \mid \nu,\ \text{diag}\,(\xi)\,\Psi\text{diag}\,(\xi)) \tag{4.16}$$

  where $\nu = d + 1$ is chosen, and $\Psi = I_d$. The vector ξ is assigned a normal distribution $\mathcal{N}\,(0,\ I_d)$.
- All other variance parameters appearing in a diagonal matrix V in (4.14) are assigned, conditionally on an auxiliary parameter ξ, inverse chi-squared priors with 1 degree of freedom and scale parameter ξ^2. Each parameter ξ is assigned a $\mathcal{N}\,(0, 1)$ prior. Marginally, the standard deviation parameters have half-Cauchy priors. Gelman (2006) demonstrates that these priors are better default priors than the more common inverse chi-squared priors.

The model is fit using Markov Chain Monte Carlo (MCMC) sampling, in particular the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). The multilevel models considered belong to the class of additive latent Gaussian models with random effect terms being Gaussian Markov Random Fields (GMRFs), and we make use of the sparse matrix and block sampling techniques described in Rue and Held (2005) for efficiently fitting such models to the data. Moreover, the parametrization in terms of the aforementioned auxiliary parameters ξ (Gelman, Van Dyk, Huang and Boscardin, 2008), greatly improves the convergence of the Gibbs sampler used. See Boonstra and van den Brakel (2016) for more details on the Gibbs sampler used, including specifications of the full conditional distributions. The methods are implemented in R using the *mcmcsae* R-package (Boonstra, 2016).

For each model considered, the Gibbs sampler is run in three independent chains with randomly generated starting values. Each chain is run for 2,500 iterations. The first 500 draws are discarded as a "burn-in sample". From the remaining 2,000 draws from each chain, we keep every fifth draw to save memory while reducing the effect of autocorrelation between successive draws. This leaves 3 * 400 = 1,200 draws to compute estimates and standard errors. It was found that the effective number of independent draws was near 1,200 for most model parameters, meaning that most autocorrelation was indeed removed by the thinning. The convergence of the MCMC simulation is assessed using trace and autocorrelation plots as well as the Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992), which diagnoses the mixing of the chains. The diagnostics suggest that all chains converge well within the burnin stage, and that the chains mix well, since all Gelman-Rubin factors are close to one. Also, the estimated Monte Carlo simulation errors (accounting for any remaining autocorrelation in the chains) are small compared to the posterior standard errors for all parameters, so that the number of retained draws is sufficient for our purposes.

The estimands of interest can be expressed as functions of the parameters, and applying these functions to the MCMC output for the parameters results in draws from the posteriors for these estimands. In this paper we summarize those draws in terms of their mean and standard deviation, serving as estimates and standard errors, respectively. All estimands considered can be expressed as linear predictors, i.e., as linear combinations of the model parameters. Estimates and standard errors for the following estimands are computed:

- Signal: the vector θ_{it} including all fixed and random effects, except those associated with waves 2 to 5. These correspond to the fitted values $X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)}$ associated with each fifth row 1, 6, 11, … of $\hat{\tilde{Y}}$ and the design matrices.
- Trend: prediction of the long-term trend. This is computed by only incorporating the trend components of each model in the linear predictor. For most models considered the trend corresponds to seasonally adjusted figures, i.e., predictions of the signal with all seasonal effects removed.
- Growth of trend: the differences between trends at two consecutive months.

5 Results

The results obtained with the state space and multilevel time series representations of the STMs are described in Subsections 5.1 and 5.2, respectively. First, two discrepancy measures are defined to evaluate and compare the different models. The first measure is the Mean Relative Bias (MRB), which summarizes the differences between model estimates and direct estimates averaged over time, as percentage of the latter. For a given model M, the MRB_i is defined as

$$\text{MRB}_i = \frac{\sum_t \left(\hat{\theta}_{it}^M - \tilde{\tilde{Y}}_{it.}\right)}{\sum_t \tilde{\tilde{Y}}_{it.}} \times 100\%, \tag{5.1}$$

where $\tilde{\tilde{Y}}_{it.}$ are the direct estimates by province and month incorporating the ratio RGB adjustment mentioned at the end of Section 3. This benchmark measure shows for each province how much the model-based estimates deviate from the direct estimates. The discrepancies should not be too large as one may expect that the direct estimates averaged over time are close to the true average level of unemployment. The second discrepancy measure is the Relative Reduction of the Standard Errors (RRSE) and measures the percentages of reduction in estimated standard errors between model-based and direct estimates, i.e.,

$$\text{RRSE}_i = 100\% \times \frac{1}{m_T} \sum_t \left(\text{se}\left(\tilde{\tilde{Y}}_{it.}\right) - \text{se}\left(\hat{\theta}_{it}^M\right)\right) \Big/ \text{se}\left(\tilde{\tilde{Y}}_{it.}\right), \tag{5.2}$$

for a given model M. Here the estimated standard errors for the direct estimates follow from a variance approximation for $\tilde{\tilde{Y}}_{it.}$, whereas the model-based standard errors are posterior standard deviations or follow from the Kalman filter/smoother. Posterior standard deviations, standard errors obtained via the Kalman filter and standard errors of the direct estimators come from different frameworks and are formally spoken

not comparable. They are used in (5.2) to quantify the reduction with respect to the direct estimator only but not intended as model selection criteria.

5.1 Results state space models

Ten different state space models are compared. Four different trend models are distinguished. The first trend component is a smooth trend model without correlations between the domains (4.3), abbreviated as T1. The second trend model, T2, is a smooth trend model (4.3) with a full correlation matrix for the slope disturbances (4.9). The third trend component, T3, is a common smooth trend model for all provinces with eleven local level trend models for the deviation of the domains from this overall trend ((4.10) in combination with (4.11)). The fourth trend model, T4, is a common smooth trend model for all provinces with eleven smooth trend models for the deviation of the domains from this overall trend ((4.10) in combination with (4.3)). In T3 and T4 the province Groningen is taken equal to the overall trend. The component for the RGB (4.4) can be domain specific (indicated by letter "R" in the model's name) or chosen equal for all domains (no "R" in the model's name). An alternative simplicfication is to assume that RGB for waves 2, 3, 4 and 5 are equal but domain specific (indicated by "R2"). In a similar way the seasonal component can be chosen domain specific (indicated by "S") or taken equal for all domains. All models share the same component for the survey error, i.e., an AR(1) model with time varying autocorrelation coefficients for wave 2 through 5 to model the autocorrelation in the survey errors. The following state space models are compared:

T1SR: Smooth trend model and no correlation between slope disturbances; seasonal and RGB domain specific.

T2SR: Smooth trend model with a full correlation matrix for the slope disturbances; seasonal and RGB domain specific.

T2S: Smooth trend model with a full correlation matrix for the slope disturbances; seasonal domain specific, RGB equal over all domains.

T2R: Smooth trend model with a full correlation matrix for the slope disturbances; seasonal equal over all domains, RGB domain specific.

T3SR: One common smooth trend model for all domains plus eleven local levels for deviations from the overall trend; seasonal and RGB domain specific.

T3R: One common smooth trend model for all domains plus eleven local levels for deviations from the overall trend; seasonal equal over all domains, RGB domain specific.

T3R2: One common smooth trend model for all domains plus eleven local levels for deviations from the overall trend; seasonal equal over all domains, RGB is domain specific but assumed to be equal for the four follow-up waves.

T3: One common smooth trend model for all domains plus eleven local levels for deviations from the overall trend; seasonal and RGB equal over all domains.

T4SR: One common smooth trend model for all domains plus eleven smooth trend models for deviations from the overall trend; seasonal and RGB domain specific.

T4R: One common smooth trend model for all domains plus eleven smooth trend models for deviations from the overall trend; seasonal equal over all domains, RGB domain specific.

For all models, the ML estimates for the hyperparameters of the RGB and the seasonals tend to zero, which implies that these components are time invariant. Also the ML estimates for the variance components of the white noise of the population domain parameters tend to zero. This component is therefore removed from model (4.2). The ML estimates for the variance components of the survey errors in the first wave vary between 0.93 and 1.90. For the follow-up waves, the ML estimates vary between 0.86 and 1.80. The variances of the direct estimates are pooled over the domains (3.2), which might introduce some bias, e.g., underestimation of the variance in domains with high unemployment rates. Scaling the variances of the survey errors with the ML estimates for $\sigma^2_{v_{ip}}$ is neccessary to correct for this bias. The ML estimates for the hyperparameters for the trend components can be found in Boonstra and van den Brakel (2016).

Models are compared using the log likelihoods. To account for differences in model complexity, Akaike Information Criteria (AIC) and Bayes Information Criteria (BIC) are used, see Durbin and Koopman (2012), Section 7.4. Results are summarized in Table 5.1. Parsimonious models where the seasonals or RGB are equal over the domains are preferred by the AIC or BIC criteria. Note, however, that the likelihoods are not completely comparable between models. To obtain comparable likelihoods, the first 24 months of the series are ignored in the computation of the likelihood for all models. Some of the likelihoods are nevertheless odd. For example the likelihood of T2SR is smaller than the likelihood of T2S, although T2SR contains more model parameters. This is probably the result of large and complex time series models in combination with relatively short time series, which gives rise to flat likelihood functions. Also from this point of view, sparse models that avoid over-fitting are still favorable, which is in line with the results of the AIC and BIC values in Table 5.1.

Table 5.1
AIC and BIC for the state space models

Model	log likelihood	states	hyperparameters	AIC	BIC
T1SR	9,813.82	204	24	-399.41	-390.52
T2SR	9,862.86	204	35	-400.99	-391.68
T2S	9,879.03	160	35	-403.50	-395.90
T2R	9,859.97	83	35	-405.92	-401.32
T3SR	9,855.35	193	24	-401.60	-393.14
T3R	9,851.62	72	24	-406.48	-402.74
T3R2	9,871.65	36	24	-408.82	-406.48
T3	9,881.16	28	24	–409.55	-407.52
T4SR	9,857.47	204	24	-401.23	-392.34
T4R	9,853.65	83	24	-406.11	-401.94

Modeling correlations between slope disturbances of the trend results in a significant model improvement. Model T1SR, e.g., is nested within T2SR and a likelihood ratio test clearly favours the latter. For model T2SR it follows that the dynamics of the trends for these 12 domains can be modeled with only 2 underlying common trends, since the rank of the 12×12 covariance matrix equals two. As a result the full covariance matrix for the slope disturbances of the 12 domains is actually modeled with 23 instead of

78 hyperparameters. This shows that the correlations between the slope disturbances are very strong. Correlations indeed vary between 1.00 and 0.98. See Boonstra and van den Brakel (2016) for the ML estimates of the full covariance matrix.

Table 5.2 shows the MRB, defined by (5.1). Models that assume that the RGB is equal over the domains, i.e., T2S and T3, have large relative biases for some of the domains. Large biases occur in the domains where unemployment is large (e.g., Groningen) or small (e.g., Utrecht) compared to the national average. A possible compromise between parsimony and bias is to assume that the RGB is equal for the four follow-up waves but still domain specific (T3R2). For this model the bias is small, with the exception of Gelderland.

Table 5.2
Mean Relative Bias averaged (5.1) over time (%), per province for state space models

	Grn	Frs	Drn	Ovr	Flv	Gld	Utr	N-H	Z-H	Zln	N-B	Lmb
T1SR	1.1	0.5	2.0	-0.2	0.1	3.4	0.1	0.6	1.7	-2.1	0.5	2.1
T2SR	1.2	0.7	2.2	-0.1	0.2	3.5	0.2	0.6	1.7	-2.1	0.5	2.1
T2S	-3.1	3.1	0.7	0.9	-4.4	2.8	2.4	0.8	0.5	1.7	1.8	1.5
T2R	0.9	0.8	1.8	-0.2	-0.4	3.4	0.1	0.6	1.7	-1.6	0.6	2.2
T3SR	0.8	0.6	2.0	-0.2	-0.3	3.5	0.3	0.5	1.7	-2.0	0.6	2.0
T3R2	-0.1	1.3	2.1	-0.6	-0.8	3.6	0.9	0.6	1.5	-1.1	1.0	1.2
T3R	0.5	0.7	1.8	-0.2	-0.8	3.5	0.3	0.5	1.6	-1.5	0.7	2.1
T3	-4.0	2.5	0.1	0.9	-5.0	2.8	2.3	0.7	0.6	2.5	2.0	1.3
T4SR	0.8	0.7	2.1	-0.2	-0.0	3.5	0.2	0.6	1.7	-1.9	0.5	2.1
T4R	0.6	0.7	1.8	-0.2	-0.6	3.4	0.1	0.6	1.7	-1.3	0.7	2.1

In Figure 5.1 the smoothed trends and standard errors of models T1SR, T2SR and T2S are compared. The month-to-month development of the trend and the standard errors for these three models are compared in Figure 5.2. The smoothed trends obtained with the common trend model are slightly more flexible compared to a model without correlation between the slope disturbances. This is clearly visible in the month-to-month change of the trends. Modeling the correlation between slope disturbances clearly reduces the standard error of the trend and the month-to-month change of the trend. Assuming that the RGB is equal for all domains (model T2S) affects the level of the trend and further reduces the standard error, mainly since the number of state variables are reduced. The difference between the trend under T2SR and T2S is a level shift. This follows from the month-to-month changes of the trend under model T2SR and T2S, which are exactly equal. According to AIC and BIC the reduction of the number of state variables by assuming equal RGB for all domains is an improvement of the model. In this application, however, interest is focused on the model fit for the separate domains. Assuming that the RGB is equal over all domains is on average efficient for overall goodness of fit measures, like AIC and BIC, but not necessarily for all separate domains. The bias introduced in the trends of some of the domains by taking the RGB equal over the domains is undesirable.

In Figure 5.3 the smoothed trends and standard errors of models T2SR, T3SR and T4SR are compared. The month-to-month developments of the trend and the standard errors can be found in Boonstra and van den Brakel (2016). The trends obtained with one overall smooth trend plus eleven trends for the domain

deviations of the overall trend resemble trends obtained with the common trend model. In this application the dynamics based on the two common trends of model T2SR are reasonably well approximated by the alternative trends of models T3SR and T4SR. This is an empirical finding that may not generalize to other situations, particularly when more common factors are required. The common trend model, however, has the smallest standard errors for the trend. Furthermore, the trends under the model with a local level for the domain deviations from the overall trend are in some domains more volatile compared to the other two models. This is most obvious in the month-to-month changes of the trend. It is a general feature for trend models with random levels to have more volatile trends, see Durbin and Koopman (2012), Chapter 3. The more flexible trend model of T3 also results in a higher standard error of the month-to-month changes.

Assuming that the seasonals are equal for all domains is another way of reducing the number of state variables and avoid over-fitting of the data. This assumption does not affect the level of the trend since the MRB is small (see Table 5.2) and results in a significant improvement of the model according to AIC and BIC. Particularly if interest is focused on trend estimates, some bias in the seasonal patterns is acceptable and a model with a trend based on T2, or T4, with the seasonal component assumed equal over the domains, might be a good compromise between a model that accounts sufficiently for differences between domains and model parsimony to avoid over-fitting of the data.

Model T3 is the most parsimonious model that is the best model according to AIC and BIC. Particularly the assumption of equal RGB results in biased trend estimates in some of the domains (see Table 5.2). See Boonstra and van den Brakel (2016) for a comparison of the trend and the month-to-month development of the trend of models T2R, T3 and T4R. Assuming that the seasonals are equal over the domains, results in a less pronounced seasonal pattern. See Boonstra and van den Brakel (2016) for a comparison of the signals for models T2SR and T2R.

In Boonstra and van den Brakel (2016) results for year-to-year change of the trends under models T2R and T3R2 are included. Time series estimates for year-to-year change are very stable and precise and greatly improve the direct estimates for year-to-year change.

Table 5.3 shows the RRSE, defined by (5.2), for the ten state space models. Recall that the RRSE quantifies the reduction with respect to the direct estimator and is not intended as a model selection criterion. Table 5.4 contains the averages of standard errors for signal, trend, and growth (month-to-month differences of trend). The average is taken over all months and provinces. Modeling the correlation between the trends explicitly (T2) or implicitly (T3 or T4) reduces the standard errors for the trend and signal significantly. The time series modeling approach is particularly appropriate to estimate month-to-month changes through the trend component. The precision of the month-to-month changes, however, strongly depends on the choice of the trend model. A local level trend model (T3) results in more volatile trends and has a clearly larger standard error for the month-to-month change. Parsimonious models where RGB or the seasonal components are assumed equal over the domains result in further strong standard error reductions at the cost of introducing bias in the trend or the seasonal patterns.

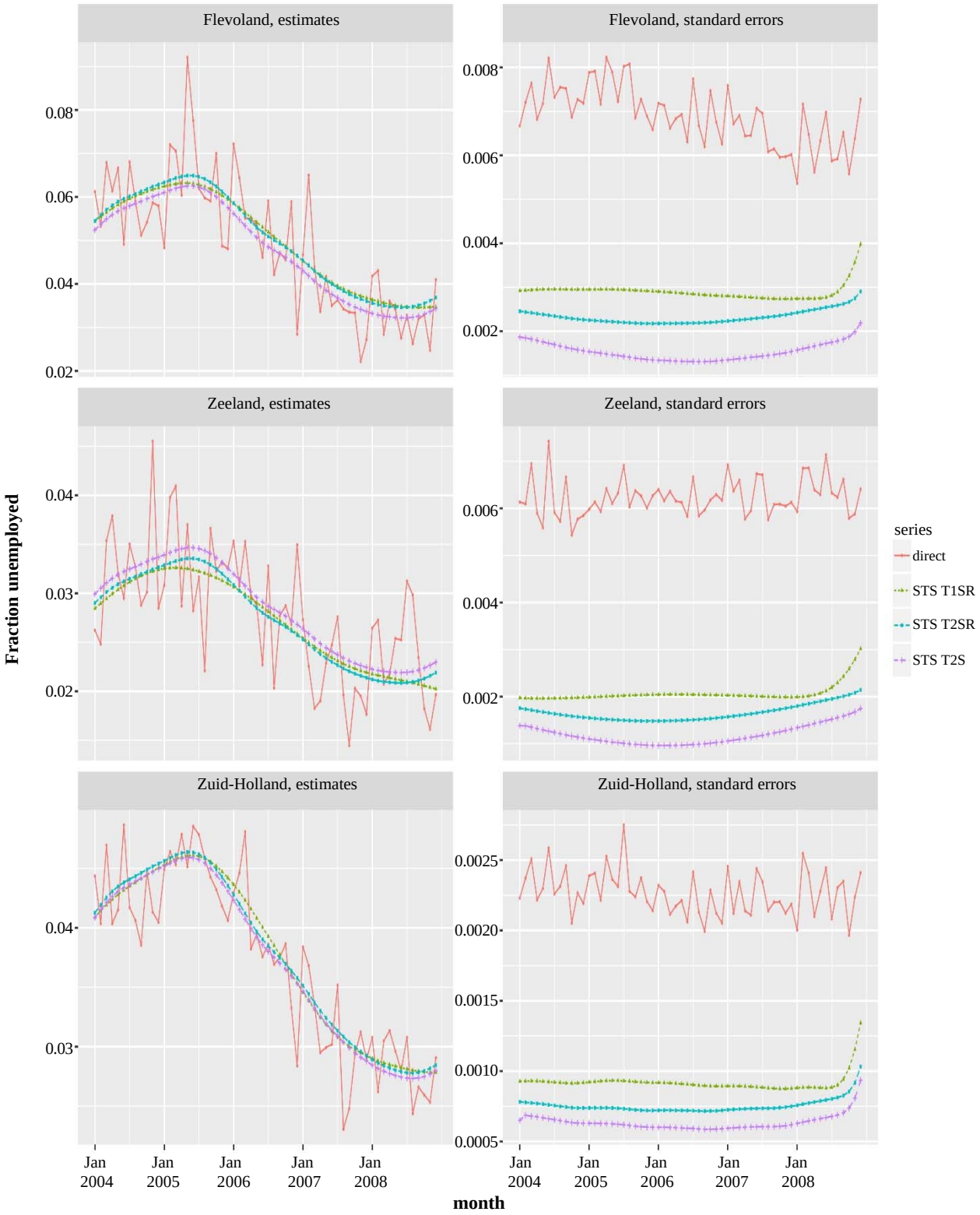


Figure 5.1 Comparison of direct estimates and smoothed trend estimates for three models (left) and their estimated standard errors (right).

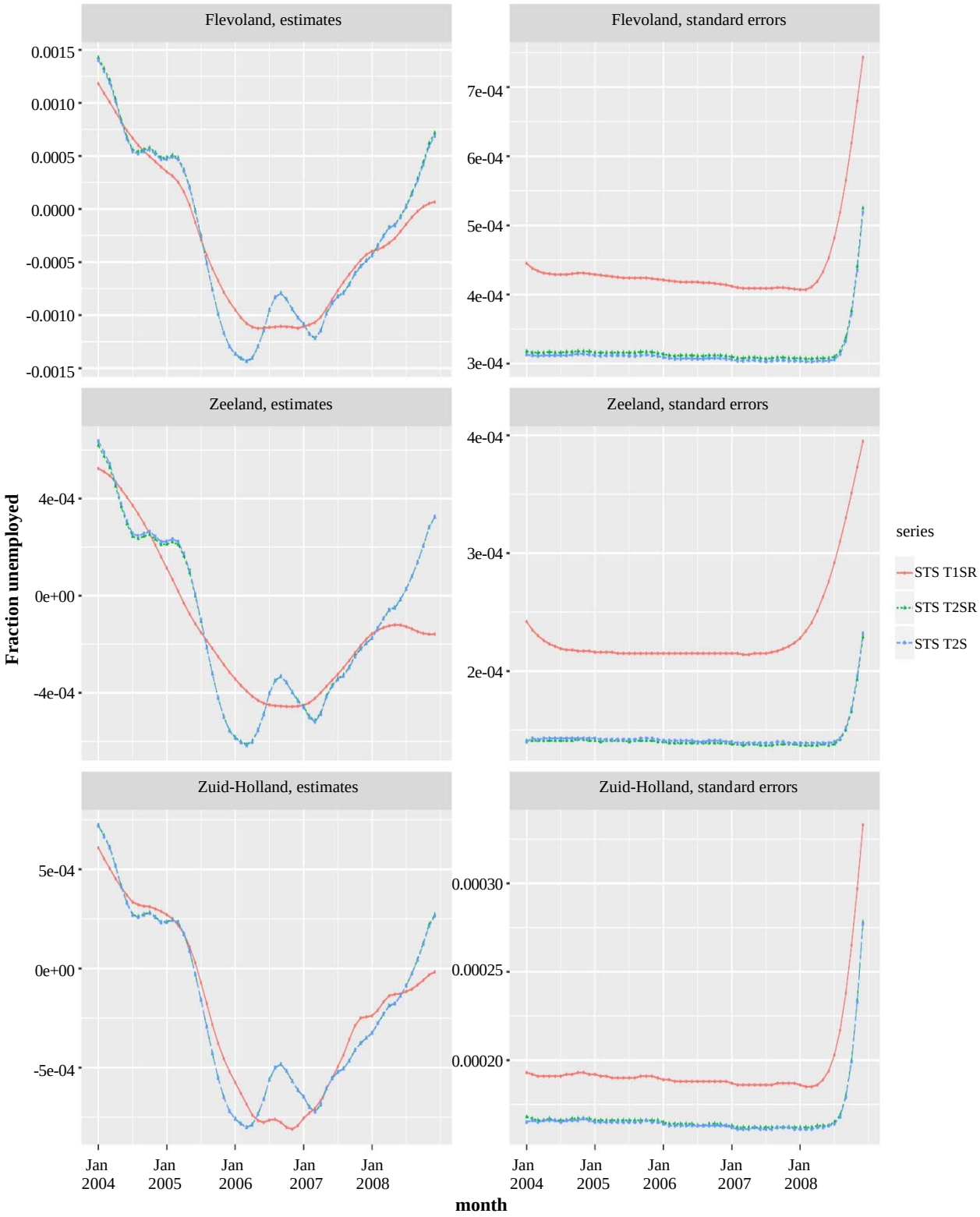


Figure 5.2 Comparison of smoothed month-to-month developments (left) and their standard errors (right).

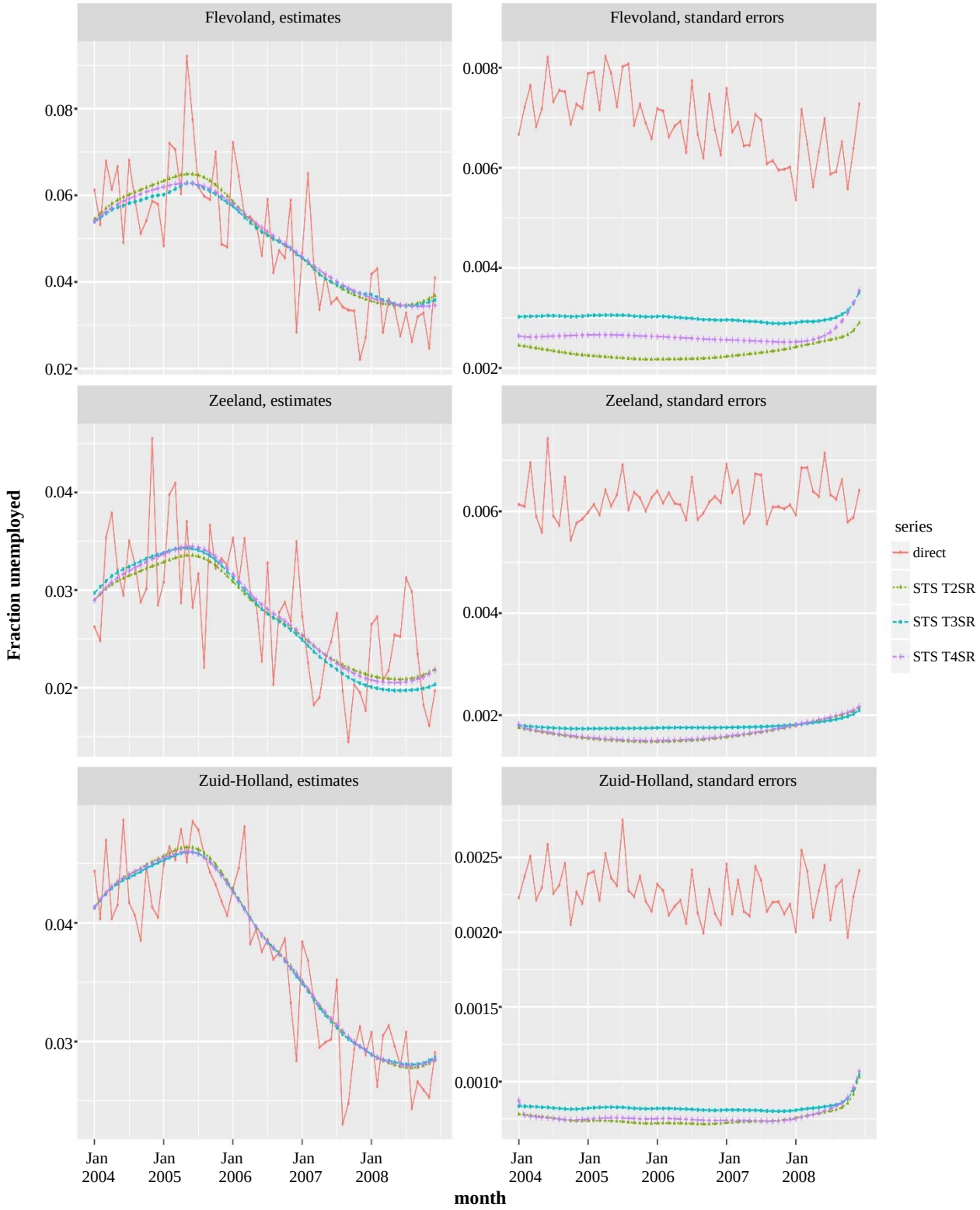


Figure 5.3 Comparison of direct estimates and smoothed trend estimates for three models (left) and their estimated standard errors (right).

Table 5.3
Relative reductions in standard errors (5.2) of the signal estimates based on the state space models compared to those of the direct estimates (%), per province

	Grn	Frs	Drn	Ovr	Flv	Gld	Utr	N-H	Z-H	Zln	N-B	Lmb
T1SR	36	36	38	42	43	44	47	47	45	50	47	43
T2SR	43	42	43	48	49	49	53	53	50	54	53	48
T2S	49	48	51	53	55	54	58	56	54	58	56	54
T2R	64	63	62	65	66	63	68	68	63	73	67	64
T3SR	45	41	45	48	42	51	49	50	48	53	49	50
T3R	67	62	63	66	56	61	62	64	65	70	60	66
T3R2	68	63	64	67	57	62	62	65	65	70	60	67
T3	79	74	76	75	65	69	69	69	69	76	63	76
T4SR	43	41	45	48	45	50	49	53	50	54	51	49
T4R	65	63	64	65	62	62	63	68	63	73	63	65

Table 5.4
Means of standard errors over all months and provinces relative to the mean of the direct estimator's standard errors (%) for the state space models

	se(signal)	se(trend)	se(growth)
direct	100		
T1SR	57	41	6
T2SR	51	33	4
T2S	46	23	4
T2R	34	33	4
T3SR	53	35	9
T3R	36	35	9
T3R2	36	34	9
T3	28	26	9
T4SR	52	34	4
T4R	35	34	4

5.2 Results multilevel models

The ten models T1SR to T4R on pages 408-409 fitted as a state space model with the Kalman filter have also been fitted using the Bayesian multilevel approach using a Gibbs sampler. See Boonstra and van den Brakel (2016) for a detailed description of the fixed effect design matrices and random effect design and precision matrices corresponding to these models. The Bayesian approach accounts for uncertainty in the hyperparameters by considering their posterior distributions, implying that variance parameters do not actually become zero, as frequently happens for the ~~M~~ML estimates in the state space approach. For comparison purposes, however, effects absent from the state space model due to zero ML estimates have also been suppressed in the corresponding multilevel models. In addition to these ten models we consider one more model with extra terms including a dynamic RGB component as well as a white noise term.

Differences between state space and multilevel estimates based on the ten models considered can arise because of

- the different estimation methods, ML versus MCMC,

- the different modeling of survey errors. In the multilevel models the survey errors' covariance matrix is taken to be $\Sigma = \oplus_{i=1}^{m_A} \lambda_i \Phi_i$ with Φ_i the covariance matrix of estimated design variances for the initial estimates for province i, and λ_i scaling factors, one for each province. In the state space models the survey errors are allowed to depend on more parameters though eventually an AR(1) model is used to approximate these dependencies,
- the slightly different parameterizations of the trend components. For the trend in model T3, for example, the province of Groningen is singled out by the state space model used, because no local level component is added for that province.

The estimates and, to a lesser extent, the standard errors based on the multilevel models are quite similar to the results obtained with the state space models. We show this only for the smoothed signals of model T2R in Figure 5.4, as the qualitative differences between state space and multilevel results are quite consistent over all models. More comparisons for signals, trends and month-to-month developments for models T2R and T3R2 can be found in Boonstra and van den Brakel (2016).

The small differences between the state space and multilevel signal estimates are due to slightly more flexible trends in the estimated multilevel models. Larger differences can be seen in the standard errors of the signal: the multilevel models yield almost always larger standard errors for provinces with high unemployment levels (Flevoland and Zuid-Holland in the figure), whereas for provinces with smaller unemployment levels (e.g., Zeeland) the differences are somewhat less pronounced.

The larger flexibility of the multilevel model trends is most likely due to the relatively large uncertainty about the variance parameters for the trend, which is accounted for in the Bayesian multilevel approach but ignored in the ML approach for the state space models. The posterior distributions for the trend variance parameters are also somewhat right-skewed. The posterior means for the standard deviations are always larger than the ML estimates for the corresponding hyperparameters of the state space models (compare Table 2 and Table 8 in Boonstra and van den Brakel (2016)). For the models with trend T2, i.e., with a fully parametrized covariance matrix over provinces, the multilevel models show positive correlations among the provinces, as do the state space ML estimates, but the latter are much more concentrated near 1, whereas the posterior means for correlations in the corresponding multilevel model T2SR are all between 0.45 and 0.8.

Table 5.5 contains values of the DIC model selection criterion (Spiegelhalter, Best, Carlin and van der Linde, 2002), the associated effective number of model parameters p_{eff}, and the posterior mean of the log-likelihood. The parsimonious model T3 is selected as the most favourable model by the DIC criterion. So in this case the DIC criterion selects the same model as the AIC and BIC criteria do for the state space models. An advantage of DIC is that it uses an effective number of model parameters depending on the size of random effects, instead of just the number of model parameters used in AIC/BIC. That said, the numbers p_{eff} are in line with the totals of the numbers of states and hyperparameters in Table 5.1 for the state space models.

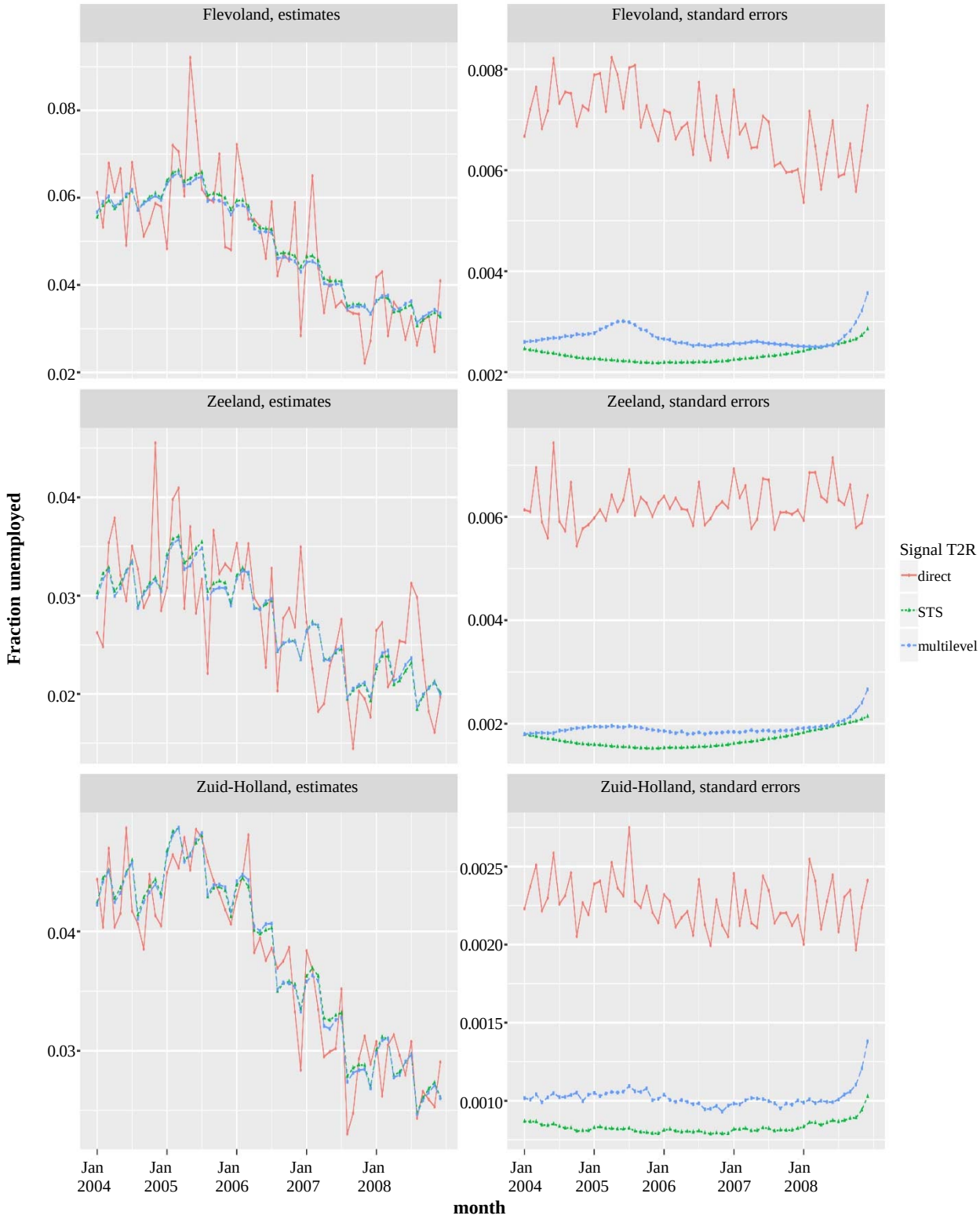


Figure 5.4 Comparison between smoothed signals (left) and their standard errors (right) obtained using state space (STS) model T2R and the corresponding multilevel model.

Table 5.5
DIC, effective number of model parameters and posterior mean of log likelihood

	DIC	p_{eff}	mean llh
T1SR	-29,054	255	14,655
T2SR	-29,076	235	14,656
T2S	-29,129	196	14,662
T2R	-29,164	118	14,641
T3SR	-29,081	242	14,662
T3R	-29,174	126	14,650
T3R2	-29,217	94	14,655
T3	-29,230	82	14,656
T4SR	-29,084	228	14,656
T4R	-29,170	109	14,640

As was the case for the state space models, the parsimonious model T3 comes with larger average bias over time for the provinces Groningen and Flevoland, which have the highest rates of unemployment. Model T3R2 has much smaller average biases for Groningen and Flevoland and since its DIC value is not that much higher than for model T3, model T3R2 seems to be a good compromise between models T3 and T3R, being more parsimonious than T3R and respecting provincial differences better than model T3.

Table 5.6 contains the average standard errors for signal, trend and month-to-month differences in the trend, in comparison to the average for the direct estimates. The average is taken over all months and provinces. The results are again similar to the results obtained with the state space models, see Table 5.4, although especially the standard errors of month-to-month changes are larger under the multilevel models.

Table 5.6
Means of standard errors over all months and provinces relative to the mean of the direct estimator's standard errors (%) for the multilevel time series models

	se(signal)	se(trend)	se(growth)
direct	100		
T1SR	55	41	8
T2SR	52	37	6
T2S	49	33	7
T2R	39	38	6
T3SR	53	38	15
T3R	39	38	15
T3R2	39	38	15
T3	34	32	15
T4SR	51	36	6
T4R	37	36	6

Finally, a multilevel model based on model T3R2 but with additional random effects has been fitted to the data. This extended model includes a white noise term, the balanced dummy seasonal (equivalent to the trigonometric seasonal), and a dynamic RGB component. These components were seen to be absent or time independent in the state space approach due to zero ML hyperparameter estimates, and therefore were also

not included in the multilevel models considered so far. In addition, the extended multilevel model includes season by province random effects, as a compromise between fixed provincial seasonal effects and no such interaction effects at all. More details and figures comparing the estimation results from this extended model to those from multilevel models T3R2 and T3SR can be found in Boonstra and van den Brakel (2016). It was found that most additional random effects were small so that the estimates based on the extended model are quite close to the estimates based on model T3R2, and the estimated standard errors are only slightly larger than those for model T3R2. A DIC value of -29,260 was found, well below the DIC value for model T3R2. This improvement in DIC was seen to be almost entirely due to the dynamic RGB component. Apparently, modeling the RGB as time-dependent results in a better fit. This seems to be in line with the temporal variations in differences between first wave and follow-up wave survey regression estimates, visible from Figure 3 in Boonstra and van den Brakel (2016).

6 Discussion

A time series small area estimation model has been applied to a large amount of survey data, comprising 6 years of Dutch LFS data, to estimate monthly unemployment fractions for 12 provinces over this period. Two different estimation approaches for structural time series models (STM) are applied and compared. The first one is a state space approach using a Kalman filter, where the unknown hyperparameters are replaced by their ML estimates. The second one is a Bayesian multilevel time series approach, using a Gibbs sampler.

The time series models that do not account for cross-sectional correlations and borrow strength over time only, already show a major reduction of the standard errors compared to the direct estimates. A further small decrease of the standard errors is obtained by borrowing strength over space through cross-sectional correlations in the time series models. Another great advantage of the time series model approach concerns the estimation of change. Under the multilevel model estimates of change and their standard errors can be easily computed, especially when the model fit is in the form of an MCMC simulation. Under the state space approach, estimates of change follow directly from the Kalman filter recursion by keeping the required state variables from the past in the state vector. The desired estimate for change, including its standard error, follows from the contrast of the specific state variables. Month-to-month and year-to-year change of monthly data are very stable and precise, which is a consequence of the strong positive correlation between level estimates. However, the stability of the estimates of change strongly depends on the choice of the trend model. Local level models result in more volatile trend estimates and thus also more volatile estimates of change and naturally have a higher standard error compared to smooth trend models.

In this paper different trend models are considered that model correlation between domains with the purpose to borrow strength over time and space. The most complex approach is to specify a full covariance matrix for the disturbance terms of the trend component. One way to construct parsimonious models is to take advantage of cointegration. In the case of strong correlation between domains the covariance matrix

will be of reduced rank, which means that the trends of the m_A domains are driven by less than m_A common trends. In this application two common trends are sufficient to model the dynamics of the twelve provinces, resulting in a strong reduction of the number of hyperparameters required to model the cross-sectional correlations between the domains. In order to further reduce the number of state and hyperparameters, alternative trend models are considered that implicitly account for cross-sectional correlations. Under this approach all domains share an overall trend. Each domain has a domain-specific trend to account for the deviation from the overall trend. This can be seen as a simplified form of a common trend model. In this application the alternative trend model results in comparable estimates for the trends and standard errors. So this approach might be a practical attractive alternative for common trend models. For example if the number of domains is large or the number of common factors is larger, then the proposed trend models are less complex compared to general common trend models. More research into the statistical properties of these alternative trend models is necessary for better understanding the implied covariance structures.

Several differences between the time series multilevel models fitted in an hierarchical Bayesian framework and state space models fitted with the Kalman filter with a frequentist approach can be observed. Within the multilevel Bayesian framework different STMs are compared using DIC as a formal model selection criterion. Since the state space models are fitted in a frequentist framework, STMs are compared with AIC or BIC. An advantage of the DIC criterion used in the Bayesian multilevel approach is that it uses the effective number of degrees of freedom as a penalty for model complexity. This implies that the penalty for a random effect increases with the size of the variance components of this random factor and varies between zero if the variance component equals zero and the number of levels of this factor if the variance component tends to infinity. The penalty in AIC or BIC for a random component always equals one, regardless the size of its variance component and therefore does not account properly for model complexity. Note that for multilevel models fitted in a frequentist framework the so-called conditional AIC is proposed (Vaida and Blanchard, 2005) where the penalty for model complexity is also based on the effective degrees of freedom. In this case the penalty for a random effect increases as the size of its variance component increases in a similar way as with the DIC. For state space models fitted in a frequentist framework such model selection criteria seem less readily available.

A difference between the multilevel models and state space models is that under the former model components are more often found to be time varying while under the state space approach most components, with the exception of the trend, are estimated as time invariant. This is a result of the method of model fitting. Under the frequentist approach applied to the state space models, ML estimates for many hyperparameters are on the border of the parameter space, i.e., zero for variance components and one for correlations between slope disturbance terms. Under the hierarchical Bayesian approach the entire distribution of the (co)variance parameters is simulated resulting in mean values for these hyperparameters that are never exactly on the border of the parameter space, e.g., always positive in the case of variance components. A consequence of this feature is that the variances of the trend hyperparameters are higher and that the covariances between the trend disturbances are smaller than one under the hierarchical Bayesian

approach. Another remarkable observation is that the DIC prefers models with time varying RGB and time varying seasonal components as well as a white noise term for the population parameter. This results in this application in models with a higher degree of complexity under the hierarchical Bayesian multilevel models compared to the state space models fitted in a frequentist approach. Differences in estimates for the trend and the signals are, however, small.

An advantage of the hierarchical Bayesian approach is that the standard errors of the domain predictions account for the uncertainty about the hyperparameters. As a result the standard errors obtained under the hierarchical Bayesian approach of comparable models are slightly higher and less biased compared to the state space approach. For the state space approach several bootstrap methods are available to account for hyperparameter uncertainty (Pfeffermann and Tiller, 2005) but these methods significantly increase the computational cost.

From a computational point of view there are some differences between the methods too. The Kalman filter approach applied to state space models can be used online, producing new filtered estimates by updating previous predictions when data for a new month arrives and is from that point of view computationally very efficient. The numerical optimization procedure for ML estimation of the hyperparameters, on the other hand, can be cumbersome for large multivariate models if the number of hyperparameters is large. The Gibbs sampler multilevel approach used here produces estimates for the whole time series at once. It must be re-estimated completely when data for a new month arrives. However, due to the use of sparse matrices and redundant parameterization the multilevel approach is quite competitive computationally, see also Knorr-Held and Rue (2002). An advantage of the simultaneous multilevel estimation is that constraints over time can easily be imposed. For example, imposing sum-to-zero constraints over time allows to include local level provincial trends for all provinces in addition to a global smooth trend with no resulting identification issues.

In this application there is a preference for the time series multilevel models in the hierarchical Bayesian framework. One reason is the relatively simple way the DIC criterion can be computed, which better accounts for model complexity than AIC or BIC. Also, the Gibbs sampler under the Bayesian approach is better suited to fit complex multivariate STMs with large numbers of hyperparameters. In addition, the standard errors for the domain predictions obtained under the multilevel models account for the uncertainty about the hyperparameters, also in a straightforward way.

The time series estimates are quite smooth, and a more thorough model evaluation is necessary to find out whether that is appropriate or whether the time series model underfits the unemployment data or is open to improvement in other ways. There are many ways in which the time series SAE model may be extended to further improve the estimates and standard errors. For example, it may be an improvement to use a logarithmic link function in the model formulation as in You (2008). Effects would then be multiplicative instead of additive. Another possible improvement would come from a more extensive modeling of the sampling variances (You and Chapman, 2006; You, 2008; Gómez-Rubio, Best, Richardson, Li and Clarke, 2010). The models can also be improved by including additional auxiliary information at the province by

month level, for instance registered unemployment. In Datta et al. (1999) similar effects associated with unemployment insurance are modeled as varying over areas, although not over time.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. This work was funded by the European Union under grant no. 07131.2015.001-2015.257. We are grateful to the unknown reviewers and the Associate Editor for careful reading of a former draft of this manuscript and providing useful comments.

References

Andrieu, C., Poucet, R. and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B*, 72, 269-342.

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 349, 23-30.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 401, 28-36.

Bollineni-Balabay, O., van den Brakel, J., Palm, F. and Boonstra, H.J. (2016). Multilevel hierarchical bayesian vs. state-space approach in time series small area estimation; the dutch travel survey. Technical Report 2016-03, https://www.cbs.nl/en-gb/our-services/methods/papers, Statistics Netherlands.

Boonstra, H.J. (2016). *mcmcsae: MCMC Small Area Estimation*. R package version 0.7, available on request from the author.

Boonstra, H.J., and van den Brakel, J. (2016). Estimation of level and change for unemployment using multilevel and structural time-series models. Technical Report 2016-10, https://www.cbs.nl/en-gb/our-services/methods/papers, Statistics Netherlands.

Chan, J.C.C., and Jeliazkov, I. (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1, 101-120.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 448, 1074-1082.

Diallo, M.S. (2014). *Small Area Estimation Under Skew-Normal Nested Error Models*. PhD thesis, School of Mathematics and Statistics, Carleton University, Ottawa, Canada.

Doornik, J.A. (2009). *An Object-oriented Matrix Programming Language Ox 6*. Timberlake Consultants Press.

Durbin, J., and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 366, 269-277.

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.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 3, 515-533.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 4, 457-472.

Gelman, A., Van Dyk, D.A., Huang, Z. and Boscardin, W.J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17, 1, 95-122.

Geman, S., and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, 6, 721-741.

Gómez-Rubio, V., Best, N., Richardson, S., Li, G. and Clarke, P. (2010). Bayesian statistics for small area estimation. Technical Report http://www.bias-project.org.uk/research.htm.

Harvey, A.C. (2006). Seasonality and unobserved components models: an overview. Technical Report ISSN 1725-4825, Conference on seasonality, seasonal adjustment and their implications for shortterm analysis and forecasting, 10-12 May 2006, Luxembourg.

Kish, L. (1965). *Survey Sampling*. Wiley.

Knorr-Held, L., and Rue, H. (2002). On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29, 4, 597-614.

Koopman, S.J., Harvey, A.C., Doornik, J.A. and Shephard, A. (1999). *STAMP: Structural Time Series Analyser, Modeller and Predicter*. Timberlake Consultants, Press London.

Koopman, S.J., Shephard, A. and Doornik, J.A. (1999). Statistical algorithms for models in state-space form using ssfpack 2.2. *Econometrics Journal*, 2, 113-166.

Koopman, S.J., Shephard, A. and Doornik, J.A. (2008). *Ssfpack 3.0: Statistical algorithms for models in state-space form*. Timberlake Consultants, Press London.

Krieg, S., and van den Brakel, J.A. (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics and Data Analysis*, 56, 2918-2933.

McCausland, W.J., Miller, S. and Pelletier, D. (2011). Simulation smoothing for state-space models: A computational efficiency analysis. *Computational Statistics and Data Analysis*, 55, 199-212.

O'Malley, A.J., and Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103, 484, 1405-1418.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 163-175.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14534-eng.pdf.

Pfeffermann, D., and Tiller, R. (2005). Bootstrap approximation to prediction mse for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26, 893-916.

Pfeffermann, D., and Tiller, R. (2006). Small area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Piepho, H.-P., and Ogutu, J.O. (2014). Simple state-space models in a mixed model framework. *The American Statistician*, 61, 3, 224-232.

Proietti, T. (2000). Comparing seasonal components for structural time series models. *International Journal of Forecasting*, 16, 247-260.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Wiley-Interscience.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

Rue, H., and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC.

Ruiz-Cárdenas, R., Krainski, E.T. and Rue, H. (2012). Direct fitting of dynamic models using integrated nested laplace approximations - inla. *Computational Statistics and Data Analysis*, 56, 1808-1828.

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

Spiegelhalter, D.J., Best, N.G. Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64, 4, 583-639.

Vaida, F., and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, 95, 2, 351-370.

van den Brakel, J.A., and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, 35, 2, 177-190. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11040-eng.pdf.

van den Brakel, J.A., and Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41, 2, 267-296. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14231-eng.pdf.

van den Brakel, J.A., and Krieg, S. (2016). Small area estimation with state-space common factor models for rotating panels. *Journal of the Royal Statisitcal Society*, 179, 763-791.

Woodruff, R. (1966). Use of a regression technique to produce area breakdowns of the monthly national estimates of retail trade. *Journal of the American Statistical Association*, 61, 314, 496-504.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 1, 19-27. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10614-eng.pdf.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 1, 25-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6602-eng.pdf.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 427-450
Statistics Canada, Catalogue No. 12-001-X

Robust variance estimators for generalized regression estimators in cluster samples

Timothy L. Kennel and Richard Valliant[1]

Abstract

Standard linearization estimators of the variance of the general regression estimator are often too small, leading to confidence intervals that do not cover at the desired rate. Hat matrix adjustments can be used in two-stage sampling that help remedy this problem. We present theory for several new variance estimators and compare them to standard estimators in a series of simulations. The proposed estimators correct negative biases and improve confidence interval coverage rates in a variety of situations that mirror ones that are met in practice.

Key Words: Jackknife variance estimator; Hat matrix adjustment; Leverage adjustment; Superpopulation model; Two-stage sample; Sandwich variance estimator.

1 Introduction

Generalized regression (GREG) estimation is a common technique used to calibrate estimates, reduce sampling errors, and correct for nonsampling errors. Official surveys of households often use generalized regression to calibrate sample-based estimates to population controls, assure consistent estimates of demographic characteristics across surveys, and reduce nonresponse and undercoverage errors. GREG estimation is also frequently used because it draws strength from auxiliary data, resulting in smaller sampling errors than other design-based estimators.

Popular techniques used to estimate the sampling errors of calibrated estimators from complex samples either require extensive computational resources or tend to underestimate the true sampling errors, especially with small to moderate sample sizes. Two popular techniques used to estimate the sampling variance of GREG estimators are linearization and replication. Linearization estimators (Särndal, Swensson and Wretman, 1989) may not converge to the true sampling error fast enough to produce accurate results in small to moderate samples. Särndal, Swensson and Wretman (1992, page 176) remark that "For complex statistics such as an estimator of a population variance, covariance, or correlation coefficient, fairly large samples may be required before the bias is negligible." On the other hand, alternative replication techniques such as the jackknife and the bootstrap that generally produce larger variance estimates can be computationally demanding.

Leverage-adjusted sandwich estimators provide an alternative approach to estimating design-based sampling errors that also have model-based justifications. Royall and Cumberland (1978) applied this approach to develop estimators of the prediction variance of estimators of finite population totals. From a model-based framework, Long and Ervin (2000) and MacKinnon and White (1985) demonstrated how the sandwich estimator could be used for variance estimation for estimators of regression parameters even when

1. Timothy L. Kennel is an Assistant Division Chief for Statistical Methods in the Decennial Statistical Studies Division, U.S. Census Bureau. E-mail: timothy.l.kennel@census.gov; Richard Valliant is Research Professor Emeritus, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. E-mail: valliant@umich.edu.

the variance component of the working model was misspecified. Valliant (2002) took this approach to estimate the design-based variance of GREG estimators under one stage of sampling. This paper extends Valliant's work to clustered sample designs.

In Section 2, we introduce the GREG estimator and present several alternative variance estimators for it. All derivations are contained in the Appendix. In Section 3, we show how the new variance estimators perform in several simulations. In Section 4, we summarize our findings with a conclusion.

2 Theoretical results

Suppose that the population has $i = 1, 2, \ldots, M$ clusters. In cluster i there are N_i elements so that there are $N = \sum_{i=1}^{M} N_i$ elements in the population. The universe of clusters is denoted as U and the universe of elements in cluster i is U_i. An analysis variable y_{ik} is associated with element k in cluster i. The population total of y is $t_{Uy} = \sum_{i=1}^{M} \sum_{k=1}^{N_i} y_{ik}$. Each population element also has a p-vector of auxiliary variables, $\mathbf{x}_{ik}$, that can be used in estimation. A two-stage sample is selected without replacement at the first and second stages. The selection probability of cluster i is π_i, and $\pi_{k|i}$ is the conditional selection probability of element k in cluster i. The overall selection probability of element ik is $\pi_{ik} = \pi_i \pi_{k|i}$. Denote the set of sample clusters by s and the set of sample elements within cluster i by s_i. The number of sample clusters is m while the number of sample elements selected from sample cluster i is n_i. The total sample size of elements is $n = \sum_{i \in s} n_i$.

As a working model, suppose that $\mathbf{Y}_U$, the N-vector of analysis variables, follows the following linear model:

$$\begin{aligned} E_\xi(\mathbf{Y}_U) &= \mathbf{X}\boldsymbol{\beta} \\ \operatorname{cov}_\xi(\mathbf{Y}_U) &= \boldsymbol{\Psi} \end{aligned} \tag{2.1}$$

where the subscript ξ denotes expectation with respect to a model; $\mathbf{X} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_M^\top]^\top$ is the $N \times p$ matrix of auxiliaries with $\mathbf{X}_i$ being the $N_i \times p$ matrix of auxiliaries for the N_i elements in cluster i; and $\boldsymbol{\beta}$ is a parameter vector of length p. Elements within clusters are assumed to be correlated while elements in different clusters are independent under the model. Thus, the covariance matrix $\boldsymbol{\Psi}$ is an $N \times N$ block diagonal matrix with diagonal matrices $\boldsymbol{\Psi}_i = [\psi_{ik}]_{N_i \times N_i}$. A key feature of the variance estimators we propose is that the particular form of ψ_{ik} does not have to be known to construct variance estimators. The proposed variance estimators will be consistent regardless of the form of $\boldsymbol{\Psi}$.

Särndal et al. (1992, Chapter 8) discuss three different GREG estimators that can be used in clustered samples. These three estimators depend on the available data. We consider their case B which occurs when unit-level data are available for the complete sample and control totals are available for the population. In this case, the GREG estimator is

$$\begin{aligned} \hat{t}_y^{gr} &= \hat{t}_{y\pi} + \hat{\mathbf{B}}^\top \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi} \right) \\ &= \mathbf{g}^\top \boldsymbol{\Pi}^{-1} \mathbf{y}_s \end{aligned} \tag{2.2}$$

where $\mathbf{y}_s$ is the n-vector of y's for the sample elements, $\hat{t}_{y\pi}$ is the π-estimator of the total of the y's, $\mathbf{t}_{Ux}$ is the p-vector of population totals of the x's, $\hat{\mathbf{t}}_{x\pi}$ is the π-estimator of $\mathbf{t}_{Ux}$, and (if $\boldsymbol{\Psi}$ is known) $\hat{\mathbf{B}} = \mathbf{A}^{-1}\mathbf{X}_s^\top\boldsymbol{\Psi}_s^{-1}\boldsymbol{\Pi}^{-1}\mathbf{y}_s$ with $\mathbf{A} = \mathbf{X}_s^\top\boldsymbol{\Psi}_s^{-1}\boldsymbol{\Pi}^{-1}\mathbf{X}_s$, $\mathbf{X}_s$ the matrix of sample auxiliaries, and $\boldsymbol{\Pi} = \text{diag}[\pi_{ik}]$ $(i \in s, k \in s_i)$; $\boldsymbol{\Psi}_s$ is the part of $\boldsymbol{\Psi}$ associated with the sample elements; and $\mathbf{g}^\top = \mathbf{1}_n^\top + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1}\mathbf{X}_s^\top\boldsymbol{\Psi}_s^{-1}$ where $\mathbf{1}_n$ is a vector of n 1's.

The component of the g-weight for sample cluster i is $\mathbf{g}_i^\top = \mathbf{1}_{n_i}^\top + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1}\mathbf{X}_{si}^\top\boldsymbol{\Psi}_{si}^{-1}$ with $\mathbf{X}_{si}^\top = \left[\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\right]$ being the $p \times n_i$ matrix of auxiliaries for sample elements in sample cluster i, $\boldsymbol{\Psi}_{si}$ is the $n_i \times n_i$ part of $\boldsymbol{\Psi}_i$ for sample elements in sample cluster i, and $\mathbf{1}_{n_i}$ is a vector of n_i 1's. Since $\boldsymbol{\Psi}$ is generally unknown, a surrogate value $\mathbf{Q}$ may be used for $\boldsymbol{\Psi}_s^{-1}$; $\mathbf{Q} = \mathbf{I}$ is a common choice. Below, we assume that a general $\mathbf{Q}$ is used in the GREG rather than $\boldsymbol{\Psi}_s^{-1}$.

2.1 Current variance estimators

Särndal et al. (1992, Result 8.9.1) present an estimator of the design variance of $\hat{t}_y^{gr}$, which involves joint selection probabilities of clusters and elements within clusters. In the case of Poisson sampling at both stages, their estimator is

$$\upsilon_g = \sum_{i\in s}\frac{(1-\pi_i)}{\pi_i^2}\left(\hat{t}_{e,i}^{g}\right)^2 + \sum_{i\in s}\frac{1}{\pi_i}\sum_{k\in s_i}\frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2}g_{ik}^2e_{ik}^2 \tag{2.3}$$

where $\hat{t}_{e,i}^{g} = \sum_{s_i} g_{ik}e_{ik}\big/\pi_{k|i}$, g_{ik} is the k^{th} component of the $\mathbf{g}_i$ vector, and $e_{ik} = y_{ik} - \mathbf{x}_{ik}^\top\hat{\mathbf{B}}$. This estimator is computationally simpler than the general form that uses joint selection probabilities and may perform reasonably well for πps designs where the variance of estimators can be approximated by formulas that assume independence between selections.

An estimator that is appropriate if the first-stage sample is selected with replacement is

$$\upsilon_{wr} = \frac{m}{m-1}\sum_{i\in s}\left(e_{1i} - \overline{e}_1\right)^2 \tag{2.4}$$

with $e_{1i} = \sum_{k\in s_i} e_{ik}\big/\pi_{ik}$ and $\overline{e}_1 = m^{-1}\sum_{i\in s} e_{1i}$. The jackknife linearization estimator is (Yung and Rao, 1996)

$$\upsilon_{JL} = \frac{m-1}{m}\sum_{i\in s}\left(e_{2i} - \overline{e}_2\right)^2 \tag{2.5}$$

where $e_{2i} = \sum_{k\in s_i} g_{ik}e_{ik}\big/\pi_{ik}$ and $\overline{e}_2 = m^{-1}\sum_{i\in s} e_{2i}$ with g_{ik} being the k^{th} component of the $\mathbf{g}_i$ vector.

The jackknife is another popular variance estimation technique. Krewski and Rao (1981) present several asymptotically equivalent ways of writing the jackknife. The following form of the jackknife estimator is a convenient starting point for the calculations that follow:

$$\upsilon_{\text{Jack}} = \frac{m-1}{m}\sum_{i\in s}\left(\hat{t}_{y(i)}^{gr} - \hat{t}_{y(\cdot)}^{gr}\right)^2 \tag{2.6}$$

where $\hat{t}_{y(i)}^{gr}$ is the value of the GREG estimator after removing cluster i and $\hat{t}_{y(\cdot)}^{gr}$ is the average of all $\hat{t}_{y(i)}^{gr}$ estimates. Using (2.6) can be computationally demanding because m different estimates of $\hat{t}_{y(i)}^{gr}$ must be

computed. The estimators, v_{Jack}, v_{wr}, and v_{JL} are all design-consistent under the conditions in Krewski and Rao (1981) and Yung and Rao (1996). One of their key conditions is that clusters be selected with replacement. This assumption simplifies theoretical calculations but is only a convenience since the theoretical results have been shown in many empirical studies to be good predictors of estimator performance in without-replacement designs as long as the first-stage sampling fraction is small.

2.2 New variance estimators

We use the model-based framework to construct new variance estimators. First, we derive the model-based variance of $\hat{t}_y^{gr}$. Assume that model (2.1) holds and that sampling is ignorable in the sense that the probability of a unit's being in the sample given $\mathbf{Y}_U$ and $\mathbf{X}$ depends only on $\mathbf{X}$ (e.g., see discussion in Valliant, Dorfman and Royall, 2000, Section 2.6.2 and the additional references therein). Then, we construct estimators of the model variance, using hat-matrix adjustments to account for heterogeneity in the data. We evaluate the design-based properties of the new variance estimators in a simulation.

To calculate the model variance of $\hat{t}_y^{gr}$, define $\mathbf{y}_i$ as the population vector of analysis variables for cluster i, and $\mathbf{y}_{si}$ as the vector for sample elements. As shown in Appendix A.2, under model (2.1) the model-based variance of $\hat{t}_y^{gr}$ is

$$\begin{aligned}\text{var}_\xi\left(\hat{t}_y^{gr} - t_{Uy}\right) &= \sum_{i\in s}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\boldsymbol{\Psi}_{si}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i - 2\sum_{i\in s}\left[\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\text{cov}_\xi\left(\mathbf{y}_{si},\,\mathbf{y}_i\right)\mathbf{1}_{N_i}\right] + \mathbf{1}_N^\top\boldsymbol{\Psi}\mathbf{1}_N \\ &= L_1 - 2L_2 + L_3\end{aligned}$$

where $\text{var}_\xi\left(\mathbf{y}_{si}\right) = \boldsymbol{\Psi}_{si}$, the part of $\boldsymbol{\Psi}$ associated with elements in s_i, and $\mathbf{1}_{N_i}$ and $\mathbf{1}_N$ are vectors of N_i and N 1's.

The model-based error variance of $\hat{t}_y^{gr}$ requires knowledge of $\boldsymbol{\Psi}$ for the full population. Without some strong assumptions that link the sample and nonsample covariance structures, components of $\boldsymbol{\Psi}$ associated with the nonsample cannot be estimated from the sample. However, as shown in Appendix A.2, under some reasonable conditions the orders of the terms are $L_1 = O\left(M^2/m\right)$ and $L_2 = L_3 = O\left(M\right)$ so that L_1 dominates the variance as the number of sample and population clusters increase. Thus,

$$\text{av}_\xi\left(\hat{t}_y^{gr} - t_{Uy}\right) = \sum_{i\in s}\mathbf{g}_i^\top\boldsymbol{\Pi}_i^{-1}\boldsymbol{\Psi}_{si}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i \tag{2.7}$$

where av_ξ denotes asymptotic model variance under the assumptions in Appendix A.1. A robust estimator of the right-hand side of (2.7) can be formed even when $\boldsymbol{\Psi}_{si}$ is unknown. On the other hand, if the number of population clusters increases at the same rate as sample clusters, (i.e., $f = m/M$ converges to a non-zero constant), then L_1, L_2, and L_3 may all contribute importantly to the asymptotic variance. In this paper, we will only consider estimation of L_1.

Unless the true variance matrix of $\mathbf{y}_s$ is known, $\boldsymbol{\Psi}_i$ must be estimated. In Appendix A.3 we show that in large samples $\text{var}_\xi\left(\mathbf{e}_i\right) \approx \boldsymbol{\Psi}_i$ where $\mathbf{e}_i = \mathbf{y}_{si} - \hat{\mathbf{y}}_{si}$ with $\hat{\mathbf{y}}_{si} = \mathbf{X}_{si}\hat{\mathbf{B}}$ and $\mathbf{X}_{si}$ being the $n_i \times p$ matrix of auxiliaries for sample elements in sample cluster i. Substituting $\mathbf{e}_i\mathbf{e}_i^\top$ for $\boldsymbol{\Psi}_{si}$ in (2.7) yields the sandwich estimator

$$\upsilon_R = \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i . \tag{2.8}$$

Based on results in Appendix A.3, υ_R is approximately unbiased for $\text{av}_\xi \left(\hat{t}_y^{gr} - t_{Uy} \right)$ in large samples. This sandwich estimator is also closely related to the design-based, ultimate cluster estimator for a sample design in which clusters are selected with replacement, which is, in turn, similar to both υ_g and υ_{JL} in with replacement sampling. Consequently, υ_R has both desirable design-based and model-based properties.

In small to moderate-sized samples, υ_R will be model-biased and will often underestimate the true variance. A hat-matrix adjustment can be made as a correction. As shown in Appendix A.3,

$$\text{E}_\xi \left(\mathbf{e}_i \mathbf{e}_i^\top \right) = \text{var}_\xi \left(\mathbf{e}_i \right) = \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right) \mathbf{\Psi}_{si} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^\top + \sum_{j \neq i;\, i,\, j \in s} \mathbf{H}_{ij} \mathbf{\Psi}_{sj} \mathbf{H}_{ij}^\top \tag{2.9}$$

where $\mathbf{H}_{ij} = \mathbf{X}_{si}^\top \mathbf{A}^{-1} \mathbf{X}_{sj} \mathbf{Q}_j \mathbf{\Pi}_j^{-1}$ $(i,\ j = 1, \ldots, m)$ with $\mathbf{Q}_j$ and $\mathbf{\Pi}_j$ being the $n_j \times n_j$ parts of $\mathbf{Q}$ and $\mathbf{\Pi}$ associated with sample cluster j. As in (Li and Valliant, 2009; Valliant, 2002), the $\mathbf{H}_{ij}$ can be collected into a survey weighted hat matrix:

$$\begin{aligned} \mathbf{H} &= \mathbf{X}_s \mathbf{A}^{-1} \mathbf{X}_s^\top \mathbf{Q} \mathbf{\Pi}^{-1} \\ &= \begin{bmatrix} \mathbf{X}_{s1} \mathbf{A}^{-1} \mathbf{X}_{s1}^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \ldots & \mathbf{X}_{s1} \mathbf{A}^{-1} \mathbf{X}_{sm}^\top \mathbf{Q}_m \mathbf{\Pi}_m^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_{sm} \mathbf{A}^{-1} \mathbf{X}_{s1}^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \ldots & \mathbf{X}_{sm} \mathbf{A}^{-1} \mathbf{X}_{sm}^\top \mathbf{Q}_m \mathbf{\Pi}_m^{-1} \end{bmatrix}. \end{aligned} \tag{2.10}$$

Based on the assumptions in Appendix A.1, $\mathbf{H} = O\left(m^{-1} \right)$, from which we conclude that $\text{var}_\xi \left(\mathbf{e}_i \right) \approx \mathbf{\Psi}_{si}$. The diagonal submatrices $\mathbf{H}_{ii}$ are matrix analogs to leverages in single-stage sampling. In ordinary least squares regression, the vector of predicted values can be written as $\hat{\mathbf{y}} = \mathbf{H}_{\text{OLS}} \mathbf{y}$ with $\mathbf{H}_{\text{OLS}} = \mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T$. Leverages are diagonals of the hat matrix, $\mathbf{H}_{\text{OLS}}$, and can be used to correct for a small sample bias in $e_i^2 = \left(y_i - \hat{y}_i \right)^2$ as an estimator of $\text{var}_\xi \left(y_i \right)$. We use the $\mathbf{H}_{ii}$ in an analogous way below.

To adjust for the fact that $\mathbf{e}_i \mathbf{e}_i^\top$ is model-biased for small to moderate samples, we make leverage-like adjustments to $\mathbf{e}_i \mathbf{e}_i^\top$. If $\mathbf{Q} = \mathbf{I}$ and the sample is self-weighting (i.e., $\mathbf{\Pi} = c\mathbf{I}$ for some $0 < c < 1$), then $\text{var}_\xi \left(\mathbf{e}_i \right) = \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right) \mathbf{\Psi}_{si}$ (see Appendix A.3). Solving for $\mathbf{\Psi}_{si}$ and substituting into (2.8) gives the variance estimator:

$$\upsilon_D = \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i \tag{2.11}$$

which, in this special case, is also approximately unbiased since $\mathbf{H}_{ii} = O\left(m^{-1} \right)$. One undesirable feature of υ_D is that it can be negative or can have negative contributions from some clusters if $\upsilon_{Di} = \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i < 0$. For such clusters, replacing υ_{Di} with $\upsilon_{Ri} = \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i$ will assure a positive variance estimator. This adjustment is used in the simulation in Section 3.

In Appendices A.4 and A.5, we show that the jackknife variance estimator can be written exactly as

$$\upsilon_{\text{Jack}} = \frac{m-1}{m}\left[\sum_{i\in s}\left(D_i - \bar{D}\right)^2 - 2\sum_{i\in s}\left(D_i - \bar{D}\right)F_i + \sum_{i\in s}F_i^2\right] \tag{2.12}$$

where

$$\begin{aligned}
F_i &= \left(G_i - \bar{G}\right) - \frac{1}{n}\left(K_i - \bar{K}\right) \\
D_i &= \mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i \\
K_i &= \left(\mathbf{1}_N^\top \mathbf{X}_U - m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_{si}\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right);\ \bar{K} = m^{-1}\sum_{i\in s}K_i \\
G_i &= \mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\left[\mathbf{H}_{ii}\mathbf{y}_{si} - \hat{\mathbf{y}}_{si}\right];\ \bar{G} = m^{-1}\sum_{i\in s}G_i \\
\mathbf{R}_i &= \mathbf{A}^{-1}\mathbf{X}_{si}^\top \mathbf{Q}_i \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i .
\end{aligned}$$

This form of υ_{Jack} results in a significant reduction in computations since only one GREG estimate is needed, rather than m estimates. (Of course, recomputing the GREG for every jackknife replicate may still be advantageous if an elaborate nonresponse adjustment affects the size of the true variance.)

In large samples υ_{Jack} can be approximated by

$$\upsilon_{J1} = \frac{m-1}{m}\sum_{i\in s}\left(D_i - \bar{D}\right)^2 \tag{2.13}$$

or by

$$\begin{aligned}
\upsilon_{J2} &= \frac{m-1}{m}\sum_{i\in s}D_i^2 \\
&= \frac{m-1}{m}\sum_{i\in s}\mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^\top\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\boldsymbol{\Pi}_i^{-1}\mathbf{g}_i .
\end{aligned} \tag{2.14}$$

The estimators, υ_{J1} and υ_{J2} are clustered versions of the single-stage approximations to the jackknife in Valliant (2002, equations (3.5), (3.6)).

As sketched in Appendix A.6, $\upsilon_{\text{Jack}}, \upsilon_{JL}, \upsilon_{J1}, \upsilon_{J2}, \upsilon_D,$ and υ_R are all asymptotically equivalent as $m \to \infty$. Since υ_{Jack} and υ_{JL} are design-consistent, the alternative estimators above can be expected to perform well over repeated samples when the size of the first-stage sample is large, and when model (2.1) is approximately correct. One caveat is that the sampling fraction of clusters must be small so that estimators made from a without-replacement, first-stage sample will perform as if the sample had been selected with-replacement.

None of these sandwich-like estimators includes finite population correction factors. Thus, they may tend to overestimate the sampling variance when a large proportion of the sample clusters is selected. To account for this, we can further adjust all of the variance estimators in an ad hoc fashion by multiplying the variance estimators by a finite population correction factor, denoted f_{pc}, as developed by Kott (1988). This results in the following adjusted estimators:

$$
\begin{aligned}
\upsilon_R^* &= f_{pc} \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i \\
\upsilon_D^* &= f_{pc} \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i \\
\upsilon_{\text{Jack}}^* &= f_{pc} \frac{m}{m-1} \left[\sum_{i \in s} \left(D_i - \bar{D} \right)^2 - 2 \sum_{i \in s} \left(D_i - \bar{D} \right) F_i + \sum_{i \in s} F_i^2 \right] \\
\upsilon_{J1}^* &= f_{pc} \frac{m}{m-1} \sum_{i \in s} \left(D_i - \bar{D} \right)^2 \\
\upsilon_{J2}^* &= f_{pc} \sum_{i \in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii} \right)^{-1} \mathbf{\Pi}_i^{-1} \mathbf{g}_i .
\end{aligned}
$$

When a simple random sample is selected in the first stage, $f_{pc} = 1 - m/M$. According to Kott (1988), an appropriate correction when the first stage is selected with varying probabilities is $f_{pc} = 1 - m\sum_{i=1}^{M} p_i^2$ where p_i is the single draw probability for cluster i, i.e., the probability that cluster i would be selected in a sample of size 1.

3 Simulation

We performed a series of simulation studies to test the performance of the new variance estimators in different populations. In each simulated sample, we computed the quantities listed in Table 3.1. To evaluate the variance estimators, we calculated the average of the variance estimates, compared those averages to the empirical mean square error, and computed coverage probabilities of confidence intervals based on the different variance estimates. Table 3.2 summarizes the sample designs for the 18 simulation studies. The column called Label gives the headings used in later tables. The sample designs are used in three populations described below.

Table 3.1
Statistics of interest for clustered GREG variance simulation

Statistic	Description
$\hat{t}_y^{\pi}$	Estimated total from the Horvitz-Thompson Estimator
$\hat{t}_y^{gr}$	Estimated total from the GREG
υ_E	Empirical variance
υ_g	Design-based variance estimator that assumes Poisson sampling at both stages from Särndal et al. (1992) in (2.3)
υ_{wr}	With-replacement variance estimator in (2.4)
υ_{JL}	Jackknife linearization variance estimator from Yung and Rao (1996) in (2.5)
υ_R	Sandwich estimator in (2.8)
υ_D	First hat-matrix adjusted sandwich estimator in (2.11)
υ_{Jack}	Jackknife variance estimator in (2.6)
υ_{J1}	First approximation to the jackknife variance estimator in (2.13)
υ_{J2}	Second approximation to the jackknife variance estimator in (2.14)
υ_R^*	Sandwich estimator with a finite population adjustment
υ_D^*	First hat-matrix adjusted sandwich estimator with a finite population correction
υ_{Jack}^*	Jackknife variance estimator with a finite population correction
υ_{J1}^*	First approximation to jackknife with a finite population correction
υ_{J2}^*	Second approximation to jackknife with a finite population adjustment

3.1 Data

We conducted simulations on three populations to assess the design-based performance of the variance estimators under a variety of situations. In the first population, we investigated the performance of the variance estimators when the first-stage sampling fraction was large and the sample size was moderate. The focus of the second simulation study was on the performance of the variance estimators under a relatively messy dataset and a small first-stage sample size. The final simulation study shows the performance of the variance estimators in large samples.

Table 3.2
Simulation designs for three populations

	Label	Population	First stage sample	*m*	Second stage sample	No. of samples
1	srs fixed	Third Grade	srswor	25	$n_i = 5$	1,000
2	srs fixed	Third Grade	srswor	50	$n_i = 5$	1,000
3	srs epsem	Third Grade	srswor	25	$f_i = \frac{675}{2,427}$	1,000
4	srs epsem	Third Grade	srswor	50	$f_i = \frac{675}{2,427}$	1,000
5	pps epsem	Third Grade	ppswor	25	$n_i = 5$	1,000
6	pps epsem	Third Grade	ppswor	50	$n_i = 5$	1,000
7	srs fixed	ACS	srswor	3	$n_i = 9$	5,000
8	srs fixed	ACS	srswor	15	$n_i = 9$	5,000
9	srs epsem	ACS	srswor	3	$f_i = \frac{30,430}{194,329}$	5,000
10	srs epsem	ACS	srswor	15	$f_i = \frac{30,430}{194,329}$	5,000
11	pps epsem	ACS	ppswor	3	$n_i = 9$	5,000
12	pps epsem	ACS	ppswor	15	$n_i = 9$	5,000
13	srs fixed	Simulated	srswor	300	$n_i = 2$	1,000
14	srs fixed	Simulated	srswor	1,500	$n_i = 2$	100
15	srs epsem	Simulated	srswor	300	$f_i = \frac{60,000}{195,164}$	1,000
16	srs epsem	Simulated	srswor	1,500	$f_i = \frac{60,000}{195,164}$	100
17	pps epsem	Simulated	ppswor	300	$n_i = 3$	1,000
18	pps epsem	Simulated	ppswor	1,500	$n_i = 3$	100

3.1.1 Third grade population

The first simulation study used the Third Grade population from Appendix B.6 of Valliant et al. (2000). This dataset contained the mathematics achievement scores for 2,427 third graders in 135 schools. The relatively small number of schools in this population and the fairly constant number of students in each school made it ideal for studying samples with large sampling fractions.

We used GREG to estimate the average mathematics achievement score for third graders. Altogether, we selected 1,000 samples in each of six sample designs listed in Table 3.2. In the first sample design, we selected 1,000 simple random samples without replacement (srswor) of 25 schools. Within each sampled school, we selected exactly five students via srswor. Because the number of students in each school varied

from school to school, this sample design resulted in different unconditional probabilities of selection, but a fixed sample size of 125 students. The second sample design was similar to the first, except we selected 50 schools. Selecting 50 of the 135 schools resulted in a large first-stage sampling fraction of 0.37, necessitating a finite population correction factor. Both the samples of $m = 25$ and 50 might be considered to be of "moderate" size.

In the third sample design, we selected 1,000 simple random samples of 25 schools without replacement. Within each sampled school, we selected students at a constant rate of $\frac{675}{2,427}$, yielding 1,000 samples with random sizes centered around 125 students. The result of this design was that each student had the same unconditional probability of selection. The fourth sample design was similar to the third, except we selected 50 schools. The sample sizes were also random under this design, with an average of 250 students. Since the third and fourth sample designs resulted in every unit getting the same chance of selection, these sample designs are labeled srs epsem (equal probability selection mechanism) in subsequent tables.

In the fifth design, we selected 1,000 samples of 25 schools with probabilities proportional to the number of students in each school. Within each sampled school, we selected exactly five students, yielding 1,000 samples with exactly 125 students each. The sixth sample design was similar to the fifth, except we selected 50 schools. We selected 1,000 samples of size 250 students using this design. The fifth and sixth designs are epsem. Like the second and fourth sample designs, this sample design also had a large sampling fraction and warranted the need for a finite population correction factor to adjust the variance estimators.

From each sample, we estimated the average achievement scores for the finite population using a GREG estimator and assuming that the number of students in the population was known. The assisting model was meant to replicate the clustered linear regression model in Section 9.6 of Valliant et al. (2000). The eleven explanatory variables used to model each student's math achievement score were: an intercept, sex (male or female), ethnicity (White/Asian, Black, Native American/Other, or Hispanic), language spoken at home is the same as the test (Always, Sometimes/Never), type of community (Outskirts of a town or city, Village/City), and school enrollment. The total mathematics achievement estimated with the GREG estimator was divided by the number of students in the population, 2,427, to get the average achievement score. The average achievement score for the population was 477.7. For the full population, the R-squared for the student-level linear model was 0.9735, indicating a very strong linear relationship.

3.1.2 American Community Survey population

The second simulation study used Census 2000 Summary File 3 data and American Community Survey (ACS) 2005 - 2009 Summary File data. The goal was to estimate the total number of housing units in the U.S. state of Alabama as reported in the ACS Summary File. Block group counts from Census 2000 were used as covariates in the assisting model.

To create the population, first all block group data were extracted from the ACS Summary File and the Census 2000 Summary File 3. Then, the two files were merged at the block group level. Block groups with 1,000 or more housing units in Census 2000 were removed because such large block groups had different characteristics than the majority of blocks. In many sampling designs such large units would be placed in a separate, certainty stratum and not contribute to the variance of estimates. Also, block groups with extreme

growth in the total number of housing units were also removed. Specifically block groups that had gained more than 10 units over twice the 2000 census count were removed.

Clusters were defined as counties and block groups were treated as units. Treating block group as a unit is motivated by the common task of selecting a sample of blocks, listing them, and then using the listings to estimate the total number of housing units in the finite population.

Clusters with fewer than 10 block groups or more than 120 block groups in them were removed from the frame of clusters. Overall, there were 61 clusters (counties) containing a total of 2,051 block groups and 1,109,499 housing units in the edited dataset. Altogether, six counties and 1,278 block groups containing 1,030,471 housing units were removed from the Alabama file.

Figure 3.1 shows two scatterplots. The first plot shows the total number of housing units in the block group as reported on the ACS summary file as a function of the 2000 Census housing unit count. Each point represents one of the 2,051 block groups in the finite population. The diagonal line is a nonparametric smoother, indicating a strong relationship between the two variables. The plot also shows some evidence of heteroscedasticity because the points appear to fan out as the 2000 census count increases. The second plot shows the residuals obtained by regressing the 2000 census housing unit count on the ACS housing unit count using ordinary least squares (OLS) plotted versus the ACS housing unit count. As the number of housing units reported on the ACS file increases, the model predictions appear to seriously underestimate the true number of housing units. This suggests some degree of nonlinearity in the mean function. In addition, there is noticeable heteroscedasticity in variance.

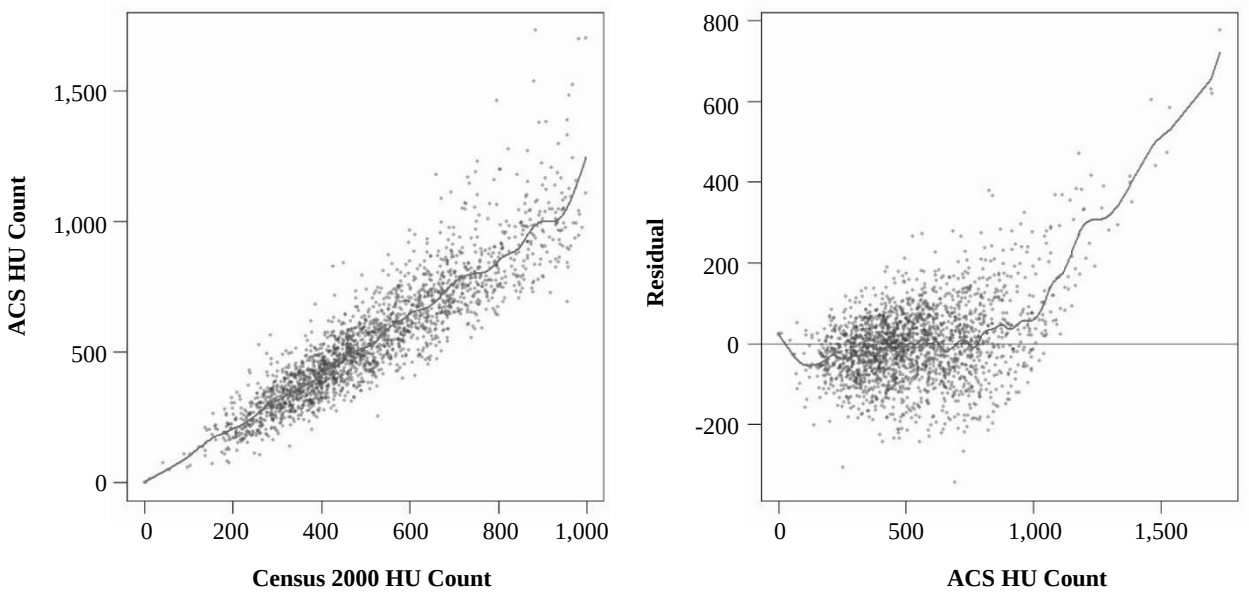


Figure 3.1 Scatter plot and residual plot for ACS population. Gray lines are nonparametric smoothers.

As in the first simulation study, we tested six different sample designs. We selected 5,000 samples in each of six different selection mechanisms listed in Table 3.2. In the first sample design, we selected 5,000 simple random samples of 3 clusters without replacement. In large national surveys, it is not uncommon to

select a small number of primary sampling units in each stratum. In this case, we treat Alabama as if it were a single design stratum and its 61 counties as clusters. Three counties within that stratum were sampled. Within each cluster, we selected nine block groups using srswor. The second design was similar with 15 clusters and 9 block groups per cluster. The first two sample designs resulted in highly variable weights. The other designs (rows 9-12) were parallel to those in rows 3-6 for the Third Grade population. The sample sizes of $m = 3$ and 15 are small so that theoretical, large sample properties are less likely to hold.

From each sample, we estimated the total number of housing units in the finite population using a GREG estimator. The assisting model included an intercept and the Census 2000 count of housing units; the heteroscedasticity noted above was not accounted for in the GREG. For the full population, the R-squared was 0.819, again indicating a strong linear relationship.

3.1.3 Simulated population

A population was created with a large number of clusters to assess the asymptotic characteristics of the variance estimators. Generated using a classic linear model, a total of 30,000 clusters were created, each with a random number of units. The number of units in each cluster was determined by adding three to a uniform random integer between 0 and 7. This created clusters ranging in size from 3 to 10 units. Altogether, the population contained 195,164 units within 30,000 clusters. For each unit, a positive covariate was created as $x_k \sim 1{,}000 \exp N(0, 1)$ where $N(0, 1)$ is a normal random variate with mean of 0 and standard deviation of 1. A random response was created such that $y_k \sim N\left(1{,}000 + 2x_k, \frac{x_k}{2}\right)$. Figure 3.2 shows scatter plots of the relationship between x_k and y_k for the finite population.

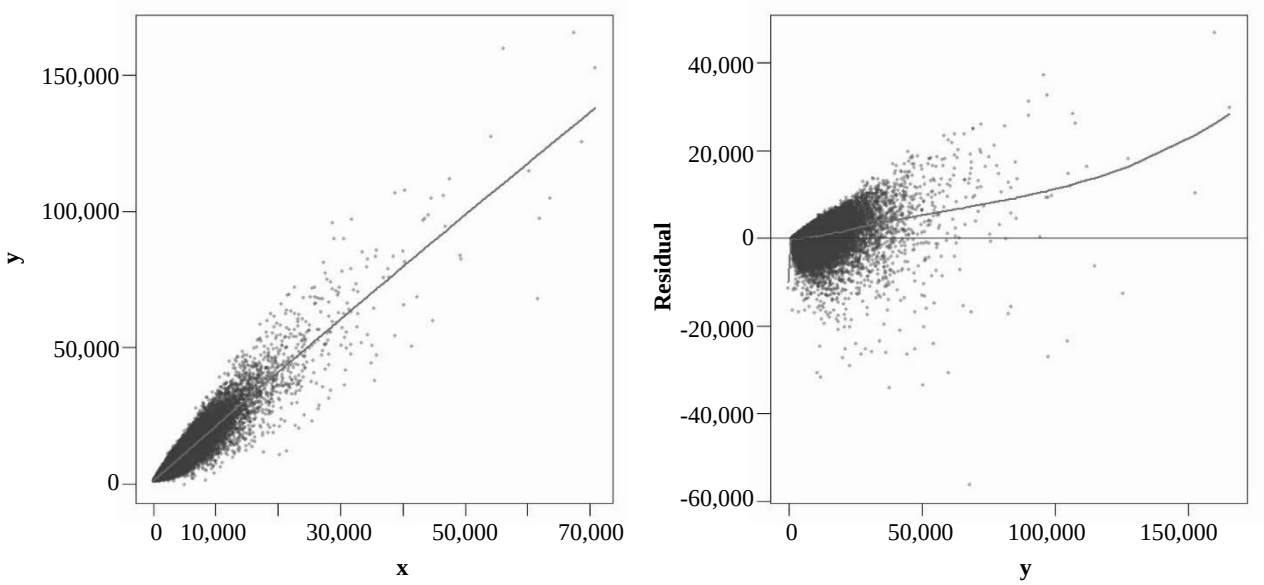


Figure 3.2 Scatter plot and residual for simulated population. Gray lines are nonparametric smoothers.

We selected samples using the six different probability selection mechanisms listed in rows 13-18 of Table 3.2. The types of sample designs are parallel to those used for the Third Grade and ACS populations.

In designs 14, 16, and 18, we selected 100 simple random samples of 1,500 clusters without replacement. We only selected 100 samples due to the excessive amount of computer time it took to select and process each sample. The sample sizes of $m = 300$ and 1,500 are large so that theoretical, large sample properties should hold.

From each sample, we estimated the total of the response using a GREG estimator. The true finite population total was 839,149,969. The assisting model included an intercept and x with $\mathbf{Q} = \mathbf{I}$. For the full population, the R-squared was 0.953, indicating a very strong linear relationship. Figure 3.2 shows a scatter plot of the population as well as a residual plot based on an OLS regression of x_k on y_k for the full population. There is clear evidence of heteroscedasticity of errors.

3.2 Results

We explored the bias, variability, and confidence interval coverage of the new and existing variance estimators. We only show tables for some of the simulations to conserve space. Table 3.3 shows the means of the π-estimator and the GREG estimator as well as the ratios of the average values of the variance estimators to the empirical mse's for all populations and sample size combinations across all simulations. Both the π-estimator and the GREG estimator are approximately unbiased; however, the GREG estimator is much more efficient.

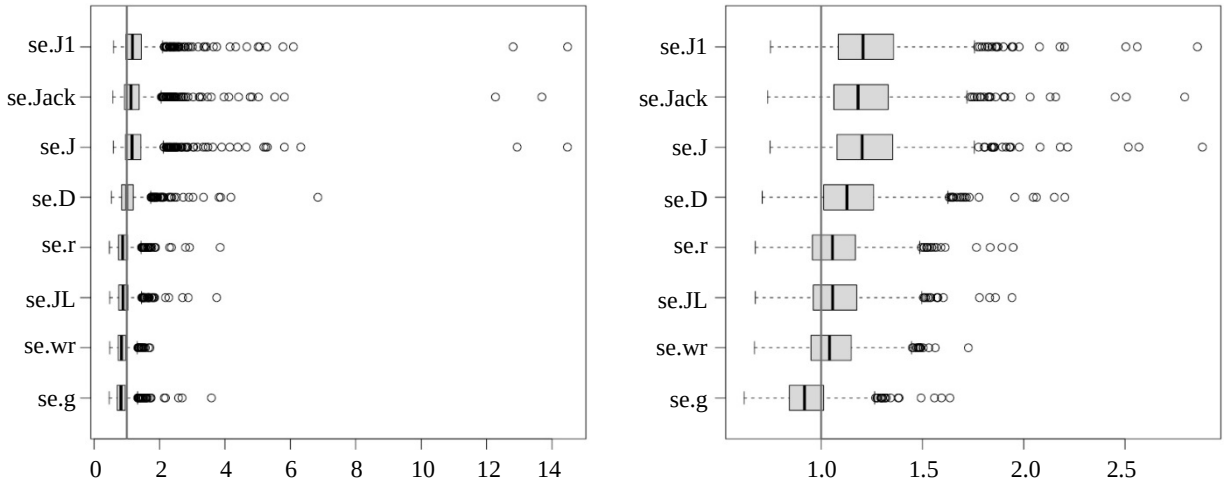
Table 3.3
Simulation Results for estimates for means and variance estimators for three populations and six sample designs in each population. Values in rows for variance estimators are ratios of mean estimated variance to empirical mse of the GREG. See Table 3.1 for descriptions of the variance estimators

Estimator	srs fixed		srs epsem		pps epsem	
			Third Grade Population			
	$m = 25$	$m = 50$	$m = 25$	$m = 50$	$m = 25$	$m = 50$
Average $\hat{t}_y^{\pi} / N$	477.23	477.11	476.29	476.85	477.31	477.75
mse $\hat{t}_y^{\pi} / N$	663.12	264.75	2,013.90	981.54	142.93	53.17
Average $\hat{t}_y^{g} / N$	474.27	476.37	476.95	477.24	477.50	477.85
mse $\hat{t}_y^{g} / N$	218.96	66.66	114.08	50.10	121.57	41.32
$v_g / \text{mse}(\hat{t}_y^{g})$	0.76	0.87	0.73	0.82	0.66	0.91
$v_{wr} / \text{mse}(\hat{t}_y^{g})$	0.75	1.11	0.79	1.06	0.73	1.19
$v_{JL} / \text{mse}(\hat{t}_y^{g})$	0.88	1.16	0.85	1.10	0.78	1.24
$v_R / \text{mse}(\hat{t}_y^{g})$	0.87	1.15	0.82	1.08	0.74	1.22
$v_D / \text{mse}(\hat{t}_y^{g})$	1.26	1.32	1.09	1.25	0.95	1.36
$v_{J2} / \text{mse}(\hat{t}_y^{g})$	2.22	1.54	1.50	1.46	1.23	1.54
$v_{\text{Jack}} / \text{mse}(\hat{t}_y^{g})$	2.03	1.49	1.44	1.43	1.19	1.51
$v_{J1} / \text{mse}(\hat{t}_y^{g})$	2.22	1.55	1.56	1.49	1.28	1.57
$v_R^* / \text{mse}(\hat{t}_y^{g})$	0.71	0.73	0.67	0.68	0.60	0.74
$v_D^* / \text{mse}(\hat{t}_y^{g})$	1.02	0.83	0.88	0.79	0.76	0.83
$v_{J2}^* / \text{mse}(\hat{t}_y^{g})$	1.81	0.97	1.22	0.92	0.99	0.93
$v_{\text{Jack}}^* / \text{mse}(\hat{t}_y^{g})$	1.66	0.94	1.17	0.90	0.95	0.92
$v_{J1}^* / \text{mse}(\hat{t}_y^{g})$	1.81	0.98	1.27	0.94	1.03	0.95

Table 3.3 (continued)
Simulation Results for estimates for means and variance estimators for three populations and six sample designs in each population. Values in rows for variance estimators are ratios of mean estimated variance to empirical mse of the GREG. See Table 3.1 for descriptions of the variance estimators

Estimator	srs fixed		srs epsem		pps epsem	
	ACS Population (numbers in thousands)					
	$m = 3$	$m = 15$	$m = 3$	$m = 15$	$m = 3$	$m = 15$
Average $\hat{t}_y^{\pi} / N$	1,119.13	1,108.23	1,112.89	1,113.89	1,111.48	1,109.02
mse $\hat{t}_y^{\pi} / N$	181,329.24	27,650.01	201,618.77	32,926.98	15,991.69	2,619.32
Average $\hat{t}_y^{g} / N$	1,081.68	1,103.34	1,104.45	1,108.45	1,106.36	1,108.46
mse $\hat{t}_y^{g} / N$	11,220.86	921.82	2,111.84	408.19	1,874.39	352.65
$\upsilon_g / \text{mse}(\hat{t}_y^{g})$	2.70	0.90	0.44	0.83	0.53	0.92
$\upsilon_{wr} / \text{mse}(\hat{t}_y^{g})$	1.17	0.98	0.68	1.03	0.87	1.14
$\upsilon_{JL} / \text{mse}(\hat{t}_y^{g})$	2.18	0.91	0.65	0.99	0.79	1.11
$\upsilon_R / \text{mse}(\hat{t}_y^{g})$	2.80	1.00	0.43	0.92	0.53	1.03
$\upsilon_D / \text{mse}(\hat{t}_y^{g})$	6.09	1.32	0.84	1.08	0.89	1.15
$\upsilon_{J2} / \text{mse}(\hat{t}_y^{g})$	17,191.52	1.85	2.36	1.27	1.64	1.29
$\upsilon_{\text{Jack}} / \text{mse}(\hat{t}_y^{g})$	4,678.25	1.47	1.37	1.19	1.05	1.21
$\upsilon_{J1} / \text{mse}(\hat{t}_y^{g})$	17,190.86	1.72	3.07	1.36	2.35	1.38
$\upsilon_R^* / \text{mse}(\hat{t}_y^{g})$	2.66	0.76	0.41	0.70	0.49	0.68
$\upsilon_D^* / \text{mse}(\hat{t}_y^{g})$	5.79	0.99	0.80	0.82	0.83	0.76
$\upsilon_{J2}^* / \text{mse}(\hat{t}_y^{g})$	16,346.03	1.40	2.25	0.96	1.52	0.85
$\upsilon_{\text{Jack}}^* / \text{mse}(\hat{t}_y^{g})$	4,448.17	1.11	1.30	0.90	0.97	0.80
$\upsilon_{J1}^* / \text{mse}(\hat{t}_y^{g})$	16,345.41	1.30	2.92	1.03	2.19	0.91
	Simulated Population (numbers in millions)					
	$m = 300$	$m = 1{,}500$	$m = 300$	$m = 1{,}500$	$m = 300$	$m = 1{,}500$
Average $\hat{t}_y^{\pi} / N$	838.91	838.71	838.13	843.13	838.74	839.06
mse $\hat{t}_y^{\pi} / N$	1,588.43	250.20	2,303.19	563.77	1,218.73	253.13
Average $\hat{t}_y^{g} / N$	838.57	839.10	838.81	840.01	839.39	839.08
mse $\hat{t}_y^{g} / N$	156.29	23.07	117.18	19.63	105.64	25.24
$\upsilon_g / \text{mse}(\hat{t}_y^{g})$	0.91	1.11	0.91	1.13	1.01	0.89
$\upsilon_{wr} / \text{mse}(\hat{t}_y^{g})$	0.94	1.13	0.91	1.17	1.01	0.90
$\upsilon_{JL} / \text{mse}(\hat{t}_y^{g})$	0.91	1.13	0.92	1.15	1.02	0.90
$\upsilon_R / \text{mse}(\hat{t}_y^{g})$	0.91	1.13	0.92	1.14	1.02	0.90
$\upsilon_D / \text{mse}(\hat{t}_y^{g})$	1.03	1.15	0.96	1.16	1.07	0.91
$\upsilon_{J2} / \text{mse}(\hat{t}_y^{g})$	1.50	1.17	1.03	1.18	1.13	0.93
$\upsilon_{\text{Jack}} / \text{mse}(\hat{t}_y^{g})$	1.48	1.17	1.03	1.18	1.12	0.93
$\upsilon_{J1} / \text{mse}(\hat{t}_y^{g})$	1.50	1.17	1.03	1.18	1.13	0.93
$\upsilon_R^* / \text{mse}(\hat{t}_y^{g})$	0.90	1.07	0.91	1.09	1.01	0.85
$\upsilon_D^* / \text{mse}(\hat{t}_y^{g})$	1.02	1.09	0.96	1.11	1.05	0.86
$\upsilon_{J2}^* / \text{mse}(\hat{t}_y^{g})$	1.48	1.11	1.02	1.12	1.12	0.88
$\upsilon_{\text{Jack}}^* / \text{mse}(\hat{t}_y^{g})$	1.47	1.11	1.01	1.12	1.11	0.88
$\upsilon_{J1}^* / \text{mse}(\hat{t}_y^{g})$	1.48	1.11	1.02	1.13	1.12	0.88

The performance of the variance estimators depends on the sample design and the population. Some of the estimates in Table 3.3 from the ACS population with the simple random sample of 3 clusters and 9 units in each cluster stand out as being extremely poor. The inverses of the probabilities of selection vary quite a bit for this sample design. The variability of these weights, coupled with some extreme observations in the population, causes instability for some of the variance estimators. Namely, $\upsilon_{J2}, \upsilon_{\text{Jack}}, \upsilon_{J1}, \upsilon^*_{J2}, \upsilon^*_{\text{Jack}}, \upsilon^*_{J1}$ are extreme overestimates on average. All six of these estimators contain explicit or implicit hat matrix adjustments which can be quite large and seriously inflate the variance estimators when coupled with large sampling weights. On the other hand, υ_D, which also has a hat matrix adjustment, performs reasonably well for all populations and sample sizes. Noteworthy is the result that υ_D is much less of an overestimate for the mse in the combination (ACS, srs fixed, $m = 3$, $n_i = 9$) whereas other hat-matrix adjusted estimators were extreme overestimates. The estimators, $\upsilon_g, \upsilon_{wr}$, and to a lesser extent, υ_R and υ_{JL}, tend to be underestimates at the smaller sample sizes in the Third Grade and ACS populations and for all sample designs in those populations, but the problem diminishes for the larger sample sizes.



(a) Samples of size $m = 25$ (b) Samples of size $m = 50$

Figure 3.3 Boxplots of ratios of standard error estimates to the empirical standard errors for 1,000 SRS samples from Third Grade population. Vertical reference lines at 1.

The boxplots in Figure 3.3 show the variability of the estimators more clearly for srs's of size $m = 25$ and 50 from the Third Grade population. The boxplots depict the estimated standard errors (SEs) as a fraction of the empirical SE for the samples in each simulation. A ratio of 1 means that the estimated variance was equal to the empirical variance. Some samples yield large SE estimates, even though the majority of samples are much closer to the empirical variance. The degree of overestimation and the incidence of extreme values decreases substantially with the larger sample size as is evident by comparing the figures. The hat-matrix adjusted estimators also tend to somewhat overestimate the true variance, as evinced by the

boxes that are shifted above the reference lines drawn at 1. This can be an advantage for confidence interval coverage.

Table 3.4 shows the six-number summaries of the ratios of the SE estimates, $\sqrt{v}$, to the square root of the empirical variance, $\sqrt{v_E}$, for the Third Grade population for four of the sample designs. As indicated by the median value of the ratios for υ_{J2}, υ_{Jack}, υ_{J1}, υ^*_{J2}, υ^*_{Jack}, and υ^*_{J1}, they are generally centered near the empirical SEs, but can have extremely large values in some samples that affect their averages. (The problem of outlying values is even more severe in the ACS population; details are not shown here.) The estimators that are least affected by extremes are υ_g, υ_{wr}, υ_{JL}, υ_R, υ_D, υ^*_R, and υ^*_D. However, the estimators that incorporate *fpc*'s often are underestimates except in the case of srs and $m = 25$.

Table 3.4
Six-number summaries for alternative standard error estimators for Third Grade population in four sample designs. v_E is empirical variance across simulated samples. See Table 3.1 for descriptions of the variance estimators

	$\sqrt{v}$	Distribution of $\sqrt{v}/\sqrt{\upsilon_E}$					
		Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
srs $m = 25$	$\sqrt{\upsilon_g}$	0.46	0.71	0.82	0.86	0.96	3.59
	$\sqrt{\upsilon_{wr}}$	0.48	0.73	0.84	0.87	0.97	1.71
	$\sqrt{\upsilon_{JL}}$	0.48	0.75	0.88	0.92	1.03	3.75
	$\sqrt{\upsilon_R}$	0.47	0.74	0.87	0.92	1.02	3.85
	$\sqrt{\upsilon_D}$	0.53	0.84	1.00	1.08	1.20	6.84
	$\sqrt{\upsilon_{J2}}$	0.59	0.96	1.16	1.31	1.43	14.47
	$\sqrt{\upsilon_{\text{Jack}}}$	0.57	0.93	1.13	1.26	1.38	13.69
	$\sqrt{\upsilon_{J1}}$	0.59	0.97	1.17	1.32	1.44	14.48
	$\sqrt{\upsilon^*_R}$	0.42	0.67	0.79	0.83	0.92	3.48
	$\sqrt{\upsilon^*_D}$	0.48	0.76	0.90	0.97	1.08	6.17
	$\sqrt{\upsilon^*_{J2}}$	0.53	0.87	1.05	1.18	1.29	13.06
	$\sqrt{\upsilon^*_{\text{Jack}}}$	0.52	0.84	1.02	1.14	1.25	12.35
	$\sqrt{\upsilon^*_{J1}}$	0.54	0.88	1.06	1.19	1.30	13.07
srs $m = 50$	$\sqrt{\upsilon_g}$	0.62	0.84	0.92	0.94	1.01	1.64
	$\sqrt{\upsilon_{wr}}$	0.67	0.95	1.04	1.06	1.15	1.73
	$\sqrt{\upsilon_{JL}}$	0.68	0.96	1.06	1.08	1.18	1.94
	$\sqrt{\upsilon_R}$	0.68	0.96	1.06	1.07	1.17	1.95
	$\sqrt{\upsilon_D}$	0.71	1.01	1.13	1.15	1.26	2.20
	$\sqrt{\upsilon_{J2}}$	0.75	1.08	1.20	1.24	1.35	2.88
	$\sqrt{\upsilon_{\text{Jack}}}$	0.74	1.06	1.18	1.22	1.33	2.79
	$\sqrt{\upsilon_{J1}}$	0.75	1.09	1.21	1.24	1.36	2.86
	$\sqrt{\upsilon^*_R}$	0.54	0.76	0.84	0.85	0.93	1.55
	$\sqrt{\upsilon^*_D}$	0.56	0.80	0.89	0.91	1.00	1.75
	$\sqrt{\upsilon^*_{J2}}$	0.59	0.86	0.95	0.98	1.07	2.29
	$\sqrt{\upsilon^*_{\text{Jack}}}$	0.58	0.84	0.94	0.97	1.06	2.22
	$\sqrt{\upsilon^*_{J1}}$	0.60	0.86	0.96	0.99	1.08	2.27

Table 3.4 (continued)
Six-number summaries for alternative standard error estimators for Third Grade population in four sample designs. v_E is empirical variance across simulated samples. See Table 3.1 for descriptions of the variance estimators

	$\sqrt{v}$	Distribution of $\sqrt{v}/\sqrt{v_E}$					
		Min	1st Qu.	Median	Mean	3rd Qu.	Max
pps $m = 25$	$\sqrt{v_g}$	0.48	0.71	0.79	0.80	0.88	1.33
	$\sqrt{v_{wr}}$	0.51	0.76	0.84	0.84	0.92	1.30
	$\sqrt{v_{JL}}$	0.50	0.76	0.86	0.87	0.96	1.46
	$\sqrt{v_R}$	0.49	0.75	0.84	0.85	0.94	1.43
	$\sqrt{v_D}$	0.53	0.83	0.94	0.96	1.06	1.66
	$\sqrt{v_{J2}}$	0.59	0.94	1.06	1.09	1.21	2.15
	$\sqrt{v_{\text{Jack}}}$	0.57	0.92	1.04	1.07	1.18	2.10
	$\sqrt{v_{J1}}$	0.60	0.96	1.08	1.11	1.23	2.19
	$\sqrt{v_R^*}$	0.43	0.67	0.76	0.76	0.84	1.30
	$\sqrt{v_D^*}$	0.47	0.75	0.84	0.86	0.95	1.51
	$\sqrt{v_{J2}^*}$	0.52	0.84	0.95	0.98	1.08	1.90
	$\sqrt{v_{\text{Jack}}^*}$	0.51	0.82	0.93	0.96	1.06	1.86
	$\sqrt{v_{J1}^*}$	0.53	0.86	0.97	1.00	1.10	1.93
pps $m = 50$	$\sqrt{v_g}$	0.72	0.88	0.95	0.95	1.01	1.28
	$\sqrt{v_{wr}}$	0.78	1.00	1.09	1.09	1.16	1.47
	$\sqrt{v_{JL}}$	0.81	1.01	1.11	1.11	1.19	1.52
	$\sqrt{v_R}$	0.80	1.00	1.09	1.09	1.18	1.50
	$\sqrt{v_D}$	0.84	1.06	1.15	1.16	1.25	1.64
	$\sqrt{v_{J2}}$	0.88	1.11	1.22	1.23	1.33	1.83
	$\sqrt{v_{\text{Jack}}}$	0.88	1.10	1.21	1.22	1.31	1.81
	$\sqrt{v_{J1}}$	0.89	1.13	1.23	1.24	1.34	1.85
	$\sqrt{v_R^*}$	0.62	0.78	0.85	0.85	0.92	1.16
	$\sqrt{v_D^*}$	0.65	0.82	0.90	0.90	0.97	1.28
	$\sqrt{v_{J2}^*}$	0.68	0.87	0.95	0.96	1.03	1.43
	$\sqrt{v_{\text{Jack}}^*}$	0.67	0.86	0.94	0.95	1.02	1.42
	$\sqrt{v_{J1}^*}$	0.69	0.88	0.96	0.97	1.04	1.44

Lastly, Table 3.5 shows the 95% confidence interval coverage for all of the estimators based on t-distributions. That is, we computed, $\left[\hat{t}_y^{gr} - t_{0.975,\, m-1}\sqrt{v},\ \hat{t}_y^{gr} + t_{0.975,\, m-1}\sqrt{v}\ \right]$ where $t_{0.975,\, m-1}$ is the 97.5th percentile from a t-distribution with $m-1$ degrees of freedom. We then noted how often the true value fell below, above, and inside this range. In addition to the new and old estimators, Table 3.5 also shows the confidence interval coverage attained when the empirical variance, v_E, was used to form the confidence intervals. Ideally, the population total should be within the estimated 95% confidence interval for 95% of the samples. The true total should be below the 95% confidence bounds for 2.5% of the samples and above the confidence bounds for the same percentage of samples.

The jackknife-based estimators, v_D^*, v_{Jack}^*, and v_{J2}, cover at higher rates than the other variance estimators because they are larger. In small samples, jackknife-based estimators cover above the nominal level. The traditional variance estimators, v_g, v_{wr}, and v_{JL} under-covered in a number of cases, although their coverage was almost always higher than 90%. Note that v_g is generally an improvement over v_R due to the hat-matrix adjustment that makes v_D larger.

The variance estimators that incorporate hat matrix adjustments (v_D, v_{J2}, v_{Jack}, and v_R^*) generally increase CI coverage rates compared to the other choices. This advantage was especially noticeable for the ACS population where, for example, v_{wr} covers in less than 90% of samples in the combinations, (v_{Jack}^*, $m = 3$), (srs epsem, $m = 3$), and (srs epsem, $m = 15$). Although, in principal, an *fpc* would seem useful in some of the population and sample size combinations, CIs based on the variance estimators with *fpc*'s cover at lower rates than their counterparts without the *fpc*'s. For example, in ACS (srs epsem, $m = 15$) the coverage rates for v_R^*, v_D^*, v_{J2}^*, v_{Jack}^*, and v_{J1}^* range from 86.1 to 90.6% while the rates for the versions without *fpc*'s range from 90.2 to 93.4%.

Table 3.5
Coverage of 95% confidence intervals for population totals based on *t*-distributions and alternative variance estimators. See Table 3.1 for descriptions of the variance estimators

Variance est.	Third Grade			ACS			Simulation			Third Grade			ACS			Simulation		
	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper
	srs $m = 25$			srs $m = 3$			srs $m = 300$			srs $m = 50$			srs $m = 15$			srs $m = 1{,}500$		
v_E	2.9	95.6	1.5	0.7	99.3	0	2.7	95.0	2.3	3.4	95.1	1.5	3.3	95.8	1.0	1.0	96.0	3.0
v_a	7.4	90.7	1.9	2.4	97.3	0.4	4.3	93.5	2.2	5.9	92.8	1.3	6.6	92.3	1.0	1.0	95.0	4.0
v_{wr}	7.0	90.5	2.5	9.2	88.8	2.0	3.9	92.8	3.3	4.1	95.0	0.9	7.5	91.0	1.5	1.0	96.0	3.0
v_{JL}	5.5	93.2	1.3	6.5	92.1	1.4	4.4	93.4	2.2	3.3	96.1	0.6	7.2	91.4	1.4	1.0	95.0	4.0
v_R	5.9	92.7	1.4	3.1	96.3	0.6	4.3	93.5	2.2	3.4	96.0	0.6	6.5	92.5	1.0	1.0	95.0	4.0
v_D	3.8	95.4	0.8	1.6	98.0	0.4	3.7	94.2	2.1	2.4	97.1	0.5	5.1	94.3	0.6	1.0	95.0	4.0
v_{J2}	1.7	98.0	0.3	0.6	99.3	0.1	3.6	94.4	2.0	2.0	97.7	0.3	3.9	95.7	0.4	1.0	95.0	4.0
v_{Jack}	2.1	97.6	0.3	3.2	95.9	0.8	3.6	94.4	2.0	2.0	97.7	0.3	5.6	93.7	0.7	1.0	95.0	4.0
v_{J1}	1.6	98.1	0.3	1.6	98.0	0.3	3.6	94.4	2.0	2.0	97.7	0.3	4.5	95.0	0.5	1.0	95.0	4.0
v_R^*	8.6	89.4	2.0	3.4	96.0	0.7	4.4	93.4	2.2	7.8	89.8	2.4	9.5	88.5	2.0	1.0	95.0	4.0
v_D^*	5.5	93.3	1.2	1.6	98.0	0.4	3.8	94.1	2.1	6.4	92.2	1.4	7.5	91.1	1.4	1.0	95.0	4.0
v_{J2}^*	2.9	96.6	0.5	0.6	99.3	0.1	3.6	94.4	2.0	5.2	93.8	1	5.8	93.3	0.8	1.0	95.0	4.0
v_{Jack}^*	3.7	95.7	0.6	3.4	95.7	0.9	3.6	94.4	2.0	5.5	93.4	1.1	7.9	90.6	1.6	1.0	95.0	4.0
v_{J1}^*	2.7	96.9	0.4	1.7	97.9	0.4	3.6	94.4	2.0	5.0	93.9	1.1	6.6	92.3	1.1	1.0	95.0	4.0
	srs epsem $m = 25$			srs epsem $m = 3$			srs epsem $m = 300$			srs epsem $m = 50$			srs epsem $m = 15$			srs epsem $m = 1{,}500$		
v_E	1.7	96.2	2.1	0.0	99.9	0.1	2.4	94.7	2.9	2.3	95.5	2.2	1.1	97.1	1.8	3.0	94.0	3.0
v_a	5.6	91.2	3.2	6.5	91.5	2.0	2.6	94.1	3.3	5.1	92.2	2.7	8.3	90.4	1.3	3.0	96.0	1.0
v_{wr}	5.8	91.2	3.0	9.6	87.2	3.2	3.1	93.3	3.6	3.4	95.1	1.5	9.3	89.7	1.1	3.0	95.0	2.0
v_{JL}	5.1	92.4	2.5	6.5	91.2	2.3	2.6	94.1	3.3	2.8	96.0	1.2	8.2	90.9	0.9	3.0	96.0	1.0
v_R	5.2	92.3	2.5	8.4	88.3	3.3	2.6	94.1	3.3	2.9	95.7	1.4	8.8	90.2	1.0	3.0	96.0	1.0
v_D	3.7	94.3	2.0	5.5	92.8	1.7	2.5	94.3	3.2	2.3	96.9	0.8	7.8	91.6	0.7	3.0	96.0	1.0
v_{J2}	1.9	97.3	0.8	2.6	96.7	0.7	2.3	94.9	2.8	2.0	97.9	0.1	6.9	92.6	0.5	3.0	96.0	1.0
v_{Jack}	2.2	96.8	1.0	4.7	94.0	1.3	2.3	94.9	2.8	2.1	97.8	0.1	7.3	92.1	0.6	3.0	96.0	1.0
v_{J1}	1.8	97.5	0.7	2.5	96.9	0.6	2.3	94.9	2.8	2.0	97.9	0.1	6.2	93.4	0.4	3.0	96.0	1.0
v_R^*	6.6	89.5	3.9	8.9	87.8	3.4	2.7	93.9	3.4	7.7	88.7	3.6	11.7	86.1	2.2	3.0	96.0	1.0
v_D^*	5.1	92.5	2.4	5.7	92.4	1.9	2.5	94.3	3.2	6.0	91.6	2.4	10.6	88.0	1.5	3.0	96.0	1.0
v_{J2}^*	3.4	94.9	1.7	2.8	96.5	0.7	2.3	94.9	2.8	4.6	93.7	1.7	9.2	89.7	1.1	3.0	96.0	1.0
v_{Jack}^*	3.5	94.8	1.7	4.9	93.7	1.4	2.3	94.9	2.8	4.7	93.3	2	9.9	89.0	1.2	3.0	96.0	1.0
v_{J1}^*	3.0	95.4	1.6	2.6	96.8	0.6	2.3	94.9	2.8	4.6	93.7	1.7	8.6	90.6	0.8	3.0	96.0	1.0

Table 3.5 (continued)
Coverage of 95% confidence intervals for population totals based on *t*-distributions and alternative variance estimators. See Table 3.1 for descriptions of the variance estimators

Variance est.	Third Grade			ACS			Simulation			Third Grade			ACS			Simulation		
	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper
	pps $m = 25$			*pps* $m = 3$			*pps* $m = 300$			*pps* $m = 50$			*pps* $m = 9$			*pps* $m = 1{,}500$		
v_E	1.7	95.9	2.4	0	100.0	0.0	2.9	94.2	2.9	2.3	95.3	2.4	0.7	98.0	1.3	2.0	95.0	3.0
v_a	6.2	90.0	3.8	4.7	94.3	1.0	2.9	93.9	3.2	3.1	94.1	2.8	5.1	94.4	0.5	2.0	92.0	6.0
v_{wr}	5.1	91.1	3.8	5.6	92.8	1.5	3.1	93.6	3.3	2.0	97.0	1.0	5.3	94.3	0.4	3.0	92.0	5.0
v_{JL}	4.9	92.0	3.1	4.9	93.5	1.5	2.9	94.0	3.1	1.9	96.9	1.2	4.9	94.7	0.3	2.0	92.0	6.0
v_R	5.3	91.5	3.2	7.2	90.5	2.3	2.9	93.9	3.2	2.0	96.8	1.2	5.6	94.1	0.4	2.0	92.0	6.0
v_D	3.8	94.1	2.1	4.4	94.4	1.1	2.7	94.7	2.6	1.7	97.4	0.9	4.8	94.9	0.3	2.0	92.0	6.0
v_{J2}	2.7	96.1	1.2	2.6	97.0	0.4	2.6	95.0	2.4	1.6	97.9	0.5	4.3	95.5	0.2	2.0	92.0	6.0
v_{Jack}	2.8	95.8	1.4	4.2	94.9	0.9	2.6	95.0	2.4	1.6	97.9	0.5	4.7	95.1	0.2	2.0	92.0	6.0
v_{J1}	2.2	96.7	1.1	2.1	97.5	0.4	2.6	95.0	2.4	1.5	98.0	0.5	3.9	96.0	0.1	2.0	92.0	6.0
v_R^*	7.4	87.8	4.8	7.6	90.0	2.4	2.9	93.9	3.2	5.0	90.6	4.4	8.9	89.8	1.3	2.0	92.0	6.0
v_D^*	5.3	91.6	3.1	4.7	94.0	1.3	2.7	94.5	2.8	4.1	92.2	3.7	8.1	90.9	1.0	2.0	92.0	6.0
v_{J2}^*	3.6	94.3	2.1	2.8	96.8	0.4	2.6	95.0	2.4	3.0	94.1	2.9	7.2	92.0	0.7	2.0	92.0	6.0
v_{Jack}^*	4.0	93.7	2.3	4.5	94.5	1.0	2.6	95.0	2.4	3.1	94.0	2.9	7.9	91.1	1.0	2.0	92.0	6.0
v_{J1}^*	3.5	94.6	1.9	2.2	97.4	0.4	2.6	95.0	2.4	2.9	94.4	2.7	6.8	92.6	0.6	2.0	92.0	6.0

One feature of v_D and v_D^* is that both the cluster-specific contributions, $v_{D,i}$ and $v_{D,i}^*$, as well as the overall variance estimates can be negative. In the simulations, the adjustment described after (2.11) was used to avoid negative contributions. Negative estimates were more common when the second stage sample sizes were small and the weights were quite variable. For example, for the ACS population, almost 28% of the simple random samples of 3 clusters and $m_i = 9$ resulted in at least one negative variance contribution for a cluster. More commonly, about 10% of the samples contained at least one negative variance estimate for a cluster. In the Third Grade population, 16% to 27% of the samples had at least one negative value of v_{Di}. In the simulated population with large sample sizes, v_{Di} was negative in less than 5% of the samples. With the ad hoc correction of setting $I_i - H_{ii}$ to I_i, v_D is one of the most attractive variance estimators because it tends to slightly overestimate the empirical variance, has some of the best confidence interval coverage, and has reasonable variability compared to other variance estimators.

4 Conclusion

Leverage adjustments to standard variance estimators have been shown to reduce bias and improve confidence interval coverage based on general regression estimators in single-stage samples. This paper extends those results to two-stage samples by presenting new adjustments based on hat matrices. Our theory provides the justification for the adjustments and illustrates that some of the proposed estimators are related to the delete-a-cluster jackknife that is a common procedure in survey estimation.

To test the theory, we conducted a series of simulation studies on three populations designed to assess performance in a variety of situations. In a school population a large sampling fraction of first-stage units was used. In a second population, based on American Community Survey data, the effects of small sample

sizes were tested. In a third simulated population, we examined large sample performance. Both simple random sampling and probability proportional to size sampling of clusters were used.

The relationships of the variance estimators were similar across all sample designs. The with-replacement variance estimator, υ_{wr}, which is the default choice in survey software packages, the jackknife linearization estimator, υ_{JL}, and the design-based variance estimator, υ_g, that assumes Poisson sampling at each stage as a computational convenience, are often negatively biased leading to confidence intervals that cover at less than the desired rate. Some of the jackknife-related estimators – υ_{Jack}, υ_{J1}, and υ_{J2} – which explicitly or implicitly include hat-matrix adjustments, are prone to producing large, outlying values when the first-stage sample is small. This is especially true when the first-stage is selected by *srs* but is less so in *pps* sampling when an efficient measure of size is used.

The variance estimators proposed here, particularly υ_D, provide alternatives to estimating the variance of GREG estimators in complex samples. At the expense of somewhat inflating the variability of the variance estimator, the hat-matrix adjusted sandwich estimators, denoted here by v_D, v_{J1}, and v_{J2}, give confidence interval coverage that is closer to the nominal value in small to moderate samples. Depending on the sample design and population characteristics, hat-matrix adjusted estimators can produce less biased variance estimates and better inferences when compared to the standard methods.

Acknowledgements

The authors thank the associate editor and two referees whose comments substantially improved the presentation.

Appendix

Theoretical results

A.1 Assumptions

The assumptions used to obtain asymptotic results are listed below. The number of population and sample clusters approach infinity; however, the number of population clusters increases at a faster rate than the number of sample clusters. Certain population quantities are assumed to be bounded.

A.1.1 $m/M \to 0$ *as* $m \to \infty$ *and* $M \to \infty$.

A.1.2 *All* N_i *and* n_i *are bounded.*

A.1.3 $\pi_{ik} = O(m/M)$ *for all* ik.

A.1.4 *All elements of* $\mathbf{X}$, $\mathbf{\Psi}$, *and* $\mathbf{Q}$ *are bounded.*

A.1.5 *The sample design is such that* $\frac{\sqrt{m}}{M}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_{Ux}\right) \xrightarrow{d} N(0, \mathbf{V})$, *where* $\mathbf{V}$ *is a* $p \times p$ *positive definite matrix, i.e.,* $\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_{Ux}\right) = O_p\left(M/\sqrt{m}\right)$.

Since $\mathbf{\Pi} = O\left(\frac{m}{M}\right)$ elementwise and $\mathbf{A} = \mathbf{X}_s^\top \mathbf{Q}^{-1} \mathbf{\Pi}^{-1} \mathbf{X}_s$ can be written as the sum of n terms and n_i is bounded while $m \to \infty$, $\mathbf{A} = O(M)$. By definition $\mathbf{g}_i^\top = \mathbf{1}_{n_i} + \left(\mathbf{t}_{Ux} - \hat{\mathbf{t}}_{x\pi}\right)^\top \mathbf{A}^{-1} \mathbf{X}_i^\top \mathbf{Q}_i$. The second term in g_i is $O_p\left(m^{-1/2}\right)$; consequently $\mathbf{g}_i$ converges to a vector of 1's. Using $\mathbf{A} = O(M)$ along with assumptions A.1.3 and A.1.4, $\mathbf{H}_{ij}$ is $O\left(m^{-1}\right)$ elementwise.

A.2 Model variance of GREG

Let $\mathbf{y}_{si}$ be the vector of all sample elements in cluster i and $\mathbf{y}_i$ be the vector of all elements in cluster i. The variance of the GREG, with respect to the working model (2.1) is:

$$\begin{aligned}
\mathrm{var}_\xi\left(\hat{t}_y^{gr} - t_y\right) &= \mathrm{var}_\xi\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{y}_{si} - \sum_{i\in U} \mathbf{1}_{N_i}^\top \mathbf{y}_i\right) \\
&= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_{si} \mathbf{\Pi}_i^{-1} \mathbf{g}_i - 2\mathrm{cov}_\xi\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{y}_{si}, \sum_{i\in U} \mathbf{1}_{N_i}^\top \mathbf{y}_i\right) + \mathbf{1}_N^\top \mathbf{\Psi} \mathbf{1}_N.
\end{aligned}$$

Since $\sum_{i\in U} \mathbf{1}_i^\top \mathbf{y}_i = \sum_{i\in s} \mathbf{1}_i^\top \mathbf{y}_i + \sum_{i\in(U-s)} \mathbf{1}_i^\top \mathbf{y}_i$ and elements in different clusters are uncorrelated, we have,

$$\begin{aligned}
\mathrm{var}_\xi\left(\hat{t}_y^{gr} - t_y\right) &= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_{si} \mathbf{\Pi}_i^{-1} \mathbf{g}_i - 2\sum_{i\in s}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathrm{cov}_\xi\left(\mathbf{y}_{si}, \mathbf{y}_i\right) \mathbf{1}_{N_i}\right] + \mathbf{1}_N^\top \mathbf{\Psi} \mathbf{1}_N \\
&= L_1 - 2L_2 + L_3.
\end{aligned}$$

Since $\mathbf{A}^{-1} = O\left(M^{-1}\right)$ and $\mathbf{g}_i$ and $\mathbf{\Psi}_{si}$ are bounded, we have $L_1 = O\left(M^2/m\right)$. Because $\mathbf{\Psi}_{si}$ is bounded, $\mathrm{cov}_\xi\left(\mathbf{y}_{si}, \mathbf{y}_i\right) = O(1)$ and $L_2 = O(M)$. L_3 is the sum of N terms. Since the N_i are bounded, $L_3 = O(M)$. Thus, L_1 is the dominant term of the prediction variance.

A.3 Proof that $\mathrm{var}_\xi\left(\mathbf{e}_i\right) \approx \mathbf{\Psi}_{si}$

In this section in order to simplify the notation, we omit the subscript s on $\mathbf{y}_{si}$, $\hat{\mathbf{y}}_{si}$, and $\mathbf{\Psi}_{si}$. The residual can be written in terms of a hat matrix as follows.

$$\begin{aligned}
\mathbf{e}_i &= \mathbf{y}_i - \hat{\mathbf{y}}_i \\
&= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j\neq i;\, i,\, j\in s} \mathbf{H}_{ij} \mathbf{y}_j
\end{aligned}$$

where $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix. The model variance of $\mathbf{e}_i$ is then

$$\begin{aligned}
\mathrm{var}_\xi\left(\mathbf{e}_i\right) &= \mathrm{var}_\xi\left[\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j\neq i} \mathbf{H}_{ij} \mathbf{y}_j\right] \\
&= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right) \mathrm{var}_\xi\left(\mathbf{y}_i\right)\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i} \mathbf{H}_{ij} \mathrm{var}_\xi\left(\mathbf{y}_j\right) \mathbf{H}_{ij}^\top \\
&= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right) \mathbf{\Psi}_i \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i} \mathbf{H}_{ij} \mathbf{\Psi}_j \mathbf{H}_{ij}^\top. \qquad \text{(A.1)}
\end{aligned}$$

As noted above, $\mathbf{H}_{ii} = O\left(m^{-1}\right)$. Thus, $\mathrm{var}_\xi\left(\mathbf{e}_i\right) = \mathbf{\Psi}_i + O\left(m^{-1}\right)$.

To justify υ_D, note that the second term of (A.1) can be written as

$$\sum_{j\neq i} \mathbf{H}_{ij} \mathbf{\Psi}_j \mathbf{H}_{ij}^\top = \sum_{j\in s} \mathbf{H}_{ij} \mathbf{\Psi}_j \mathbf{H}_{ij}^\top - \mathbf{H}_{ii} \mathbf{\Psi}_i \mathbf{H}_{ii}^\top.$$

The sum over the full cluster sample is

$$\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = \mathbf{X}_i\mathbf{A}^{-1}\left(\sum_{j\in s}\mathbf{X}_j^\top\mathbf{Q}_j\boldsymbol{\Pi}_j^{-1}\boldsymbol{\Psi}_j\boldsymbol{\Pi}_j^{-1}\mathbf{Q}_j\mathbf{X}_j\right)\mathbf{A}^{-1}\mathbf{X}_i^\top.$$

In the special case of $\mathbf{Q}_j = \boldsymbol{\Psi}_j^{-1}$ and $\boldsymbol{\Pi}_i = c\mathbf{I}_{n_i}$ for some constant $c \in (0, 1)$ (i.e., the sample is self-weighting), we have

$$\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = c^{-2}\mathbf{X}_i\mathbf{A}^{-1}\left(\sum_{j\in s}\mathbf{X}_j^\top\boldsymbol{\Psi}_j^{-1}\mathbf{X}_j\right)\mathbf{A}^{-1}\mathbf{X}_i^\top,$$

along with $\mathbf{H}_{ii} = c\mathbf{X}_i\mathbf{A}^{-1}\mathbf{X}_i^\top\boldsymbol{\Psi}_i^{-1}$ and $\mathbf{A} = c^{-1}\mathbf{X}\boldsymbol{\Psi}^{-1}\mathbf{X}$. Using these simplifications, $\sum_{j\in s}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top = \mathbf{H}_{ii}\boldsymbol{\Psi}_i$. Substituting this result in (A.1) and simplifying gives

$$\begin{aligned}\text{var}_\xi(\mathbf{e}_i) &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^\top + \sum_{j\neq i}\mathbf{H}_{ij}\boldsymbol{\Psi}_j\mathbf{H}_{ij}^\top \\ &= \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\boldsymbol{\Psi}_i. \qquad \text{(A.2)}\end{aligned}$$

This is the basis for the adjustment of υ_R to obtain υ_D.

A.4 Proof that $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} - \mathbf{R}_i$ for cluster samples

In this section, we omit the subscript s on $\mathbf{X}_s$, $\mathbf{y}_s$, $\mathbf{X}_{si}$, $\mathbf{y}_{si}$, $\mathbf{X}_{s(i)}$, and $\mathbf{y}_{s(i)}$ to simplify the notation. The subscript (i) denotes removal of the i^{th} cluster from the full sample matrix or vector. For example, $\hat{\mathbf{B}}_{(i)}$ is an estimate of $\mathbf{B}$ based on all sample clusters except cluster i and is

$$\hat{\mathbf{B}}_{(i)} = \left(\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}$$

where $\mathbf{W}_{(i)} = \mathbf{Q}_{(i)}\boldsymbol{\Pi}_{(i)}^{-1}$. Using Lemma 9.5.1 in Valliant et al. (2000), we have

$$\hat{\mathbf{B}}_{(i)} = \left(\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}^{-1}\right)\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}.$$

Since $\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)} = \mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i^\top\mathbf{W}_i\mathbf{y}_i$ and $\hat{\mathbf{B}} = \mathbf{A}^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{y}$, we have

$$\begin{aligned}\hat{\mathbf{B}}_{(i)} &= \mathbf{A}^{-1}\left(\mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i^\top\mathbf{W}_i\mathbf{y}_i\right) \\ &\quad + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}^{-1}\left(\mathbf{X}^\top\mathbf{W}\mathbf{y} - \mathbf{X}_i\mathbf{W}_i\mathbf{y}_i\right) \\ &= \hat{\mathbf{B}} - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)\mathbf{y}_i + \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\hat{\mathbf{y}}_i \\ &\quad - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{H}_{ii}\mathbf{y}_i \\ &= \hat{\mathbf{B}} - \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i.\end{aligned}$$

That is, $\hat{\mathbf{B}}_{(i)} = \hat{\mathbf{B}} - \mathbf{R}_i$.

A.5 Jackknife variance estimator of clustered GREG in terms of leverages

We now simplify the delete-a-cluster Jackknife variance estimator of the clustered GREG. As in Sections A.3 and A.4, we omit the subscript s on various terms to simplify the notation. The estimated total after removing the i^{th} cluster is defined as

$$\begin{aligned}
\hat{t}_{y(i)}^{gr} &= \frac{m}{m-1}\hat{t}_{y(i)}^{\pi} + \left[\mathbf{t}_{Ux} - \frac{m}{m-1}\hat{\mathbf{t}}_{x(i)}^{\pi}\right]\hat{\mathbf{B}}_{(i)} \\
&= \frac{m\mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{y}}{m-1} - \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} + \left[\mathbf{1}_N^\top \mathbf{X}_U - \frac{m\mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{X}}{m-1} + \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i}{m-1}\right]\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) \\
&= \frac{m\mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{y}}{m-1} - \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} \\
&\quad + \frac{m}{m-1}\left(\mathbf{1}_N^\top \mathbf{X}_U - \mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{X}\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) - \frac{1}{m-1}\left(\mathbf{1}_N^\top \mathbf{X}_U - m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\right)\left(\hat{\mathbf{B}} - \mathbf{R}_i\right) \\
&= \frac{m}{m-1}\hat{t}_y^{gr} - \frac{m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{y}_i}{m-1} - \frac{m}{m-1}\left(\mathbf{1}_N^\top \mathbf{X}_U - \mathbf{1}_n^\top \boldsymbol{\Pi}^{-1}\mathbf{X}\right)\mathbf{R}_i - \frac{1}{m-1}K_i.
\end{aligned}$$

Adding and subtracting $\frac{m}{m-1}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i$ and doing a substantial amount of simplification leads to

$$\hat{t}_{y(i)}^{gr} = \frac{m}{m-1}\hat{t}_y^{gr} - \frac{m}{m-1}\mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i + \frac{m}{m-1}G_i - \frac{1}{m-1}K_i.$$

Taking the difference between the delete-one estimates and the average of those estimates gives

$$\begin{aligned}
\hat{t}_{y(i)}^{gr} - \hat{t}_{y(\cdot)}^{gr} &= -\frac{m}{m-1}\left(D_i - \bar{D}\right) + \frac{m}{m-1}\left(G_i - \bar{G}\right) - \frac{1}{m-1}\left(K_i - \bar{K}\right) \\
&= -\frac{m}{m-1}\left(D_i - \bar{D}\right) + \frac{m}{m-1}\left[\left(G_i - \bar{G}\right) - \frac{1}{m}\left(K_i - \bar{K}\right)\right].
\end{aligned}$$

Letting $F_i = \left(G_i - \bar{G}\right) - m^{-1}\left(K_i - \bar{K}\right)$ leads to the formula for υ_{Jack} in equation (2.12). Next, since $\mathbf{H}_{ii} = O\left(m^{-1}\right)$ and $\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\mathbf{B}}$,

$$\begin{aligned}
F_i &= \left(G_i - \bar{G}\right) - \frac{1}{m}\left(K_i - \bar{K}\right) \\
&\approx \left[-\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\hat{\mathbf{y}}_i + \frac{1}{m}\sum_{i\in s}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\hat{\mathbf{y}}_i\right] - \frac{1}{m}\left[-m\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\hat{\mathbf{B}} + \sum_{i\in s}\mathbf{1}_{n_i}^\top \boldsymbol{\Pi}_i^{-1}\mathbf{X}_i\hat{\mathbf{B}}\right] \\
&= \mathbf{0}.
\end{aligned}$$

Thus, $F_i = o(1)$, and υ_{Jack} in (2.6) and (2.12) is asymptotically equivalent to υ_{J1} in (2.13).

Finally, to justify υ_{J2} in (2.14), we write υ_{J1} in the computational form

$$\upsilon_{J1} = \frac{m}{m-1}\left[\sum_{i\in s}\left(\mathbf{g}_i^\top U_i\mathbf{e}_i\right)^2 - \frac{1}{m}\left(\sum_{i\in s}\mathbf{g}_i^\top U_i\mathbf{e}_i\right)^2\right] \tag{A.3}$$

where $U_i = \boldsymbol{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}$. Note that the model variance of D_i is

$$\begin{aligned}
\text{var}_\xi\left(D_i\right) &= \text{var}_\xi\left(\mathbf{g}_i^\top U_i\mathbf{e}_i\right) \\
&= \mathbf{g}_i^\top U_i^\top \text{var}_\xi\left(\mathbf{e}_i\right)U_i\mathbf{g}_i.
\end{aligned}$$

Because $U_i = O\left(M/m\right)$ and the sum in $\sum_{i\in s}\text{var}_\xi\left(D_i\right)$ contains $n = m\bar{n}$ terms, the variance of $\sum_{i\in s}\mathbf{g}_i^\top U_i\mathbf{e}_i$ is $O\left(M^2/m\right)$. Next, scaling υ_{J1} to be appropriate for a mean, the first term in the brackets

in (A.3) is $N^{-2}\sum_{i\in s} D_i^2 = O(m^{-1})$. Since the second term in brackets has model expectation 0 and variance that is $O(m^{-1})$, it converges in probability to 0, and v_{J2} is asymptotically equivalent to v_{J1}.

A.6 Asymptotic equivalence of variance estimators

In this appendix we sketch arguments for why several variance estimators are asymptotically equivalent. Using design-based arguments, Yung and Rao (1996, Appendix) showed that the jackknife linearization estimator, v_{JL}, for the GREG is asymptotically equivalent to the design-consistent estimator, v_{Jack}, in stratified multistage designs with a large number of strata and a bounded number of sample clusters selected from each stratum. Using regularity conditions in Rao and Shao (1985), that result can be extended to cover designs in which either (i) the number of strata is large and the number of clusters per stratum is bounded or (ii) the number of strata is limited and the number of sample clusters per stratum is large, as is the case in this article.

The jackknife linearization estimator in Section 2 can be expanded as

$$N^{-2}v_{JL} = N^{-2}\sum_{i\in s}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{e}_i\mathbf{e}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{g}_i - N^{-2}m\left(m^{-1}\sum_{i\in s}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{e}_i\right)^2. \quad \text{(A.4)}$$

The first term in (A.4) equals v_R. Because, under some reasonable assumptions, $\mathbf{g}_i$ and $\mathbf{e}_i$ are bounded, and $\mathbf{\Pi}_i^{-1} = O(M/m)$ by assumptions A.1.2 and A.1.3, the first term in (A.4) is $O(1/m)$. The second term is also $O(1/m)$, but the model expectation of $\bar{\mathbf{e}}_2 = m^{-1}\sum_{i\in s}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{e}_i$ is zero as long as (2.1) holds. Since $\bar{\mathbf{e}}_2$ is a mean, its model-variance will approach 0 as $m\to\infty$. Thus, the second term in (A.4) will converge in probability to 0 and $v_{JL}\approx v_R$.

In Section A.5 it was shown that v_{Jack} and v_{J1} are asymptotically equivalent. Under A.1.1-A.1.4, $\mathbf{H}_{ii} = O(m^{-1})$. Consequently, v_{J2} and v_D are approximately the same as v_R as $m\to\infty$. Thus, $v_{\text{Jack}}\approx v_{JL}$ by extension of Yung and Rao (1996), both of which are design-consistent. Further, v_{JL} is asymptotically equivalent to v_{J1}, v_{J2}, v_D, and v_R. As a result, the alternative variance estimators considered here all have both model-based and design-based justifications.

References

Kott, P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika*, 75(4), 797-799.

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*, 9(5), 1010-1019.

Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35, 1, 15-24. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10881-eng.pdf.

Long, J.S., and Ervin, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3), 217-224.

MacKinnon, J.G., and White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305-325.

Rao, J.N.K., and Shao, J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80(391), 620-630.

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

Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.

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

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey Methodology*, 28, 1, 103-114. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002001/article/6424-eng.pdf.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley Series in Probability and Statistics: Survey Methodology Section. New York: John Wiley & Sons, Inc.

Yung, W., and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 1, 23-31. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996001/article/14388-eng.pdf.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 451-463
Statistics Canada, Catalogue No. 12-001-X

A note on propensity score weighting method using paradata in survey sampling

Seho Park, Jae Kwang Kim and Kimin Kim[1]

Abstract

Paradata is often collected during the survey process to monitor the quality of the survey response. One such paradata is a respondent behavior, which can be used to construct response models. The propensity score weight using the respondent behavior information can be applied to the final analysis to reduce the nonresponse bias. However, including the surrogate variable in the propensity score weighting does not always guarantee the efficiency gain. We show that the surrogate variable is useful only when it is correlated with the study variable. Results from a limited simulation study confirm the finding. A real data application using the Korean Workplace Panel Survey data is also presented.

Key Words: Unit Nonresponse; Smoothed weight; Surrogate variable.

1 Introduction

Paradata provides additional information on the quality of the collected survey data. The term paradata was coined by Couper (1998) to refer to the process data automatically generated from the data collection. It has been expanded to include various types of data about the data collection process in sample surveys (Kreuter, 2013).

One possibly useful paradata is the respondent behavior during the survey interview. Response time to survey can be one of the respondent behaviors. Knowles and Condon (1999) and Bassili (2003) found that response time has a negative correlation with the tendency of positive answer. It is called acquiescence bias (Couper and Kreuter, 2013). Longer response times were found to be an indicator of uncertainty and response error (Draisma and Dijkstra, 2004). Such paradata is helpful when we want to build a model for non-responses. Increasing non-response may cause non-response biases and has become a serious problem in recent years. Using the paradata that may be related to response model, non-response adjustment can be used to handle unit nonresponse effectively (Kott, 2006).

In addition to the auxiliary variables, Data Collection Process (DCP) variables are considered for estimation of non-response propensity (Beaumont, 2005). The DCP variable is treated as fixed in Holt and Elliot (1991) and the DCP variable, sometimes refer to the paradata, is used for non-response adjustment. On the other hand, Beaumont (2005) suggests to use DCP variable as a random variable and to be included in the non-response model. They show that using the paradata does not introduce additional bias and variance. Moreover, if the paradata variable is related to the study variable and the non-response, it reduces the non-response bias when the study variable is related to the non-response mechanism directly.

In our study, we show that using the paradata when it is conditionally independent with study variable given auxiliary variables inflates the variance as it brings unnecessary noise. While such phenomenon has

1. Seho Park, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, U.S.A. E-mail: seho.park@dartmouth.edu; Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-mail: jkim@iastate.edu; Kimin Kim, Korea Labor Institute, Sejong-si, 30147, Korea. E-mail: kimin1104@kli.re.kr.

been recognized in the literature (Little and Vartivarian, 2005), up to the knowledge of authors, it is not fully investigated theoretically. We investigate the effect of including the paradata into the nonresponse model using a rigorous theory.

This paper is motivated by a real survey data from Korean Workplace Panel Survey (KWPS). In the KWPS data, the reaction of the interviewee at the first contact was recorded during the data collection process. We investigate possible use of such paradata to enhance the quality of the data analysis.

The paper is organized as follows. In Section 2, basic setup is introduced and the main theoretical results are presented in Section 3. In Section 4, results of simulation studies are presented and a real data application is presented in Section 5. Concluding remarks are made in Section 6.

2 Basic setup

Consider a finite population of size N, where N is known. The finite population $\mathcal{F}_N = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N\}$, $\mathbf{u}_i = (x_i, y_i)$ is assumed to be a random sample from a superpopulation distribution $F(x, y)$. In addition, we assume that x is always observed and y is subject to missingness. Let δ be the response indicator function that takes the value one if y is observed and takes the value zero otherwise. Note that x, y and δ are all considered as random.

Suppose a sample of size n is drawn from the finite population using a probability sampling design, where inclusion in the sample is represented by the indicator variables I_i, with $I_i = 1$ if unit i is included in the sample and $I_i = 0$ otherwise. Let A be the index set of the sample and $w_i = \pi_i^{-1}$ be the design weight, where π_i is the first-order inclusion probability.

We are interested in estimating parameter θ that is implicitly defined through an estimating equation $E\{U(\theta; X, Y)\} = 0$. Under complete response, an estimator of θ is obtained by solving

$$\sum_{i \in A} w_i U(\theta; x_i, y_i) = 0.$$

In the presence of missing data, assuming that the response probabilities are known, the propensity-score adjusted estimator is obtained by solving

$$\sum_{i \in A} w_i \frac{\delta_i}{p_i} U(\theta; x_i, y_i) = 0, \tag{2.1}$$

where p_i is the response probability of unit i. Unfortunately, (2.1) is not applicable in practice because p_i are generally unknown.

Now suppose that there exists additional variable z obtained from paradata, which is always observed and satisfies

$$P(\delta_i = 1 | x_i, y_i, z_i) = P(\delta_i = 1 | x_i, z_i). \tag{2.2}$$

As x, y, δ and z are considered as random, we can use z to make inference about θ under nonresponse. Such variable z is sometimes called surrogate variable (Chen, Leung and Qin, 2008). By including a

suitable surrogate variable, we can make the response mechanism missing at random (MAR) in the sense of Rubin (1976). We call assumption (2.2) as the Augmented MAR (AMAR) since MAR holds only under the augmented model that includes surrogate variable z.

Under (2.2), we can build a parametric model for the response mechanism and construct a propensity score weighted (PSW) estimator that is obtained from

$$\sum_{i \in A} w_i \frac{\delta_i}{\hat{\pi}(x_i, z_i)} U(\theta; x_i, y_i) = 0,$$

where $\hat{\pi}(x_i, z_i)$ is a consistent estimator of $\pi(x_i, z_i) = P(\delta_i = 1 | x_i, z_i)$. Such PSW approach incorporating z variable has been discussed in Peress (2010) and Kreuter and Olson (2013).

In survey sampling, the surrogate variable z can be obtained from paradata which is not of direct interest. The information on z, however, can be helpful in making model assumptions for the response mechanism. In some cases, the surrogate variable z can satisfy

$$f(y | x, z) = f(y | x). \tag{2.3}$$

Condition (2.3) means that the surrogate variable z is not related to the study variable y that is subject to missingness. The model satisfying (2.3) can be called the reduced outcome model. If condition (2.3) does not hold, we call $f(y | x, z)$ the full outcome model.

If condition (2.3) holds in addition to condition (2.2), we can use this information to obtain a more efficient PSW estimator. Note that, by (2.2) and (2.3), we can establish

$$\begin{aligned}
P(\delta = 1 | x, y) &= \int P(\delta = 1 | x, y, z) f(z | x, y) dz \\
&= \int P(\delta = 1 | x, z) f(z | x, y) dz \\
&= \frac{\int P(\delta = 1 | x, z) f(y | x, z) f(z | x) dz}{\int f(y | x, z) f(z | x) dz} \\
&= \frac{\int P(\delta = 1 | x, z) f(y | x) f(z | x) dz}{\int f(y | x) f(z | x) dz} \\
&= P(\delta = 1 | x),
\end{aligned}$$

where the second equality follows from assumption (2.2) and the fourth equality follows from assumption (2.3). Thus, assumption (2.2) and (2.3) imply

$$f(y | x, \delta = 1) = f(y | x). \tag{2.4}$$

Under the reduced model assumption (2.3), then we can use another type of PSW estimator of the form

$$\sum_{i \in A} w_i \frac{\delta_i}{\hat{\pi}_1(x_i)} U(\theta; x_i, y_i) = 0, \tag{2.5}$$

where $\hat{\pi}_1(x_i) = \int \hat{\pi}(x_i, z_i)\, \hat{f}(z_i \mid x_i)\, dz_i$ and $\hat{f}(z \mid x)$ is an estimated conditional density of z given x. The estimator obtained from (2.5) can be called the smoothed PSW estimator (Beaumont, 2008). Note that $\hat{\pi}_1(x)$ is the smoothed version of $\hat{\pi}(x, z)$ averaged over the conditional distribution $f(z \mid x)$.

The smoothed PSW estimator obtained by solving the equation (2.5) is justified under MAR condition in (2.4). In this case, use of paradata for nonresponse adjustment is not necessarily useful, which will be justified in Section 3.

3 Main result

We now establish the main result of the paper. We assume that the response indictor functions δ_i are independent of each other. To avoid unnecessary details, we assume that $P(\delta_i = 1 \mid x_i, z_i) = \pi(x_i, z_i)$ is a known function of (x_i, z_i). Let $\hat{\theta}_{\text{PSW}}$ be the PSW estimator of θ obtained from

$$U_1(\theta) \equiv \sum_{i \in A} w_i \frac{\delta_i}{\pi(x_i, z_i)} U(\theta; x_i, y_i) = 0. \tag{3.1}$$

Also, let $\hat{\theta}_{\text{PSW2}}$ be the smoothed PSW estimator of θ obtained from

$$U_2(\theta) \equiv \sum_{i \in A} w_i \frac{\delta_i}{\pi_1(x_i)} U(\theta; x_i, y_i) = 0, \tag{3.2}$$

where $\pi_1(x_i) = P(\delta_i = 1 \mid x_i)$.

Theorem 1 *Under the assumptions (2.2) and (2.3), the smoothed PSW estimator $\hat{\theta}_{PSW2}$ from (3.2) is asymptotically unbiased and has asymptotic variance smaller than that of $\hat{\theta}_{PSW}$ from (3.1). That is,*

$$V\left(\hat{\theta}_{\text{PSW}} \,\middle|\, \mathcal{F}_N\right) \geq V\left(\hat{\theta}_{\text{PSW2}} \,\middle|\, \mathcal{F}_N\right). \tag{3.3}$$

Proof. First note that

$$E(U_2 \mid \boldsymbol{\delta}_N, \mathcal{F}_N) = \sum_{i=1}^{N} \frac{\delta_i}{\pi_1(x_i)} U(\theta; x_i, y_i),$$

where $\boldsymbol{\delta}_N = (\delta_1, \ldots, \delta_N)$. Thus, asymptotic unbiasedness of $\hat{\theta}_{\text{PSW2}}$ can be easily established by

$$\begin{aligned}
E(U_2 \mid \mathcal{F}_N) &= E\{E(U_2 \mid \boldsymbol{\delta}_N, \mathcal{F}_N) \mid \mathcal{F}_N\} \\
&= \sum_{i=1}^{N} E\left\{\frac{\delta_i}{\pi_1(x_i)} U(\theta; x_i, y_i) \mid x_i, y_i\right\} \\
&= \sum_{i=1}^{N} \frac{E(\delta_i \mid x_i, y_i)}{\pi_1(x_i)} U(\theta; x_i, y_i) \\
&= \sum_{i=1}^{N} \frac{\pi_1(x_i)}{\pi_1(x_i)} U(\theta; x_i, y_i) \\
&= \sum_{i=1}^{N} U(\theta; x_i, y_i).
\end{aligned}$$

For (3.3), it is enough to show that

$$V(U_1 \mid \mathcal{F}_N) \geq V(U_2 \mid \mathcal{F}_N). \tag{3.4}$$

Note that

$$\begin{aligned} V(U_1) &= V\{E(U_1 \mid \boldsymbol{\delta}_N, \mathcal{F}_N) \mid \mathcal{F}_N\} + E\{V(U_1 \mid \boldsymbol{\delta}_N, \mathcal{F}_N) \mid \mathcal{F}_N\} \\ &= V\left\{\sum_{i=1}^{N} \frac{\delta_i}{\pi(x_i, z_i)} U(\theta; x_i, y_i) \mid \mathcal{F}_N\right\} \\ &\quad + E\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathrm{Cov}(I_i, I_j) \frac{\delta_i}{\pi(x_i, z_i)} \frac{\delta_j}{\pi(x_j, z_j)} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \,\middle|\, \mathcal{F}_N\right\} \\ &:= V_1 + V_2. \end{aligned}$$

Now, since the δ_i are independent,

$$V_1 = E\left\{\sum_{i=1}^{N}\left(\frac{1}{\pi(x_i, z_i)} - 1\right) U(\theta; x_i, y_i)^{\otimes 2} \,\middle|\, \mathcal{F}_N\right\},$$

where $B^{\otimes 2} = BB'$. Also, writing $\Delta_{ij} = \mathrm{Cov}(I_i, I_j)$,

$$\begin{aligned} V_2 &= E\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \Delta_{ij} \frac{\delta_i}{\pi(x_i, z_i)} \frac{\delta_j}{\pi(x_j, z_j)} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \,\middle|\, \mathcal{F}_N\right\} \\ &= E\left\{\sum_{i=1}^{N} w_i^2 \Delta_{ii} \frac{E(\delta_i \mid x_i, y_i, z_i)}{\pi(x_i, z_i)^2} U(\theta; x_i, y_i)^{\otimes 2} \,\middle|\, \mathcal{F}_N\right\} \\ &\quad + E\left\{\sum_{i=1}^{N}\sum_{j\neq i} w_i w_j \Delta_{ij} \frac{E(\delta_i \mid x_i, y_i, z_i)}{\pi(x_i, z_i)} \frac{E(\delta_j \mid x_j, y_j, z_j)}{\pi(x_j, z_j)} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \,\middle|\, \mathcal{F}_N\right\} \\ &= E\left\{\sum_{i=1}^{N} (w_i - 1) \frac{1}{\pi(x_i, z_i)} U(\theta; x_i, y_i)^{\otimes 2} + \frac{1}{N^2} \sum_{i=1}^{N}\sum_{j\neq i} w_i w_j \Delta_{ij} U(\theta; x_i, y_i) U(\theta; x_j, y_j)'\right\}. \end{aligned}$$

Thus, combining the two results, we obtain

$$\begin{aligned} V(U_1) &= E\left[\sum_{i=1}^{N} w_i \left\{\frac{1}{\pi(x_i, z_i)} - 1\right\} U(\theta; x_i, y_i)^{\otimes 2} \,\middle|\, \mathcal{F}_N\right] \\ &\quad + E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \Delta_{ij} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \,\middle|\, \mathcal{F}_N\right]. \end{aligned} \tag{3.5}$$

Similarly, we can establish that

$$\begin{aligned} V(U_2) &= E\left[\sum_{i=1}^{N} w_i \left\{\frac{1}{\pi(x_i)} - 1\right\} U(\theta; x_i, y_i)^{\otimes 2} \,\middle|\, \mathcal{F}_N\right] \\ &\quad + E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \Delta_{ij} U(\theta; x_i, y_i) U(\theta; x_j, y_j)' \,\middle|\, \mathcal{F}_N\right]. \end{aligned} \tag{3.6}$$

Comparing (3.5) with (3.6), in order to show (3.4), we have only to show that

$$E\left\{\frac{1}{\pi(x,\, z)} \,\middle|\, x,\, y\right\} \;\geq\; \frac{1}{E\{\pi(x,\, z)\,|\,x,\, y\}}, \tag{3.7}$$

where $E\{\pi(x,\, z)\,|\,x,\, y\} = \pi_1(x)$. To show (3.7), note that $f(x) = 1/x$ is a convex function of $x \in (0,\, 1)$ and $\pi(\cdot)$ take values on (0, 1). We can apply Jensen's inequality to get

$$E\{f(\pi)\} \;\geq\; f\{E(\pi)\}, \tag{3.8}$$

which justifies (3.7). Here, the expectation in (3.8) is with respect to the conditional distribution of z given x and y.

By Theorem 1, under assumption (2.2) and (2.3), the smoothed PSW estimator $\hat{\theta}_{\text{PSW2}}$ leads to more efficient data analysis. Beaumont (2008) proposed the smoothed weighting for efficient estimation with survey data in a slightly different context, but the weight smoothing method of Beaumont (2008) matches with our finding when z is the design variable and δ is the sample indicator function. In this case, $P(\delta = 1\,|\,x,\, z)$ is the first order inclusion probability while $P(\delta = 1\,|\,x)$ is a smoothed version of the first order inclusion probability. Thus, if the sampling design is non-informative in the sense that $f(y\,|\,x,\, z) = f(y\,|\,x)$, then it is better to use the smoothed weight $\tilde{w}_i = \{P(\delta = 1\,|\,x)\}^{-1}$, which is consistent with the claims of Beaumont (2008) and Kim and Skinner (2013).

Under the reduced model (2.3), adding the surrogate variable z into the response propensity model can be regarded as including unnecessary noise and thus it generates inefficient estimation. For the case when the condition (2.3) is not satisfied, we can still use the smoothed PSW estimator using the weight obtained by weight smoothing conditioning on x_i, y_i, and $\delta_i = 1$, but the correct specification of the outcome model $f(y\,|\,x,\, z)$ can be challenging.

4 Simulation study

To test our theory, we perform a limited simulation study. In the simulation, we consider a situation when the augmented MAR assumption holds and check if including the surrogate variable in data analysis improves the efficiency of the final estimation.

We generate $B = 2{,}000$ Monte Carlo samples of size $n = 200$ from the outcome model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{4.1}$$

where $e_i \sim N(0,\, 1)$, $(\beta_0,\, \beta_1) = (1.2,\, 2.6)$, and $X_i \sim N(2,\, 1)$ for $i = 1, \ldots, n$.

In addition, we generate a surrogate variable Z from

$$z_i = 1 + x_i + u_i$$

with $u_i \sim N(0,\, 2^2)$. Thus, the surrogacy assumption (2.3) holds under this setup.

For the response probability, we consider the response model

$$\delta_i \sim \text{Bernoulli}(\pi_i),$$

where

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_i + \phi_2 z_i)}{1 + \exp(\phi_0 + \phi_1 x_i + \phi_2 z_i)} \tag{4.2}$$

and $(\phi_0, \phi_1, \phi_2) =$ (-1.2, 0.8, 0.4). Thus, the response mechanism satisfies the AMAR condition in (2.2). The overall response rate is 48% under this setup.

The parameters of interest are the regression coefficients in the outcome model (4.1) and the population mean of Y, $\mu = E(Y)$. We compare four methods for estimation of the parameters using Monte Carlo root mean squared error for the estimates. The four methods considered are as follows:

1. Complete case method (CC): Use the complete observations of (x_i, y_i) and estimate the parameters by the ordinary least squares method. That is, solve

$$\sum_{i=1}^{n} \delta_i U(\theta; x_i, y_i) = 0.$$

2. Propensity score weighting model method (PSW1): Use the estimated response rates as weights in estimating equation and solve the equation to estimate the parameters.
   (a) Fit a logistic regression model (4.2) for the response probability $\pi_i = \pi_i(x_i, z_i; \phi)$ and estimate $\phi = (\phi_0, \phi_1, \phi_2)$ by using the maximum likelihood method.
   (b) Parameter estimates are obtained by solving the estimating equation:

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} U(\theta; x_i, y_i) = 0,$$

   where $\hat{\pi}_i = \hat{\pi}\left(x_i, z_i; \hat{\phi}\right)$ and $\hat{\phi}$ is computed from Step (a).
3. Smoothed propensity score weighting model method (PSW2): Use the same procedure of PSW1, but the response probability is a function of explanatory variable (x) only. A response probability $\pi(x_i)$ is estimated as

$$\hat{\pi}_1(x_i) = \int \hat{\pi}(x_i, z_i)\, \hat{f}(z_i \mid x_i)\, dz_i,$$

   where $\hat{\pi}(x_i, z_i)$ is the estimated response probability in the PSW1 method. Since the estimated conditional density of z given x, $\hat{f}(z \mid x)$, is unknown, we use a nonparametric regression method for estimating $\hat{f}(z \mid x)$. Let $K_h(\cdot)$ be the kernel function satisfying certain regularity conditions and h be the bandwidth. Then, $\hat{\pi}_1(x_i)$ is obtained by

$$\hat{\pi}_1(x_i) = \frac{\sum_{j=1}^{n} \hat{\pi}(x_j, z_j) K_h(x_i, x_j)}{\sum_{j=1}^{n} K_h(x_i, x_j)}.$$

   We used the Gaussian Kernel for K_h with bandwidth $h = 1.06\, \hat{\sigma} n^{-1/5}$ chosen by the rule-of-thumb method of Silverman (1986).

4. Smoothed propensity score weighting estimator (PSW3) using logistic regression for estimating $\hat{\pi}_1(x_i)$: Use the same procedure of PSW1, but the response probability is estimated by a logistic regression model using only x_i.
   (a) Fit a logistic regression model for the response probability $\pi_i^* = \pi_i(x_i; \phi^*)$ as a function of explanatory variable (x_i) only and estimate $\phi^* = (\phi_0^*, \phi_1^*)$ by using the maximum likelihood method.
   (b) Parameter estimates are obtained by solving the estimating equation:

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i^*} U(\theta; x_i, y_i) = 0,$$

   where $\hat{\pi}_i^* = \hat{\pi}\left(x_i; \hat{\phi}^*\right)$ and $\hat{\phi}^*$ is computed from Step (a).

Table 4.1 presents the Monte Carlo biases, Monte Carlo standard errors and Monte Carlo root mean squared error of the four estimators of the three parameters, where the surrogate variable is uncorrelated with the study variable. Monte Carlo bias can be obtained by the difference between Monte Carlo mean and the true mean. Monte Carlo root squared mean squared error is the squared value of Monte Carlo mean squared error, which is a sum of squared Monte Carlo bias and Monte Carlo variance. As discussed in Section 3, the Monte Carlo root mean squared errors obtained using the smoothed propensity score weighting method (PSW2) are smaller than those of the propensity score weighting method (PSW1) as condition (2.3) is satisfied. The result confirms our theory that including the surrogate variable that is uncorrelated with the study variable may cause unnecessary noise for estimating parameters and decrease the efficiency. Also, PSW3 has larger variances due to model misspecification. Note that estimates of the regression coefficients under CC estimator are unbiased, whereas the estimator of the population mean of Y is biased.

Table 4.1
Monte Carlo biases (Bias), Monte Carlo standard errors (SE) and Monte Carlo root mean squared errors (RMSE) of point estimators

Parameter	Method	Bias	SE	RMSE
β_0	CC	0.005	0.213	0.213
	PSW1	0.001	0.247	0.247
	PSW2	0.003	0.232	0.232
	PSW3	0.005	0.239	0.240
β_1	CC	-0.001	0.088	0.088
	PSW1	0.000	0.102	0.101
	PSW2	0.001	0.096	0.096
	PSW3	-0.001	0.100	0.100
μ	CC	-0.555	0.206	0.592
	PSW1	-0.003	0.265	0.265
	PSW2	-0.049	0.219	0.225
	PSW3	0.007	0.233	0.233

5 Application

5.1 Data description

The research is motivated by real data analysis in Korean Workplace Panel Survey (KWPS) data, which is a biennial panel survey of the workplaces in Korea, sponsored by Korean Labor Institute. We used the KWPS data collected in 2007, 2009, and 2011 for our analysis.

The target population of the survey is all the companies located in South Korea with the size (= number of employees) greater than 30, except for agriculture, forestry, fishing and hunting industry. Of all the companies in the target population, which is of size 37,644 companies, 1,400 companies were selected using a stratified random sampling design.

The sampling design used for the survey is stratified sampling using the company as a sampling unit. The stratification variable is formed using 3 variables: the size of the company, the type of the company and the area where it is located. A combination of the three variables resulted in 200 strata since there are 5 levels of area, 4 levels of size of company, and 10 levels of type of company location.

From the KWPS data, we are interested in fitting a regression model for the regression of the log-scaled sales per person (Y = log(Sales)/Person) on two covariates of the company: size of company (X_1) and type of company (X_2). In the dataset, variable Y is not completely observed for all targets of the survey; they contain some missing values. However, the explanatory variables are completely observed as the size and the type of company are the characteristics that do not change easily in every two years.

The response variable (Y), the log-scaled sales per person, is a continuous variable. The two explanatory variables are categorical. The size of company variable (X_1) has four categories; 30-99 people, 100-299 people, 300-499 people, and more than 500 people. The type of company variable (X_2) contains ten categories: Light industry, chemical industry, electric/electronic industry, etc.

In the KWPS data, the variable regarding the reaction of interviewees at the first contact has been collected during the survey process and is considered as a surrogate variable in our analysis. The reaction at the first contact is categorical with three categories:

1. Friendly response $(Z = 1)$: the interviewee accepts the survey or answers the pre-questionnaire on the visit date.
2. Moderate response $(Z = 2)$: the interviewee cannot complete the survey immediately, but allows for a follow-up survey.
3. Negative response $(Z = 3)$: the interviewee who completes the survey uncooperatively or responds negatively.

Table 5.1 shows the response rates for each category of the first contact reaction. In friendly and moderate responses, response rates are 0.71 and 0.67, respectively, but the response rate for negative response is 0.45. This suggests that the surrogate variable is an important predictor for the response model.

Table 5.1
Response rate corresponding each level of reaction of interviewees

	Friendly Response	Moderate Response	Negative Response
Response Rate	0.71	0.67	0.45

From the dataset, we are interested in estimating the parameters in the regression model

$$E(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

5.2 Analysis

We first check whether condition (2.3) is satisfied. Using the idea of Fuller (1984), we test the hypothesis H_0: $\gamma = 0$ in the following model

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + e, \tag{5.1}$$

where $\mathbf{X} = (1, x_1, x_2)$ is a vector of explanatory variables, $\mathbf{Z}$ is a vector of surrogate variables, and e is a random error following $N(0, \sigma^2)$. Under H_0, we can roughly say that surrogate condition (2.3) is satisfied. Table 5.2 presents the result of the hypothesis testing. The F-statistic of the test is 0.3508 and its p-value is 0.7041, suggesting strong evidence in favor of the null hypothesis that the surrogate variables are not significant in the augmented regression model (5.1). Since the main stratification variable are included in X, the sampling design becomes noninformative (Pfeffermann, 1993). Thus, we can safely assume that the vector of surrogate variables $\mathbf{Z}$ can be treated as conditionally independent with the response variable Y given the explanatory variable X and condition (2.3) is satisfied.

Table 5.2
Test of the significance of the surrogate variable in the model (5.1)

	F statistic	p-value
H_0: $\gamma = 0$	0.3508	0.7041

Figure 5.1 also confirms the surrogacy condition (2.3). The median of three boxes seems to be almost the same around 0 and supports the result of the test that the surrogate variable is uncorrelated with response variable given explanatory variables. Hence, all of these results imply that assumption (2.3) holds for the data.

We now compare the three methods for estimating the parameters of the outcome model in (5.1), which are CC method, PSW1 method and PSW2 method. Estimated coefficients and their standard errors are presented in Table 5.3. The standard errors are calculated using bootstrap method for the stratified sampling (Rao and Wu, 1988) using $B = 1{,}000$ bootstrap replicates.

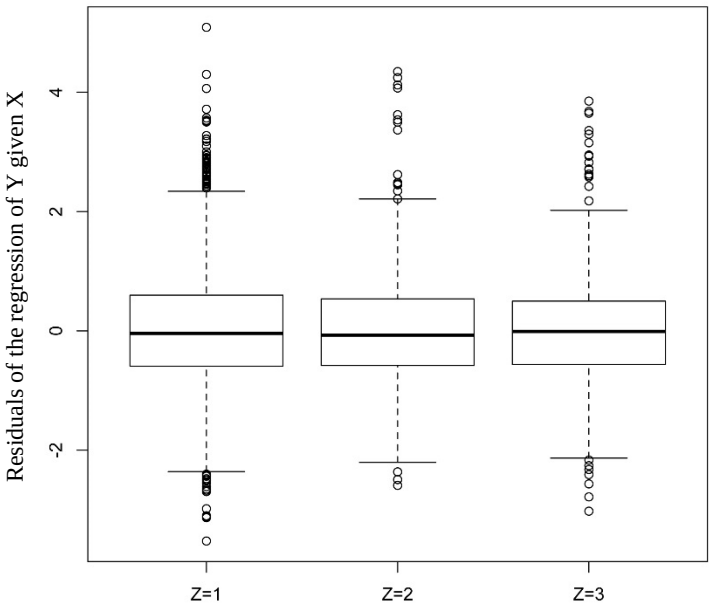


Figure 5.1 Boxplots of residuals of the regression of Y given X across each category of Z.

Table 5.3
Estimated coefficient (the standard error) from the real data analysis. (CC, complete case; PSW1, propensity score weighting method 1; PSW2, smoothed propensity score weighting method 2)

	CC	PSW1	PSW2
Intercept	5.404 (0.040)	5.408 (0.041)	5.405 (0.041)
100-299 people	0.170 (0.038)	0.165 (0.043)	0.170 (0.038)
300-499 people	0.401 (0.041)	0.342 (0.045)	0.401 (0.041)
> 500 people	0.587 (0.048)	0.528 (0.049)	0.587 (0.047)
Chemical	0.379 (0.051)	0.372 (0.053)	0.379 (0.052)
Metal/Auto	0.259 (0.045)	0.260 (0.047)	0.258 (0.046)
Elec/Electronic	-0.026 (0.051)	-0.024 (0.052)	-0.026 (0.052)
Construction	0.196 (0.075)	0.183 (0.078)	0.196 (0.077)
Personal Services	0.337 (0.055)	0.382 (0.057)	0.337 (0.054)
Transportation	-0.965 (0.064)	-0.917 (0.068)	-0.966 (0.063)
Financial Insur	-0.623 (0.073)	-0.577 (0.074)	-0.624 (0.071)
Social Services	-0.869 (0.061)	-0.839 (0.062)	-0.869 (0.060)
Elec/Gas	2.099 (0.061)	2.087 (0.059)	2.099 (0.061)

Chemical, Chemical Industry; Metal/Auto, Metal and Automobile Industry; Elec/Electronic, Electrical and Electronical Industry; Financial Insur, Finance and Insurance Services; Elec/Gas, Electric and Gas Services.

Since two explanatory variables are categorical with 4 and 10 levels, respectively, there are 13 coefficient parameters to be estimated. Table 5.3 presents the parameter estimates and their standard errors. We can see that the estimates obtained by using three methods are similar, but the standard errors obtained by using PSW2 are smaller than those of PSW1 across all levels of variables, although the efficiency gain by using PSW2 rather than PSW1 is not large. As indicated before, including the surrogate variable in calculating the propensity score weight generated unnecessary noise in estimation as the surrogate variable is uncorrelated with the study variable.

Although PSW2 shows better efficiency than PSW1, there is no real gain using PSW2 compared with CC method. Under MAR, the CC analysis provides the best estimator for the regression coefficient, although it leads to a biased estimation for the population mean or totals.

6 Conclusion

Motivated by the real survey project, we have investigated the propensity score approach incorporating the information from paradata into the response propensity model. Use of paradata in the propensity model has been advocated in the literature. However, it is not always the case. We find that using more information can decrease the efficiency of analysis, which is justified in Theorem 1. The claim is confirmed in the simulation study and the real data analysis using the KWPS data. When the surrogate variable in the paradata is conditionally independent with the study variable, conditional on the explanatory variable, it is better not to include the surrogate variable because the smoothed propensity score weight can provide more efficient estimation. In other words, it is useful to include the information from paradata only when the surrogate is correlated with the variable of interest.

References

Bassili, J.N. (2003). The minority slowness effect: Subtle inhibitions in the expression of views not shared by others. *Journal of Personality and Social Psychology*, 84(2), 261.

Beaumont, J.-F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31, 2, 227-231. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9049-eng.pdf.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95(3), 539-553.

Chen, S.X., Leung, D.H.Y. and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4), 803-823.

Couper, M.P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 48, 743-772.

Couper, M.P., and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 271-286.

Draisma, S., and Dijkstra, W. (2004). Response latency and (para) linguistic expressions as indicators of response error. *Methods for Testing and Evaluating Survey Questionnaires*, 131-147.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 1, 97-118. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1984001/article/14352-eng.pdf.

Holt, D., and Elliot, D. (1991). Methods of weighting for unit non-response. *The Statistician*, 333-342.

Kim, J.K., and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, 100(2), 385-398.

Knowles, E.S., and Condon, C.A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77(2), 379.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9547-eng.pdf.

Kreuter, F. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. New York: John Wiley & Sons, Inc.

Kreuter, F., and Olson, K. (2013). Paradata for nonresponse error investigation. In *Improving Surveys with Paradata: Analytic Uses of Process Information*, New York: John Wiley & Sons, Inc., 581, 13-42.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 2, 161-168. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9046-eng.pdf.

Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105(492), 1418-1430.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Rao, J.N.K., and Wu, C.F.J. (1988). Re-sampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC press.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 465-484
Statistics Canada, Catalogue No. 12-001-X

Suggestion of confidence interval methods for the Cronbach alpha in application to complex survey data

Jihnhee Yu, Ziqiang Chen, Kan Wang and Mine Tezal[1]

Abstract

We discuss a relevant inference for the alpha coefficient (Cronbach, 1951) - a popular ratio-type statistic for the covariances and variances in survey sampling including complex survey sampling with unequal selection probabilities. This study can help investigators who wish to evaluate various psychological or social instruments used in large surveys. For the survey data, we investigate workable confidence intervals by using two approaches: (1) the linearization method using the influence function and (2) the coverage-corrected bootstrap method. The linearization method provides adequate coverage rates with correlated ordinal values that many instruments consist of; however, this method may not be as good with some non-normal underlying distributions, e.g., a multi-lognormal distribution. We suggest that the coverage-corrected bootstrap method can be used as a complement to the linearization method, because the coverage-corrected bootstrap method is computer-intensive. Using the developed methods, we provide the confidence intervals for the alpha coefficient to assess various mental health instruments (Kessler 10, Kessler 6 and Sheehan Disability Scale) for different demographics using data from the National Comorbidity Survey Replication (NCS-R).

Key Words: Clustered data; Complex survey; Coverage-correction method; Influence function; Linearization.

1 Introduction

In this paper, we propose methods to incorporate the survey designs in confidence intervals for the alpha coefficient (Cronbach, 1951) based on the large sample approximation (linearization) and the "double" bootstrap approach. These methods have not been investigated in the related literature, even though the alpha coefficient is widely used in psychology and other relevant research areas. For a practical application of these methods, we analyze mental health instruments data from the National Comorbidity Survey Replication (NCS-R), a survey conducted between 2001 and 2003 intended to measure the prevalence of mental disorders (Kessler, Berglund, Chiu, Demler, Heeringa, Hiripi, Jin, Pennell, Walters, Zaslavsky and Zheng, 2004). In the analysis, we show the feasibility of the confidence interval method for the alpha coefficient on a survey data set.

A great deal of psychological and sociological research uses assessment instruments (i.e., questionnaires) to obtain quantitative information for a population of interest. Ideally, the different items in one instrument measure the same concepts to achieve a high internal consistency. The alpha coefficient, also known as Cronbach's alpha (henceforth referred to as α) is a popular statistic (e.g., a quick search of PubMed with the keywords "Cronbach alpha" and "scale" from the years of 2012-2016 brings up more than 700 publications) that is widely used to measure the internal consistency reliability of various instruments.

Let x denote the p-variate column vector of the observations indicating p items from an instrument, and let Σ indicate the corresponding covariance matrix. The value α is defined as

1. Jihnhee Yu, Department of Biostatistics, University at Buffalo, State University of New York, NY 14214, U.S.A. E-mail: jinheeyu@buffalo.edu; Ziqiang Chen and Kan Wang, Department of Biostatistics, University at Buffalo, State University of New York, NY 14214, U.S.A.; Mine Tezal, Department of Oral Biology, University at Buffalo, State University of New York, NY 14214, U.S.A.

$$\alpha = p/(p-1)\left(1 - \operatorname{tr}\Sigma/\mathbf{1}^T\Sigma\,\mathbf{1}\right),$$

where $\mathbf{1}$ is the conforming column vector consisting of 1, and tr indicates the trace of a matrix. The value α shows the ratio between the sum of the covariances and the sum of variances and covariances, thus a high value for α suggests that the items are highly correlated within the instrument. The theoretical values of α range from 0 to 1, where a higher value is considered to be more desirable. The estimator of α (denoted by $\hat{\alpha}$) is defined as

$$\hat{\alpha} = p/(p-1)\left(1 - \operatorname{tr}\hat{\Sigma}/\mathbf{1}^T\hat{\Sigma}\,\mathbf{1}\right),$$

where $\hat{\Sigma}$ is a consistent estimator of Σ. The estimator $\hat{\alpha}$ can take any value less than or equal to 1, including negative values.

In the literature, many confidence interval strategies for α can be found (e.g., van Zyl, Neudecker and Nel, 2000; Yuan, Guarnaccia and Hayslip, 2003; Kistner and Muller, 2004; Bonett and Wright, 2015), but discussions regarding the applications for complex survey data where observations in the data can have unequal weights due to stratifications and multistage cluster sampling (Lohr, 1999) are largely lacking.

This paper is structured as follows: In Section 2, we propose strategies for obtaining the confidence intervals of α using the linearization method and the coverage-corrected bootstrap method. In Section 3, simulation results are presented based on scenarios of stratified multi-stage cluster sampling and unequal probability sampling scenarios. In Section 4, the developed methods are applied to analyze the NCS-R data sets, and the results comparing different demographics are reported. The Section 5 is devoted to the concluding remarks.

2 Design-based confidence intervals for α

In this section, we discuss two methods to obtain the confidence interval for α, the confidence interval based on the linearization method using the influence function (Deville, 1999; Demnati and Rao, 2004) and the coverage-corrected bootstrap method (Hall, Martin and Schucany, 1989). In this discussion, we consider strategies to deal with stratification, since stratification is a common feature in surveys and may decrease the magnitude of the variances for the statistics of interest (Lohr, 1999). We note that the sampling design for the NCS-R used stratification (more details in Section 4). Later, in Section 3, we show that the linearization will be sufficient for most practical cases (e.g., scales with ordinal responses); however, the coverage rate may not be satisfactory with some non-normal distributions. The coverage-corrected bootstrap method when applied to survey data is proposed as a possible alternative to the linearization method in those cases (Section 2.2).

2.1 Linearization

A symmetric confidence interval can be obtained based on the normal approximation of an estimator for a finite population (Hájek, 1981; Sen, 1995). The linearization method is applied for the variance estimation

of complex statistics. In a survey sampling setting, we consider a population index set $U = \{1, \ldots, N\}$ with population size N. A random sample S of size n is selected from U by a sampling design $p(s) = \Pr\{S = s\}$ for all $s \subset U$. The value w_k denotes the sampling weight associated with the index $k \in s$. For probability sampling, the sampling weight for index k is the inverse of the first order inclusion probability, i.e., $w_k = [\Pr\{k \in s\}]^{-1}$. For each unit k of the population U, there is a point (or observation) x_k of $\mathbf{R}^p$, a p-dimensional real space. In a similar manner to Deville (1999), let us consider the population U that is represented by the measure M as having a mass of $1/N$ in each of the points x_k. In this way, we have $\int 1 dM = 1$ and $\int y dM = N^{-1}\sum_{k\in U} y_k$ for any vector value $y_k = y(x_k)$, where we define the integral of a vector as the integral of each component of the vector. The measure $\hat{M}$ is the estimator of M allocating a weight w_k/N to any point x_k, $k \in s$ and 0 to any other points. Following some conventional notation (e.g., Cochran, 1977), let $\int y dM = \bar{Y}$. Also let $\int y d\hat{M} = \hat{\bar{Y}}$. The influence function of a "functional" T is defined as

$$IT(M; x) = \lim_{t\to 0} \frac{T(M + t\delta_x) - T(M)}{t},$$

where δ_x denotes the added unit mass at point x (Deville, 1999), and the functional T (Krätschmer, Schied and Zähle, 2012) maps a measure to a set (e.g., the real line). The examples of the functional include $\bar{Y}$ and $\hat{\bar{Y}}$. Note that this classical definition of the influence function (Hampel, Ronchetti, Rousseeuw and Stahel, 1986; Davison and Hinkley, 1997) is slightly different from that of Deville (1999) where he defines a measure M to satisfy $\int y dM = \sum_{k\in U} y_k$. Let us define the linearized value $z_k = IT(M; x_k)$. Let $T(\hat{M})$ indicate the substitution estimator of $T(M)$ by replacing M by $\hat{M}$. Assume that the postulate of Deville (1999), i.e., $n^{-1/2}N^{-1}(\hat{X} - X)$ has a zero-mean multi-normal distribution as a limit, where X and $\hat{X}$ are the population total and the total estimator for general observation x_k, and N and n tend toward infinity. This fact leads to $\int x d\hat{M} - \int x dM = O_p(n^{-1/2})$. Assuming that T can be derived for any direction of an increase, a similar argument to Deville (1999) gives rise to the result

$$\frac{T(\hat{M}) - T(M)}{N^c} = \frac{1}{N^{c+1}} \sum_{k\in U} z_k (w_k - 1) + o(n^{-1/2}), \tag{2.1}$$

for some positive value c. Equation (2.1) results in the asymptotic variance of $T(\hat{M})$

$$\text{Avar}\{T(\hat{M})\} = \text{var}\left(\hat{\bar{Z}}\right). \tag{2.2}$$

If $T = \int x dM(x)$, then the influence function at x_k $(k \in U)$ is

$$IT(M; x_k) = \lim_{t\to 0} \frac{\sum_{i\in U} x_i/N + tx_k - \sum_{i\in U} x_i/N}{t} = x_k. \tag{2.3}$$

For a complex statistic as the functions of simple statistics, we have the influence function

$$I(f(T)) = D(f)\, IT, \tag{2.4}$$

where f is a differentiable function on the space of values for T and $D(f)$ is the matrix of the partial derivatives of f (Deville, 1999). In many cases, the linearized value z_k includes parameters to be estimated. Let $\hat{z}_k$ indicate the approximation of z_k using some statistics estimated by the sample. Deville (1999) notes that with a fixed and finite number of estimated parameters, the variance estimators based on $\hat{z}_k$ and z_k are equivalent by an asymptotically negligible quantity.

Now, we obtain the linearized value for α as follows. Consider a data set

$$\mathbf{X} = (x_1, \ldots, x_n)^T,$$

where x_k is a p-variate observation indicating p items in an instrument and n is the sample size. Let σ_{ij} and $\hat{\sigma}_{ij}$ $(i, j = 1, \ldots, p)$ denote the $(i, j)^{\text{th}}$ elements of Σ and $\hat{\Sigma}$ as defined in Section 1, respectively. Specifically, we define $\sigma_{ij} = \sum_{k=1}^{N} \left({}_i x_k - {}_i \bar{X}_k\right)\left({}_j x_k - {}_j \bar{X}_k\right)/(N-1)$ (Lohr, 1999), where ${}_i x_k$ and ${}_i \bar{X}_k$ are i^{th} element of x_k and its population mean, respectively. For simple random sampling without replacement (SRSWOR), we define $\hat{\sigma}_{ij} = \sum_{k\in s}\left({}_i x_k - {}_i \hat{\bar{X}}_k\right)\left({}_j x_k - {}_j \hat{\bar{X}}_k\right)/(n-1)$, where n is the size of the sample s and ${}_i \hat{\bar{X}}_k$ is the sample mean of ${}_i x_k$ (Lohr, 1999). For obtaining $\hat{\sigma}_{ij}$ for more complicated sampling methods including unequal probability sampling, we refer to Swain and Mishra (1994) and Patel and Bhatt (2016). In survey sampling, the sampling weights are used for correcting the disproportionality of the sample regarding the target population of interest (Pfeffermann, 1993). With the complex sampling designs often used in practice, failure to consider the sampling designs may provide biased inferences. For more of a discussion of the role of sampling weights, we refer to Pfeffermann (1993). For the variance estimation of survey data, the linearization method can be applied as in formula (2.2) incorporating the sampling weights. Following conventional notations of the vectorization of a matrix, let $\text{vech}(\mathbf{A})$ be the column vector of nonduplicated elements of the matrix $\mathbf{A}$, $\text{vec}(\mathbf{A})$ be the column vector composed of the columns of $\mathbf{A}$. Let $\hat{t}$ indicate the collection of statistics as the components of $\text{vech}^T\left(\hat{\Sigma}\right)$ and t indicate the collection of corresponding parameters. Specifically, we let $t = \left(\text{vech}^T\left(\sum_{k\in U} x_k x_k^T / N\right), \left(\sum_{k\in U} x_k / N\right)^T\right)$. Also, let the matrix K_p indicate a transition matrix that satisfies the relationship $\text{vech}(\mathbf{A}) = K_p^T \text{vec}(\mathbf{A})$, which borrows the transition matrix expression from van Zyl et al. (2000). We propose a linearized value for $\text{Var}(\hat{\alpha})$ as

$$z_k = \frac{p}{p-1}\frac{1}{\left(\mathbf{1}^T\hat{\Sigma}\mathbf{1}\right)^2}\left\{\mathbf{1}^T\hat{\Sigma}\mathbf{1}\,\text{vec}^T\left(\mathbf{I}_p\right) - \text{tr}\left(\hat{\Sigma}\right)\left(2\text{vec}^T\left(\mathbf{1}\mathbf{1}^T\right) - \text{vec}^T\left(\mathbf{I}_p\right)\right)\right\} K_p\,\hat{\mathbf{J}}\,u_k,\ k = 1, \ldots, n, \quad (2.5)$$

where a Jacobean matrix $\mathbf{J} = \partial\text{vech}^T(\Sigma)/\partial t$, $\hat{\mathbf{J}} = \left[\partial\text{vech}^T(\Sigma)/\partial t\right]_{\Sigma=\hat{\Sigma},\ t=\hat{t}}$, $u_k = \left(\text{vech}^T\left(x_k x_k^T\right),\ x_k^T\right)^T$ and $\mathbf{I}_p$ is the $p \times p$ identity matrix. We can now obtain the linearized value (2.5).

Derivation of (2.5): We consider the variance of $p/(p-1)\ \text{tr}\left(\hat{\Sigma}\right)/\mathbf{1}^T\hat{\Sigma}\mathbf{1}$ since its variance is the same as $\text{Var}(\hat{\alpha})$. Let $\alpha^* = p/(p-1)\ \text{tr}(\Sigma)/\mathbf{1}^T\Sigma\mathbf{1}$. Also, let $\text{vec}^T(\Sigma) = \left(\sigma_{11}, \ldots, \sigma_{p1}, \sigma_{12}, \ldots, \sigma_{p2}, \sigma_{13}, \ldots, \sigma_{pp}\right)$ and $\text{vech}^T(\Sigma) = \left(\sigma_{11}, \ldots, \sigma_{p1}, \sigma_{22}, \ldots, \sigma_{p2}, \sigma_{33}, \ldots, \sigma_{pp}\right)$, a p^2-vector and a $p(p+1)/2$-vector, respectively. Then, we have

$$\frac{\partial \alpha^*}{\partial t} = \frac{\partial \alpha^*}{\partial \text{vech}^T(\Sigma)} \mathbf{J} = \frac{p}{p-1} \frac{\partial}{\partial \text{vech}^T(\Sigma)} \left\{ \frac{\text{tr}(\Sigma)}{\mathbf{1}^T \Sigma \mathbf{1}} \right\} \mathbf{J}. \tag{2.6}$$

Now, in (2.6), we can show

$$\begin{aligned} \frac{\partial}{\partial \text{vech}^T(\Sigma)} \left\{ \frac{\text{tr}(\Sigma)}{\mathbf{1}^T \Sigma \mathbf{1}} \right\} &= \frac{1}{\mathbf{1}^T \Sigma \mathbf{1}} \frac{\partial \text{tr}(\Sigma)}{\partial \text{vech}^T(\Sigma)} - \frac{\text{tr}(\Sigma)}{\left(\mathbf{1}^T \Sigma \mathbf{1}\right)^2} \frac{\partial \mathbf{1}^T \Sigma \mathbf{1}}{\partial \text{vech}^T(\Sigma)} \\ &= \frac{1}{\left(\mathbf{1}^T \Sigma \mathbf{1}\right)^2} \left\{ \mathbf{1}^T \Sigma \mathbf{1} \, \text{vec}^T\left(\mathbf{I}_p\right) - \text{tr}(\Sigma) \left(2\text{vec}^T\left(\mathbf{1}\mathbf{1}^T\right) - \text{vec}^T\left(\mathbf{I}_p\right) \right) \right\} K_p. \end{aligned} \tag{2.7}$$

Using (2.6) and (2.7), we can obtain

$$\frac{\partial \alpha^*}{\partial t} = \frac{p}{p-1} \frac{1}{\left(\mathbf{1}^T \Sigma \mathbf{1}\right)^2} \left\{ \mathbf{1}^T \Sigma \mathbf{1} \, \text{vec}^T\left(\mathbf{I}_p\right) - \text{tr}(\Sigma) \left(2\text{vec}^T\left(\mathbf{1}\mathbf{1}^T\right) - \text{vec}^T\left(\mathbf{I}_p\right) \right) \right\} K_p \, \mathbf{J}. \tag{2.8}$$

Note that the expression (2.8) is a vector that consists of the derivatives of α^* with respect to the components of t. Each element in (2.8) is multiplied by the influence function corresponding to the statistics in t as in (2.4). This is accomplished by multiplying (2.8) by $u_k = \left(\text{vech}^T\left(x_k x_k^T\right), x_k^T\right)^T$, $k = 1, \ldots, n$, which is obtained by using (2.3). Now substituting Σ by $\hat{\Sigma}$ leads to the linearized value (2.5).

The formula for the new value (2.5) is easily implemented in the computer code using commonly available computer software. The relevant R code is available in the Supplementary Material.

We note that, in application to survey sampling, the estimate $\hat{\Sigma}$ should be obtained properly by incorporating the survey design. The variance is estimated by $\widehat{\text{Var}}\left(\hat{\bar{Z}}\right)$, where $\widehat{\text{Var}}$ indicates an operation to obtain the variance incorporating the weights and survey design properly, e.g., the Sen-Yates-Grundy variance estimator (Sen, 1953; Yates and Grundy, 1953), an unbiased variance estimator for the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) under designs with fixed sample sizes (e.g., Särndal, Swensson and Wretman, 1992) or the variance estimator for sampling with the replacement as a conservative approximation (Wolter, 1985). Specifically, in this paper, the variance for the NCS-R data is estimated as

$$\widehat{\text{Var}}\left(\hat{\bar{Z}}\right) = \sum_{h=1}^{H} \widehat{\text{Var}}_h\left(\hat{\bar{Z}}\right), \tag{2.9}$$

where $\widehat{\text{Var}}_h\left(\hat{\bar{Z}}\right)$ indicates the design-specific variance estimator for stratum $h\,(h = 1, \ldots, H)$. Once z_k values are obtained, standard statistical software for survey sampling such as R package "survey" (Lumley, 2004) can be used for the calculation of (2.9).

Now, consider a case that x is a random variable following a distribution and that an observation is a realization of the random variable; in addition, a sample of size n is obtained according to the random variable. In this specific case, we do not consider the finite population, where the design-based variance estimation is suitable as shown in the previous discussion. In a random variable setting, let $\hat{\bar{Z}}$ indicate the estimator with the measure $\hat{M}$ as the empirical distribution function (Fernholz, 1991). Employing the concept of a robust statistical inference based on the influence function (Davison and Hinkley, 1997), the sample variance for the population can be calculated by

$$\widehat{\mathrm{Var}}(\hat{\bar{Z}}) = n^{-1}\sum_{k=1}^{n}(z_k - \bar{z})^2 \Big/ (n-1), \tag{2.10}$$

where z_k is the linearized value (2.5) obtained from the statistic $\hat{\alpha}$ based on the sample (size n) and $\bar{z}$ is the sample mean of z_k $(k = 1, \ldots, n)$. We also note that the formula (2.10) is not constructed for infinite populations in survey methodology, where the finite population is seen as a realization from an infinite population. In that case, the outcomes of a statistical model give rise to the values of the characteristics of interest in the finite population, thus the model-based variance estimation is appropriate (Binder and Roberts, 2009). The formula (2.10) can be used for a general data analytical setting, where observations are considered as realizations of a random variable.

2.2 The coverage-corrected bootstrap method

The linearization provides reasonable estimates for the confidence intervals; however, in some cases, the coverage rate may not be satisfactory when the underlying distributions are non-normal (see Section 3). In these cases, some computer-intensive approaches such as the double bootstrap method, which is also called the coverage-corrected bootstrap may be implemented (Hall et al., 1989). We primarily discuss the double bootstrap method instead of the typical "single" bootstrap method (DiCiccio and Romano, 1988) since we observe that the single bootstrap method may not be satisfactory with non-normal underlying distributions (e.g., lognormal distribution) in terms of the coverage rate (Table 3.3).

For adjusting the bootstrap weight, the rescaling method referred to as the Rao-Wu bootstrap (Rao and Wu, 1988) is a popular approach for analyzing a lot of survey data, e.g., from Statistics Canada surveys (Mach, Saïdi and Pettapiece, 2007). The Rao-Wu bootstrap method is based on the assumption of sampling with a replacement, but is often employed for sampling without a replacement as well, when the first-stage sampling fraction is negligible (Mach et al., 2007). Herein, we propose implementing the coverage-corrected bootstrap method using the weight adjustment from Rao and Wu (1988). Among the various bootstrap confidence interval techniques (e.g., for these varieties, see Hwang, 1995), we consider the percentile bootstrap interval, which is a strictly nonparametric bootstrap approach (Hall, Martin and Schucany, 1989).

The coverage rates of the bootstrap confidence intervals can be corrected by incorporating additional bootstrap procedures. Because of bootstrapping the bootstrap sample, this kind of a procedure is referred to as the double bootstrap method (Martin, 1992). It is known that this method reduces the coverage error of two-sided confidence intervals by a factor of the order n^{-1} compared to the single bootstrap or normal-theory confidence intervals (Martin, 1992). Suppose $\hat{l}$ and $\hat{u}$ are the lower and upper bounds of the percentile bootstrap confidence interval using the original data. As proposed by Hall et al. (1989), the $100(1-q)\%$ coverage-corrected bootstrap confidence interval can be defined as $\left(\hat{l} - \delta,\ \hat{u} + \delta\right)$, where a positive value of δ satisfies

$$1 - q = \Pr\left\{\hat{\alpha} \in \left(\hat{l}^* - \delta,\ \hat{u}^* + \delta\right)\right\}. \tag{2.11}$$

The values $\hat{l}^*$ and $\hat{u}^*$ indicate the lower and upper bounds, respectively, of the confidence interval obtained by bootstrapping a resampled data set. The probability in the right-hand side of equation (2.11) is empirically evaluated as shown in the following steps.

Step 1: For each bootstrap sample $i\,(i = 1, \ldots, B)$, we obtain the intervals based on second-time resamples, $\left(\hat{l}_i^*, \hat{u}_i^*\right)$.

Step 2: We search t satisfying $\min\left\{t: \left|1 - q - \widehat{\Pr}\left\{\hat{\alpha} \in \left(\hat{l}^* - \delta, \hat{u}^* + \delta\right)\right\}\right| \geq 0\right\}$ where $\widehat{\Pr}$ indicates the empirical probability.

Step 3: The confidence interval is obtained by $\left(\max\left(\hat{l} - \delta, 0\right), \min\left(\hat{u} + \delta, 1\right)\right)$.

We use $\left(\max\left(\hat{l} - \delta, 0\right), \min\left(\hat{u} + \delta, 1\right)\right)$ since the true α is assumed to be between 0 and 1.

In the analysis, we have to resample the data without disrupting the survey design structure. The bootstrap is carried out within each stratum, and all observations in the same cluster should be kept together in a resampled data set (Lohr, 1999). For each resampled data set, new weights need to be obtained (Rao, Wu and Yue, 1992). Specifically, let n_h indicate the sample size of the primary sampling unit (PSU) in stratum $h\,(h = 1, \ldots, H)$. Suppose we resample n_h^* clusters for each stratum. Then, the rescaled weight for observation k in the resample is

$$w_k^{(b)} = w_k \left\{\left(1 - \sqrt{\frac{n_h^*}{n_h - 1}}\right) + \sqrt{\frac{n_h^*}{n_h - 1}}\,\frac{n_h}{n_h^*}\,m_k\right\}, \tag{2.12}$$

where m_k is the number of repetitions of the PSU that observation k belongs to and w_k is the original weight of observation k (Rao et al., 1992; Mach, Dumais and Robinson, 2005; Mach et al., 2007). When $n_h^* = n_h - 1$, the bootstrap weight becomes $w_k^{(b)} = w_k \left\{\frac{n_h}{n_h - 1}\,m_k\right\}$, which is a conventional bootstrap weight (Lohr, 1999). This procedure is repeated to obtain a total of B bootstrap samples. For the actual data analysis, we use $B = 500$ following the common practice of Statistics Canada surveys (Canadian Community Health Survey - Annual Component, 2007). To obtain the estimates, the stratification or cluster structure is no longer considered since the bootstrap weights take into account the survey design structure (Lohr, 1999). The percentile interval will be obtained based on the B values of the estimates of α. For each resample, α is estimated based on the sample variance and covariance matrix incorporating the weights. To obtain the coverage-corrected confidence interval, we carry out the additional bootstrap with each bootstrap sample in a similar manner to what was explained above. In the simulation and data analysis we use 200 bootstrap samples for the second round of bootstrapping. The relevant R code is provided in the Supplementary Material.

3 Simulation

We investigate the performance of the proposed methods in two scenarios; stratified two-stage cluster sampling and single-stage unequal probability sampling.

For stratified two-stage cluster sampling, the finite population is generated using three strata where each stratum includes 200 PSUs and 50 secondary sampling units (SSUs) totaling 30,000 SSUs. The underlying distributions that are used include the multi-normal distribution, multi-lognormal distribution and correlated ordinal data categorized from multi-lognormal distribution variables. The cases of $p = 5$ and $p = 10$ are considered. Different means are used for the different strata. The observations are correlated within a PSU. See the footnote of Table 3.1 for the detailed parameter information. Simple random sampling is carried out at the first-stage and second-stage, respectively, within each stratum. Thus, the appropriate weights are calculated per stratum as $(N_h M_h)/(n_h m_h)$ for each individual (SSU), where N_h, M_h, n_h and m_h are the number of PSUs per stratum, the number of SSUs per PSU, first-stage sample size per stratum, and second-stage sample size, respectively. Since the population is finite, the true value of α is known from the generated population.

For unequal probability sampling (Table 3.2), we generate a population of 30,000, where the underlying distributions of the data are the multi-normal distribution, multi-lognormal distribution and correlated ordinal data categorized from the multi-lognormal distribution variables similar to the cases found in Table 3.1. See the footnote of Table 3.2 for the detailed parameter information. Each individual i is assigned a random number x_i from the Binomial (20, 0.5) distribution, achieving the semblance of SSU sizes per PSUs. For sampling, the first-order inclusion probability is proportional to size x_i (probability proportional to size sampling). Thus, the weight for an individual i is obtained as $n^{-1}\sum_k x_k \big/ x_i$, where n is the sample size. The sample selection procedure uses the systematic sampling technique that considers first-order inclusion probabilities. For the linearization method, the variance is estimated using the usual estimator for with-replacement sampling (Mach et al., 2007) as a conservative approximation of the methods for without-replacement sampling (Wolter, 1985). Since the sampling fraction is negligible in the simulation, the finite population correction is not incorporated. The 95% confidence interval is obtained based on the normal approximation.

Table 3.1 (stratified two-stage cluster sampling) and Table 3.2 (single-stage unequal probability sampling) show the coverage rates and average widths of the confidence intervals based on the proposed linearization method and the coverage-corrected bootstrap method (1,000 simulations per scenario). The linearization method and the coverage-corrected methods are evaluated using same simulated data sets. For the coverage-corrected method, we use $B = 200$ for the first bootstrap, $B = 200$ for the second bootstrap. The linearized method shows the coverage rates as being close to the target confidence level for the multi-normal distributions and correlated ordinal data in most scenarios. We note that, in the random variable settings, the confidence intervals based on a normal approximation work well with various ordinal data once the variance is correctly obtained (Maydeu-Olivares, Coffman and Hartmann, 2007). Our simulation results show that the normal approximation works well with the ordinal data in finite population settings as well. When the underlying distribution is the multi-lognormal distribution, the coverage rates of the confidence intervals based on the normal approximation may be somewhat lower than the target coverage rate, but they improve with increasing sample sizes. For the multi-lognormal distribution, the coverage-corrected

bootstrap method using the weight adjustment by Rao and Wu (1988) shows substantially improved coverage rates comparing to the linearized method. In comparison to the linearization method, the coverage-corrected bootstrap method has slightly increased widths, and the coverage rates are reasonably close to the target confidence level for most cases in Tables 3.1 and 3.2.

We also note that for the stratified sampling cases with relatively low α values, we can identify cases that the coverage-corrected method provides less-than-desirable coverage rates, with the multi-normal or ordinal data indicating that the coverage-corrected method is not a panacea for interval estimation. Here, the linearization method is a reasonable choice over the coverage-corrected method if the underlying distribution is ordinal or normal.

Table 3.1
(Stratified two-stage cluster sampling). The coverage rates (CR) and average widths (Width) of 95% confidence intervals based on the linearization method and coverage-corrected method (Double Bt). The values of npsu and nssu are the sample sizes for PSUs and SSUs within a PSU, respectively. Two α values indicate α for $p = 5$ and $p = 10$, respectively

Method	Distribution	(npsu, nssu)	α	$p = 5$		$p = 10$	
				CR	Width	CR	Width
Linearization	Multi-normal	(10, 20)	0.91, 0.91	0.941	0.024	0.946	0.023
		(20, 20)	0.90, 0.91	0.934	0.017	0.962	0.016
		(10, 20)	0.56, 0.67	0.941	0.121	0.944	0.095
		(20, 20)	0.56, 0.67	0.953	0.084	0.961	0.064
	Multi-lognormal	(10, 20)	0.85, 0.85	0.904	0.067	0.902	0.059
		(20, 20)	0.86, 0.85	0.908	0.054	0.935	0.049
		(10, 20)	0.51, 0.53	0.913	0.163	0.924	0.154
		(20, 20)	0.51, 0.55	0.933	0.118	0.928	0.108
	Correlated ordinal	(10, 20)	0.85, 0.87	0.939	0.043	0.938	0.035
		(20, 20)	0.85, 0.87	0.939	0.030	0.954	0.025
		(10, 20)	0.48, 0.53	0.934	0.147	0.928	0.130
		(20, 20)	0.48, 0.60	0.955	0.103	0.955	0.077
Double Bootstrap	Multi-normal	(10, 20)	0.91, 0.91	0.959	0.026	0.960	0.025
		(20, 20)	0.90, 0.91	0.954	0.019	0.964	0.017
		(10, 20)	0.56, 0.67	0.939	0.120	0.909	0.084
		(20, 20)	0.56, 0.67	0.955	0.084	0.942	0.059
	Multi-lognormal	(10, 20)	0.85, 0.85	0.945	0.080	0.959	0.071
		(20, 20)	0.86, 0.85	0.948	0.063	0.963	0.057
		(10, 20)	0.51, 0.53	0.947	0.186	0.942	0.163
		(20, 20)	0.51, 0.55	0.955	0.125	0.942	0.109
	Correlated ordinal	(10, 20)	0.85, 0.87	0.964	0.047	0.955	0.038
		(20, 20)	0.85, 0.87	0.950	0.033	0.960	0.026
		(10, 20)	0.48, 0.53	0.937	0.148	0.919	0.121
		(20, 20)	0.48, 0.60	0.957	0.104	0.942	0.073

The values of α are based on the generated finite populations in all scenarios. For multi-normal data, the mean vectors consist of values of 1, 1.05 and 1.1 for strata 1, 2, and 3, respectively, and the common covariance within PSUs in addition to the covariance within multivariate data is 0.05. The covariance matrix has diagonal elements of 1 and the common off-diagonal elements to produce relevant α values. Multi-lognormal data are exponential of multi-normal data with the same mean and covariate structures. In the covariance matrix, common off-diagonal values are selected to produce relevant α values. For the correlated ordinal data, we first generate the multi-lognormal data with the same structures described above, then categorize them to 0, 1, 2 and 3 for values ≤ 2, $2 <$ values ≤ 10, $10 <$ values ≤ 15, and values > 10, respectively.

Table 3.2
(Single-stage unequal probability sampling). The coverage rates (CR) and average widths (width) of 95% confidence intervals based on the linearization method and coverage-corrected method (Double Bt). The values of n indicate the sample sizes for PSUs. Two α values indicate α for $p = 5$ and $p = 10$, respectively

Method	Distribution	n	α	$p = 5$		$p = 10$	
				CR	Width	CR	Width
Linearization	Multi-normal	100	0.90, 0.90	0.942	0.063	0.936	0.058
		200	0.90, 0.90	0.921	0.044	0.956	0.042
		100	0.50, 0.51	0.942	0.317	0.936	0.291
		200	0.50, 0.50	0.921	0.219	0.956	0.210
	Multi-lognormal	100	0.85, 0.84	0.816	0.116	0.853	0.104
		200	0.85, 0.85	0.870	0.103	0.901	0.083
		100	0.47, 0.47	0.851	0.346	0.887	0.312
		200	0.48, 0.47	0.911	0.264	0.935	0.253
	Correlated ordinal	100	0.84, 0.86	0.926	0.110	0.923	0.086
		200	0.84, 0.86	0.930	0.078	0.947	0.063
		100	0.43, 0.43	0.938	0.368	0.945	0.335
		200	0.43, 0.42	0.942	0.260	0.948	0.245
Double Bt	Multi-normal	100	0.90, 0.90	0.961	0.073	0.950	0.068
		200	0.90, 0.90	0.953	0.049	0.965	0.047
		100	0.50, 0.51	0.958	0.361	0.951	0.241
		200	0.50, 0.50	0.948	0.335	0.962	0.232
	Multi-lognormal	100	0.85, 0.84	0.912	0.166	0.943	0.138
		200	0.85, 0.85	0.940	0.136	0.948	0.107
		100	0.47, 0.47	0.954	0.436	0.946	0.382
		200	0.48, 0.47	0.946	0.318	0.965	0.295
	Correlated ordinal	100	0.84, 0.86	0.940	0.134	0.937	0.103
		200	0.84, 0.86	0.937	0.090	0.956	0.066
		100	0.43, 0.43	0.954	0.428	0.946	0.388
		200	0.43, 0.42	0.949	0.287	0.957	0.271

The values of α are based on the generated finite populations in all scenarios. For multi-normal data, the mean vectors consist of values of 1. The covariance matrix has diagonal elements of 1 and the common off-diagonal elements to produce relevant α values. Multi-lognormal data are exponential of multi-normal data with the same mean and covariate structures. Common off-diagonal values are selected to produce relevant α values. For the correlated ordinal data, we first generated the multi-lognormal data with the same structures described above, then categorize them to 0, 1, 2 and 3 for values ≤ 2, $2 <$ values ≤ 10, $10 <$ values ≤ 15, and values > 10, respectively.

Thus, we conclude that, for general ordinal data, which are typical responses for most assessment instruments, the linearization method will be satisfactory to obtain the confidence intervals. When the instruments consist of continuous data and some skewed distributions are observed, the coverage-corrected bootstrap method will generally provide more accurate confidence intervals than the normal approximation.

It may be of interest to compare the performance of the proposed confidence interval methods to other existing confidence interval methods in a random variable setting since the proposed methods can be applied to these settings, as shown in (2.10). Table 3.3 presents the comparisons of the coverage rates and widths of various confidence interval methods based on the data generated from a random variable. The existing confidence interval methods can be categorized to either using an analytical distribution based on the multi-normal distribution, or using a large sample approximation for the normal distribution of $\hat{\alpha}$ or a transformation of $\hat{\alpha}$. For the existing methods, we consider three normal-based confidence intervals and a bootstrap method, i.e., confidence intervals based on the exact F distribution using the normal data (van Zyl,

Neudecker and Nel, 2000; Kistner and Muller, 2004), a large sample approximation of $\log(1-\hat{\alpha})/2$ (van Zyl et al., 2000), a large sample approximation of $\hat{\alpha}$ based on the "distribution-free" standard error estimate (Yuan et al., 2003; Maydeu-Olivares et al., 2007), and the percentile bootstrap confidence interval with a single bootstrap (DiCiccio and Romano, 1988). These techniques are compared to the confidence intervals based on the linearization method and the coverage-corrected bootstrap method. The data are generated from the multi-normal distribution, multi-lognormal distribution, and the correlated ordinal data similar to the simulations in the previous tables. The values of α in Table 3.3 are for the random variables. In general, the results seem similar to those of finite population cases. The existing confidence interval methods, as well as the linearization method, perform unsatisfactorily with the lognormal data, yet their coverage rates are close to the target confidence levels using the ordinal data and normal distributions when the sample sizes increase. The coverage-corrected bootstrap method shows a coverage rate close to the confidence level with a lognormal distribution while providing wider confidence interval widths than the other methods. In the case of the multi-normal distribution, the coverage-corrected bootstrap method seems to have higher coverage rates than the target confidence level. In comparison with the single bootstrap method, the coverage-corrected method increases the coverage rates by 1 to 3% overall for the multi-lognormal distribution cases.

Table 3.3
The coverage rates and widths of 95% confidence intervals based on F distribution (F dist), the asymptotic distribution of the transformed $\hat{\alpha}$ (Asymp1), the asymptotic distribution by Yuan et al. (Asymp2), the linearization method (Linearization), the percentile bootstrap method with single bootstrap (Single Bt) and the coverage corrected method (Double Bt). In the first column, p, low α, and high α values are shown in the parentheses

Distribution	Approach	n	$p = 5$				$p = 10$			
			Low α		High α		Low α		High α	
			CR	Width	CR	Width	CR	Width	CR	Width
Multi-normal (5, 0.5, 0.9) (10, 0.5, 0.9)	F dist	50	0.955	0.461	0.955	0.092	0.960	0.429	0.960	0.086
		100	0.954	0.319	0.954	0.064	0.943	0.298	0.943	0.060
		200	0.948	0.222	0.948	0.044	0.954	0.208	0.042	0.954
	Asymp1	50	0.954	0.471	0.954	0.094	0.956	0.440	0.956	0.088
		100	0.947	0.322	0.947	0.064	0.939	0.302	0.939	0.060
		200	0.947	0.223	0.947	0.045	0.959	0.209	0.959	0.042
	Asymp2	50	0.937	0.432	0.937	0.086	0.931	0.407	0.931	0.081
		100	0.948	0.311	0.948	0.062	0.943	0.293	0.943	0.059
		200	0.945	0.218	0.945	0.044	0.953	0.205	0.953	0.041
	Linearization	50	0.937	0.441	0.937	0.088	0.937	0.415	0.937	0.083
		100	0.948	0.315	0.948	0.062	0.944	0.296	0.944	0.059
		200	0.946	0.219	0.946	0.044	0.953	0.206	0.953	0.041
	Single Bt	50	0.936	0.490	0.936	0.098	0.935	0.465	0.935	0.093
		100	0.944	0.334	0.944	0.067	0.939	0.314	0.939	0.063
		200	0.944	0.227	0.944	0.045	0.944	0.227	0.965	0.043
	Double Bt	50	0.959	0.498	0.960	0.107	0.959	0.484	0.960	0.103
		100	0.958	0.355	0.960	0.072	0.954	0.336	0.954	0.068
		200	0.954	0.238	0.954	0.048	0.954	0.238	0.974	0.045

The values of α are theoretical values except cases of correlated ordinal data. The α values for the correlated ordinal data are obtained based on 60,000 simulations. Structures of the mean vector and covariance matrix follow those explained in Table 3.2.

Table 3.3 (continued)
The coverage rates and widths of 95% confidence intervals based on F distribution (F dist), the asymptotic distribution of the transformed $\hat{\alpha}$ (Asymp1), the asymptotic distribution by Yuan et al. (Asymp2), the linearization method (Linearization), the percentile bootstrap method with single bootstrap (Single Bt) and the coverage corrected method (Double Bt). In the first column, p, low α, and high α values are shown in the parentheses

Distribution	Approach	n	$p = 5$				$p = 10$			
			Low α		High α		Low α		High α	
			CR	Width	CR	Width	CR	Width	CR	Width
Multi-lognormal	F dist	50	0.919	0.487	0.829	0.151	0.928	0.457	0.860	0.140
(5, 0.47, 0.85)		100	0.888	0.337	0.763	0.101	0.884	0.317	0.813	0.095
(10, 0.47, 0.84)		200	0.862	0.235	0.727	0.070	0.906	0.221	0.782	0.066
	Asymp1	50	0.921	0.497	0.827	0.155	0.923	0.469	0.859	0.143
		100	0.884	0.341	0.759	0.103	0.888	0.321	0.809	0.097
		200	0.858	0.237	0.720	0.070	0.909	0.223	0.787	0.066
	Asymp2	50	0.837	0.410	0.805	0.146	0.870	0.406	0.844	0.132
		100	0.874	0.338	0.825	0.119	0.883	0.318	0.854	0.108
		200	0.903	0.267	0.853	0.097	0.927	0.244	0.876	0.086
	Linearization	50	0.842	0.419	0.814	0.149	0.878	0.415	0.850	0.135
		100	0.877	0.342	0.828	0.120	0.885	0.321	0.862	0.109
		200	0.903	0.269	0.858	0.098	0.928	0.245	0.879	0.086
	Single Bt	50	0.929	0.472	0.887	0.174	0.930	0.464	0.889	0.158
		100	0.928	0.362	0.883	0.133	0.929	0.337	0.887	0.119
		200	0.932	0.274	0.900	0.102	0.941	0.251	0.917	0.090
	Double Bt	50	0.943	0.524	0.944	0.221	0.950	0.504	0.930	0.199
		100	0.950	0.422	0.935	0.170	0.951	0.385	0.938	0.150
		200	0.955	0.318	0.943	0.126	0.954	0.283	0.948	0.109
Correlated ordinal	F dist	50	0.941	0.424	0.926	0.149	0.950	0.256	0.931	0.075
(5, 0.84, 0.54)		100	0.931	0.292	0.929	0.102	0.939	0.177	0.904	0.052
(10, 0.91, 0.70)		200	0.938	0.203	0.917	0.071	0.956	0.123	0.933	0.036
	Asymp1	50	0.945	0.432	0.919	0.152	0.947	0.262	0.927	0.077
		100	0.930	0.295	0.922	0.103	0.938	0.179	0.907	0.053
		200	0.934	0.204	0.914	0.071	0.954	0.124	0.936	0.036
	Asymp2	50	0.922	0.432	0.911	0.144	0.928	0.242	0.920	0.074
		100	0.928	0.289	0.933	0.108	0.931	0.177	0.918	0.055
		200	0.940	0.205	0.931	0.077	0.950	0.125	0.947	0.039
	Linearization	50	0.928	0.402	0.916	0.147	0.932	0.247	0.923	0.075
		100	0.929	0.292	0.936	0.109	0.936	0.178	0.925	0.056
		200	0.940	0.206	0.931	0.078	0.950	0.126	0.950	0.039
	Single Bt	50	0.927	0.447	0.901	0.163	0.921	0.275	0.898	0.084
		100	0.928	0.308	0.929	0.116	0.935	0.189	0.908	0.059
		200	0.938	0.213	0.935	0.080	0.949	0.131	0.942	0.041
	Double Bt	50	0.950	0.476	0.927	0.189	0.945	0.310	0.934	0.101
		100	0.943	0.334	0.945	0.131	0.951	0.208	0.937	0.069
		200	0.955	0.227	0.948	0.087	0.956	0.140	0.959	0.045

The values of α are theoretical values except cases of correlated ordinal data. The α values for the correlated ordinal data are obtained based on 60,000 simulations. Structures of the mean vector and covariance matrix follow those explained in Table 3.2.

4 Application

In this section, we provide detailed information regarding the NCS-R survey and subgroup analysis using the data sets. The relevance of the instruments may vary based on the different demographic groups studied, and thus a relatively low reliability in a certain group would be an indication that the instrument items may

need some adjustments for that group. Using the data from the NCS-R, we investigate the changes of α using the Kessler 10 (K10, Kessler, Andrews, Colpe, Hiripi, Mroczek, Normand, Walters and Zaslavsky, 2002), the Kessler 6 (K6, Kessler et al., 2002) and the Sheehan Disability Scale (SDS, Sheehan, Harnett-Sheehan and Raj, 1996). More details about these scales are explained in Section 4.1.

4.1 The data

The NCS-R is a mental health survey for a nationally representative sample of English-speaking noninstitutionalized household residents in the United States (Kessler et al., 2004) and it uses the fully structured World Health Organization's (WHO) World Mental Health Survey version of the Composite International Diagnostic Interview (WMH-CIDI) (Byers, Yaffe, Covinsky, Friedman and Bruce, 2010). Using computer-assisted personal interviews, the NCS-R was carried out to obtain further information not fully covered in the previous baseline National Comorbidity Survey (NCS). A total of 9,282 participants 18 years and older completed the Part I interview, and a subsample of 5,692 participants completed the Part II instruments. The data sets are publicly accessible and downloadable on the ICPSR (Inter-university Consortium for Political and Social Research) website (https://www.icpsr.umich.edu/icpsrweb). The NCS-R is based on a stratified multi-stage probability sample design (42 strata where each stratum has two PSUs, totaling 84 PSUs), and the sample weights are provided in the data to reflect the survey design. Each PSU consists of metropolitan statistical areas or counties (Kessler et al., 2004). The final weights in the NCS-R data are adjusted for nonresponses to the survey instruments. Weights accounting for the designs of the different parts of the surveys (i.e., Parts I and II) are provided, respectively, in the NCS-R data. The weights are normalized to have a sum equal to 9,282 for Part I and 5,692 for Part II (mean weight = 1), respectively. In this case, the weights do not represent the inverse of the selection probabilities. Due to this and the fact that the sample size is quite small compared to the total population of interest, the finite population correction is not considered in the data analysis. Incorporating these weights corrects the overrepresentation of "racial minorities, females, residents of the Midwest, people with 13+ years of education, and residents of metropolitan areas" (Kessler et al., 2004).

The 10-item Kessler psychological distress scale or the K10 is an instrument used to assess the distress level of people (Kessler et al., 2002), and the K6 is an abbreviated set of six items from the K10. Both the K10 and K6 are considered effective scales for screening mental disorders (Brouwer, Cornelius, van der Klink and Groothoff, 2013). The K10 for 30-day symptoms is included in the Part II instruments. It is composed of 10 questions of a self-reported assessment of psychological distresses in the worst month of the past year for each interviewee. The questions ask feelings such as tiredness, nervousness, hopelessness, and so forth. All 10 questions produce an ordinal data scoring of 1 (all of the time) to 5 (none of the time). The final total score ranges from 10 to 50 with the higher scores showing more distress. The K10 values in the NCS-R have missing data, and the weights given by the NCS-R adjust for survey nonresponses, but they

do not adjust for items with missing data. Although these missing data may compromise the unbiasedness of the weighted estimation (Alegria, Jackson, Kessler and Takeuchi, 2007), we use only completed data and remedial approaches such as weighting class adjustment or imputation of the data are not considered in our analysis.

The SDS assesses functional impairment associated with mental disorders (Sheehan et al., 1996). The SDS in the NCS-R assesses disorder-specific role impairments (Sheehan et al., 1996; Druss, Hwang, Petukhova, Sampson, Wang and Kessler, 2009). It consists of four questions evaluating the disruption of activities associated with home, work, social and close relationship using 0 to 10 scales, with higher scores showing more severe impairment. In this paper, among the SDS scales of various mental disorders, we use the SDS for the participants with chronic conditions as a Part II instrument. Since the SDS is disorder-specific, it has missing data. For the data analysis, we use only complete data.

4.2 Subsample analysis

For the subgroups, a domain analysis may be applied. Suppose that a domain indicator function I_k^d $(d = 1, \ldots, D)$ has a value of 1 if the unit k is in a domain d (i.e., $k \in s_d$) and 0 otherwise. Then, the statistics of the domain are estimated by modifying the weight as $w_k^{(d)} = w_k I_k^d$. The procedures used to obtain the estimates and the corresponding variance or covariance are carried out with the modified weights. Since the sample size is not fixed but is rather treated as an estimate, an estimator such as the sample mean and sample variance can be considered as the ratio estimator, i.e., both the numerator and the denominator are estimated, and the variance of the estimator is obtained accordingly. However, when the sample size is large, thus the ratio between the domain sample size and the whole sample size is close to the true population ratio, it is known that the variance of the ratio estimator is approximately the same as that of the estimator with the fixed sample size using only the subgroup of interest, making "little difference in practice" regarding those estimators (Lohr, 1999, page 79). The negligible difference between the domain estimator and the estimator using only the subsample can be easily shown using the variance estimator in an unequal probability sampling with replacement setting. Let $\hat{\bar{Y}}_d$ indicate the domain estimator of the mean (Lohr, 1999) for single-stage sampling, i.e., $\hat{\bar{Y}}_d = \sum_{k=1}^n w_k I_k^d y_k \Big/ \sum_{k=1}^n w_k I_k^d = \sum_{k \in s_d} w_k y_k \Big/ \sum_{k \in s_d} w_k$, where the last term uses only the subsample. Now, for the variance estimator of $\hat{\bar{Y}}_d$ (Paben, 1999; SAS/STAT user's guide, 2010), we can show

$$\hat{V}\left(\hat{\bar{Y}}_d\right) = \sum_{k=1}^n \left\{ \frac{w_k I_d \left(y_k - \hat{\bar{Y}}_d\right)}{\sum_{l=1}^n w_l I_d} \right\}^2 \frac{n}{n-1} \simeq \sum_{k \in s_d} \left\{ \frac{w_k \left(y_k - \hat{\bar{Y}}_d\right)}{\hat{N}_d} \right\}^2 \frac{n_d}{n_d - 1}, \tag{4.1}$$

where n_d is the sample size of s_d. Here, the right-hand side of equation (4.1) uses the observation only in domain s_d. Based on this fact, the variance for a subgroup is obtained based only on the data from the subgroup of interest in this paper.

When implementing the bootstrap method, we use $n_h^* = 2,$ which produces all the positive weights in (2.12). In the subsample analysis, the bootstrap sample may contain only one PSU per stratum. In this case, the variance cannot be estimated. If we have multiple strata with one PSU, we combine those strata. If we have only one stratum with one PSU, we merge that stratum with another stratum arbitrarily. The rationale of this practice is that the variance incorporating strata is usually smaller than that without strata, thus such a practice may produce a wider (more conservative) confidence interval.

4.3 Results

The estimates of α and their confidence intervals for the whole participants are shown in Table 4.1. The table presents the confidence intervals using the coverage-corrected percentile method and the confidence interval using the linearization method for each instrument. Between the K10 and K6, it appears that the K10 has a higher α estimate. This may be explained by the fact that the removed items from the K10 are highly correlated with the remaining items in the K6, thus removing these items results in a reduced $\hat{\alpha}$ value. The coverage-corrected percentile method shows confidence intervals that are close to the linearization method, while slightly wider. Considering the ease of calculation, when an analysis deals with instruments with ordinal data, the results of the similar confidence intervals in Table 4.1 may indicate that a normal approximation using the proper variance estimation may be satisfactory for the investigated instruments, which do not include the skewed continuous data that we examined in Tables 3.1 and 3.2.

The subgroup analysis is shown in Table 4.2, where $\hat{\alpha}$ and the confidence intervals are presented for different groups by age, gender and marriage status. The age groups are defined as young (34 years and under), middle aged (35-64 years), and old aged (65 years and over) per the available literature (e.g., Sunderland, Hobbs, Anderson and Andrews, 2012), where the cut-off points for the age groups are decided by epidemiological studies and the traditional definition of old age. The marriage status is defined by grouping married and unmarried (including divorced, separated, widowed and never married). Both the coverage-corrected bootstrap method and the linearization method provide comparable confidence intervals while the coverage-corrected bootstrap produces a slightly wider confidence interval. Considering that the coverage-corrected method is computationally intensive, the linearization method may be preferred when the instruments consist of ordinal scales.

Table 4.1
Estimates of α and their 95% confidence intervals (CI) for overall sample

Instrument	$\hat{\alpha}$	Cov-Correct CI	Linearization CI	*n*
K10	0.901	(0.893, 0.911)	(0.893, 0.909)	2,378
K6	0.840	(0.829, 0.857)	(0.827, 0.852)	3,442
SDS	0.867	(0.852, 0.883)	(0.853, 0.880)	3,983

Table 4.2
Estimates of α and their 95% confidence intervals (CI) for subgroups

Instrument	Subgroups	$\hat{\alpha}$	Cov-Correct CI	Linearization CI	n
K10	Female	0.898	(0.880, 0.914)	(0.882, 0.914)	869
	Male	0.902	(0.896, 0.912)	(0.895, 0.910)	1,509
	Young age	0.888	(0.875, 0.900)	(0.875, 0.900)	890
	Middle age	0.913	(0.902, 0.925)	(0.902, 0.924)	1,281
	Old age	0.862	(0.827, 0.894)	(0.830, 0.893)	207
	Married	0.895	(0.882, 0.910)	(0.882, 0.907)	1,232
	Unmarried	0.902	(0.892, 0.913)	(0.892, 0.912)	1,146
K6	Female	0.824	(0.805, 0.849)	(0.803, 0.844)	1,288
	Male	0.848	(0.835, 0.866)	(0.835, 0.861)	2,154
	Young age	0.830	(0.810, 0.855)	(0.810, 0.849)	1,268
	Middle age	0.856	(0.842, 0.875)	(0.841, 0.870)	1,847
	Old age	0.773	(0.728, 0.821)	(0.725, 0.820)	327
	Married	0.823	(0.807, 0.844)	(0.806, 0.840)	1,805
	Unmarried	0.851	(0.833, 0.875)	(0.832, 0.869)	1,637
SDS	Female	0.874	(0.854, 0.895)	(0.853, 0.896)	1,589
	Male	0.861	(0.844, 0.880)	(0.847, 0.876)	2,394
	Young age	0.837	(0.805, 0.866)	(0.808, 0.866)	1,159
	Middle age	0.883	(0.870, 0.898)	(0.871, 0.896)	2,296
	Old age	0.849	(0.779, 0.903)	(0.796, 0.901)	555
	Married	0.886	(0.870, 0.903)	(0.871, 0.900)	2,286
	Unmarried	0.841	(0.818, 0.864)	(0.820, 0.861)	1,697

To this end, we conclude this section with a discussion of the results of the subgroups. Sizable differences in $\hat{\alpha}$ between the groups are found in the age groups with the K10 and K6 and marital status in the SDS. There are no overlaps of the confidence intervals between the middle and old-age groups in the K10 and K6. This indicates that the questions in the K10 and K6 may be relatively less consistent among the old-age group than the middle-age group. For the SDS, there is also no overlap of the confidence intervals between the married and the unmarried groups. That is, the consistency of the questions is substantially lower for the unmarried group than for the married group. We speculate that the SDS items include the impairment of a certain area that may be more relevant to the married group than the unmarried group (e.g., a disruption of activities associated with home, work, social and close relationship).

5 Concluding remarks

We explained how to obtain the confidence intervals of α in survey sampling through the linearization and coverage-corrected bootstrap methods. Through the simulation study in the setting of multi-stage cluster sampling and unequal probability sampling, the linearization method showed the workable property in terms of the coverage rate in the case of the multi-normal distribution or correlated ordinal data. When dealing with some problematic continuous data such as the multi-lognormal distribution, the coverage-corrected bootstrap method showed better performance than the linearization method in terms of the coverage rates. The discussed interval estimation methods were applied to the NCS-R data set. The application

demonstrated that both the interval estimation methods provide workable options to carry out an inference of α incorporating the survey design.

We conclude this section by noting the following recommendations. First, in the case of an unknown continuous and skewed distribution, the coverage-corrected confidence interval is a safe way to provide a confidence interval whose actual confidence level may be close to the nominal confidence level. Second, if the data are discrete with a large sample size, the normal approximation using the linearization method may provide satisfactory coverage rates and be preferred because of the easiness of computation.

Acknowledgements

The authors are grateful to the Associate Editor and two reviewers for comments and suggestions that led to a substantial improvement in this paper.

References

Alegria, M., Jackson, J.S., Kessler, R.C. and Takeuchi, D. (2007). *Collaborative Psychiatric Epidemiology Surveys (CPES)*, 2001-2003 [United States]. Retrieved from: http://doi.org/10.3886/ICPSR20240.v8.

Binder, D.A., and Roberts, G. (2009). Design-and model-based inference for model parameters. In *Handbook of Statistics*, Elsevier, 29, 33-54.

Bonett, D.G., and Wright, T.A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, 36, 3-15.

Brouwer, S., Cornelius, B.L.R., van der Klink, J.J.L. and Groothoff, J.W. (2013). The performance of the K10, K6 and GHQ-12 to screen for present state DSM-IV disorders among disability claimants. *BMC Public Health*, 13(1), 128.

Byers, A.L., Yaffe, K., Covinsky, K.E., Friedman, M.B. and Bruce, M.L. (2010). High occurrence of mood and anxiety disorders among older adults: The national comorbidity survey replication. *Archives of General Psychiatry*, 67(5), 489-496.

Canadian Community Health Survey – Annual Component (CCHS) (2007). Retrieved from: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=29539.

Cochran, W. (1977). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Cronbach, L.J. (1951). Coefficient alpha and the interval structure of tests. *Psychometrika*, 16, 297-334.

Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. New York: Cambridge University Press.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 1, 17-26. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004001/article/6991-eng.pdf.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey methodology*, 25, 2, 193-203. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999002/article/4882-eng.pdf.

DiCiccio, T.J., and Romano, J.P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society, Series B (Methodological)*, 338-354.

Druss, B.G., Hwang, I., Petukhova, M., Sampson, N.A., Wang, P.S. and Kessler, R.C. (2009). Impairment in role functioning in mental and chronic medical disorders in the United States: Results from the National Comorbidity Survey Replication. *Molecular Psychiatry*, 14(7), 728-737.

Fernholz, L.T. (1991). Almost sure convergence of smoothed empirical distribution functions. *Scandinavian Journal of Statistics*, 18(3), 255-262.

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Dekker.

Hall, P., Martin, M.A. and Schucany, W.R. (1989). Better nonparametric bootstrap confidence intervals for the correlation coefficient. *Journal of Statistical Computation and Simulation*, 33(3), 161-172.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons, Inc., 114.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.

Hwang, J.T. (1995). Fieller's problems and resampling techniques. *Statistica Sinica*, 5(1), 161-171.

Kessler, R.C., Andrews, G., Colpe, L.J., Hiripi, E., Mroczek, D.K., Normand, S.L.T., Walters, E.E. and Zaslavsky, A.M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, 32,(6) 959-976. http://dx.doi.org/10.1017/S0033291702006074.

Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B.E., Walters, E.E., Zaslavsky, A. and Zheng, H. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field, procedures. *International Journal of Methods in Psychiatric Research*, 13(2), 69-92.

Kistner, E.O., and Muller, K.E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika*, 69(3), 459-474.

Krätschmer, V., Schied, A. and Zähle, H. (2012). Qualitative and infinitesimal robustness of tail-dependent statistical functionals. *Journal of Multivariate Analysis*, 103(1), 35-47.

Lohr, S.L. (1999). *Sampling: Design and Analysis, 1st Edition*. Pacific Grove, CA: Duxbury Press.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), 1-19.

Mach, L., Dumais, J. and Robinson, A.A. (2005). A study of the properties of a bootstrap variance estimator under sampling without replacement. In *Arlington (Va): Federal Committee on Statistical Methodology (FCSM) Research Conference*.

Mach, L., Saïdi, A. and Pettapiece, R. (2007). Study of the properties of the Rao-Wu bootstrap variance estimator: What happens when assumptions do not hold. In *Proceedings of the Survey Methods Section, SSC Annual Meeting*.

Martin, M.A. (1992). On the double bootstrap. In *Computing Science and Statistics, Statistics of Many Parameters: Curves, Images, Spatial Models*, (Eds., C. Page and R. LePage), 78-78. New York: Springer.

Maydeu-Olivares, A., Coffman, D.L. and Hartmann, W.M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, 12(2), 157-176.

Paben, S. (1999). Comparison of variance estimation methods for the National Compensation Survey. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Patel, P.A., and Bhatt, S. (2016). A model-based estimation of finite population variance under PPS sampling. *Imperial Journal of Interdisciplinary Research*, 2(4).

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 317-337.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401), 231-241.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 2, 209-217. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14486-eng.pdf.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Springer Series in Statistics. Model Assisted Survey Sampling.* New York: US.

SAS Institute (2010). *SAS/STAT User's Guide: Version 9.2.* Cary, NC.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Sen, P.K. (1995). The Hájek asymptotics for finite population sampling and their ramifications. *Kybernetika*, 31(3), 251-68.

Sheehan, D.V., Harnett-Sheehan, K. and Raj, B.A. (1996). The measurement of disability. *International Clinical Psychopharmacology*, 11(3), 89-95.

Sunderland, M., Hobbs, M.J., Anderson, T.M. and Andrews, G. (2012). Psychological distress across the lifespan: Examining age-related item bias in the Kessler 6 Psychological Distress Scale. *International Psychogeriatrics*, 24(2), 231-242.

Swain, A.K.P.C., and Mishra, G. (1994). Estimation of finite population variance under unequal probability sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, 374-388.

Van Zyl, J.M., Neudecker, H. and Nel, D.G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, 65(3), 271-280.

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

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 253-261.

Yuan, K.-H., Guarnaccia, C.A. and Hayslip, B. (2003). A study of the distribution of sample coefficient alpha with the Hopkins symptom checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement*, 63(1), 5-23.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 485-511
Statistics Canada, Catalogue No. 12-001-X

Cost optimal sampling for the integrated observation of different populations

Piero Demetrio Falorsi, Paolo Righi and Pierre Lavallée[1]

Abstract

Social or economic studies often need to have a global view of society. For example, in agricultural studies, the characteristics of farms can be linked to the social activities of individuals. Hence, studies of a given phenomenon should be done by considering variables of interest referring to different target populations that are related to each other. In order to get an insight into an underlying phenomenon, the observations must be carried out in an integrated way, in which the units of a given population have to be observed jointly with related units of the other population. In the agricultural example, this means that a sample of rural households should be selected that have some relationship with the farm sample to be used for the study.

There are several ways to select integrated samples. This paper studies the problem of defining an optimal sampling strategy for this situation: the solution proposed minimizes the sampling cost, ensuring a predefined estimation precision for the variables of interest (of either one or both populations) describing the phenomenon. Indirect sampling provides a natural framework for this setting since the units belonging to a population can become carriers of information on another population that is the object of a given survey.

The problem is studied for different contexts which characterize the information concerning the links available in the sampling design phase, ranging from situations in which the links among the different units are known in the design phase to a situation in which the available information on links is very poor. An empirical study of agricultural data for a developing country is presented. It shows how controlling the inclusion probabilities at the design phase using the available information (namely the links) is effective, can significantly reduce the errors of the estimates for the indirectly observed population. The need for good models for predicting the unknown variables or the links is also demonstrated.

Key Words: Integrated surveys; Sample allocation; Indirect sampling.

1 Introduction

The need to observe together different populations related to each other is often encountered in social or economic studies. For example, in agricultural studies, the characteristics and behavior of farms can be linked to phenomena not only related to the farms themselves, but also to the social activities of individuals. This requires the study of the population of rural households, in addition to the study of the population of farms, in some integrated way. That is, to get an insight into an underlying phenomenon, the observations must be carried out in an integrated way, implying that the units of a given population have to be observed jointly with the related units of the other population. In the agricultural example, this means that a sample of rural households should be selected that have some relationship with the farm sample to be used for the study.

The integrated observation of two populations implies that if we observe the variables of the unit j of the first population, U^A, we need to observe the variables of all the units in the second population, U^B, which are linked with the j^{th} unit of U^A. The links among the units of the two populations are regulated by formal rules, contingent dependencies or relationships created for these purposes. Continuing with the

1. Piero Demetrio Falorsi and Paolo Righi, Italian National Institute of Statistics. E-mail: falorsi@istat.it, parighi@istat.it; Pierre Lavallée, E-mail: plavall1962@gmail.com.

agricultural example, these studies often refer to different statistical populations such as farms, rural households and land parcels, the units of which are linked to each other. The people of a given household may be the workers of a specific farm and those workers represent the links between the household and the farm. Furthermore, a given farm comprises specific land parcels which represent the links between that farm and the population of land parcels. The integrated observation of such populations allows the measurement global phenomena of the agricultural sector. Consider a given farm: the education level of the farm holder and the farm size, which are variables related to the population of farms, can affect the productivity of the land (a variable related to the statistical population of land parcels) which belongs to the farm. This productivity may have an impact on the risk of malnutrition of the households (population of rural households) in which the workers of the farm live. Thus, the observation of such different units in an integrated way provide insights into the relationships which link the level of education, the land productivity and the risk of malnutrition. If only aggregates are examined, then the advantage of integrated sampling is that it allows sampling from population U^B without having a frame available for it.

Another concrete example where the methodology may be of use is for firm-establishment-employee studies. For instance, the wellness of the households of people employed in firms which have a well-defined policy of social responsibility may be different from that of other types of households and the success in their children's schooling can be higher. In this case, the integrated observation allows the study the behavior of different sub-classes of households defined by a variable observable in the population of firms.

Other examples can be found in socio-demographic studies. For instance, the phenomenon of children who spend time in two households can be studied with the integrated observations of the population U^A of households and the population U^B of children.

Generally speaking, integrated observation may be of use for studying phenomena that involve variables which are correlated but belong to different statistical populations. Integrated observation allows the study of the relationships among all the variables of interest for the given phenomena, even if they belong to different populations. The independent observation of such populations would not allow the observation of the set of all the related variables of interest and hence it would not be possible to study the relationships among all the variables describing the phenomenon.

Indirect sampling (Lavallée, 2002, 2007) provides a natural framework for the estimation of the parameters of two target populations that are related to each other. In the indirect sampling framework, the units belonging to a population that are selected for a given survey can enable the collection of information on another population, through the relationship between the units in the two populations. Furthermore, indirect sampling is suitable for producing statistics of populations for which there is no sampling frame. In such a context, the sampling procedure assumes that population U^A is related to the population of interest U^B, but only the sampling frame of U^A is available. Then, a sample is selected from U^A, and using the links between the two populations, a sample of units of U^B is observed.

This paper studies the problem of sampling design for integrated observation of different populations. For this, an indirect sampling design is implemented. In particular, the focus is on the determination of the

inclusion probabilities. Since the sum of these probabilities define the expected sample size, we roughly define the problem as a *sampling allocation problem*. In fact, the two problems (determination of the inclusion probabilities and sampling allocation) coincide in stratified sampling. The allocation problem for the usual (direct) sampling setting has been dealt with in several books and papers. When one target parameter is to be estimated for the overall population, the optimal allocation in stratified sampling can be performed (Cochran, 1977, and Särndal, Swensson and Wretman, 1992). In particular, the optimal sample allocation minimizes the variance of the estimated total, subject to a given budget or, reversing the problem, a sample allocation that minimizes costs can be performed, subject to a given sampling error constraint. In multivariate cases, where more than one characteristic of each sampled unit must be measured, the optimal allocation for individual characteristics are of little practical use, unless the characteristics under study are highly correlated. This is because an allocation that is optimal for one characteristic can be far from optimal for others. The multidimensionality of the problem also leads to a compromise allocation method (Khan, Mati and Ahsan, 2010), with a loss of precision compared to the individual optimal allocations. Several authors have discussed various criteria for obtaining a feasible compromise allocation: see, for example, Kokan and Khan (1967), Chromy (1987), Bethel (1989) and Choudhry, Rao and Hidiroglou (2012).

Falorsi and Righi (2015) provide a general framework for sample design in multivariate and multi-domain surveys. This paper offers a further generalization of this framework to the case of integrated observation of two populations. Different scenarios related to the level of knowledge of the links are examined: the first scenario assumes the links between the populations are known in the design phase; the second scenario assumes the links between U^A and U^B are estimated in the design phase; in the third scenario, no links between U^A and U^B are available, but auxiliary variables on U^A can provide useful information on U^B.

Section 2 introduces the background and symbols. Section 3 and Section 4 illustrate the basic optimization problem and how it is applied in the different scenarios. Empirical evidence is shown in Section 5.

2 Background

Let U^A and U^B denote two related target populations, where U^A is the population with the available sampling frame, and U^B the survey population for which a sampling frame may or may not be available. For the agricultural example, U^A is the population of farms and U^B the population of rural households. Let s^A be a sample selected from U^A without replacement and with fixed sample size m^A, where U^A contains M^A units. Let π_j^A represent the inclusion probability of the j^{th} unit in U^A with $\pi_j^A > 0$ and $\sum_{j \in U^A} \pi_j^A = m^A$ with $\boldsymbol{\pi}^A = \left(\pi_1^A, \ldots, \pi_j^A, \ldots, \pi_{M^A}^A\right)'$. We denote by $y_{j,v}$ the value of the v^{th} $(v = 1, \ldots, V)$ characteristic on unit j and their total by Y_v^A.

We estimate the total Y_v^A according to the Horvitz-Thompson (HT) estimator,

$$\hat{Y}_v^A = \sum_{j \in s^A} w_j^A \, y_{j,v}, \tag{2.1}$$

where $w_j^A = 1/\pi_j^A$.

Many practical sampling designs define planned domains that are sub-populations in which the sample sizes are fixed before selecting the sample. Denote by $U_h^A\,(h = 1, \ldots, H)$ the planned domain of size $M_h^A = \sum_{j \in U^A} d_{j(h)}$ where $d_{j(h)} = 1$ if $j \in U_h^A$ and $d_{j(h)} = 0$ otherwise. Let us suppose that the $d_{j(h)}$ values are known and available in the sampling frame for all population units. Fixed size sampling designs are those satisfying

$$\sum_{j \in s^A} \mathbf{d}_j = \mathbf{m}^A,$$

where $\mathbf{d}_j = \left(d_{j(1)}, \ldots, d_{j(h)}, \ldots, d_{j(H)}\right)'$ and $\mathbf{m}^A = (m_1^A, \ldots, m_h^A, \ldots, m_H^A)'$ is the vector of integer numbers defining the sample sizes fixed at the design stage, with $\sum_{j \in U^A} d_{j(h)} \pi_j^A = m_h^A$. In our setting, the planned domains can overlap; therefore, the unit j may have more than one value $d_{j(h)} = 1$ (for $h = 1, \ldots, H$). Several customary fixed size sampling designs may be considered as particular cases. A well-known example is the Stratified Simple Random Sampling WithOut Replacement (SSRSWOR) design where strata are the planned domains and each $\mathbf{d}_j$ vector has $H - 1$ elements equal to zero, and one element equal to 1, which implies that each unit j can belong to one and only one planned domain. Furthermore, in this design all the units in the stratum U_h^A have a uniform inclusion probability given by $\pi_j^A = \frac{m_h^A}{M_h^A}$ for $j \in U_h^A$. If each $\mathbf{d}_j$ vector has $H - 1$ elements equal to zero and one element equal to 1, and the π_j^A values can be different in the stratum, we have a stratified sampling design, without replacement with fixed sample sizes and varying probabilities in each stratum. On the basis of the Winkler's definition (2001), if $\sum_{h=1}^{H} d_{j(h)} > 1$, we have an Incomplete Multi-Way Stratified Sampling design.

We suppose that the $M^A \times H$ matrix $\mathbf{D} = \left(\mathbf{d}_1, \ldots, \mathbf{d}_j, \ldots, \mathbf{d}_{M^A}\right)'$ is non-singular. According to this general sampling design framework, Deville and Tillé (2005) proposed an approximated expression of the variance for $\hat{Y}_v^A$ based on the Poisson sampling theory given by

$$V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \cong \left[M^A/(M^A - H)\right] \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,v}^2, \tag{2.2}$$

where $\hat{Y}_v^A \,\middle|\, \mathbf{m}^A$ is the HT estimator based on a general fixed sample size design with $\mathbf{m}^A$ units related to the vector $\boldsymbol{\pi}^A$,

$$\eta_{j,v} = y_{j,v} - \pi_j^A \mathbf{d}_j' \boldsymbol{\beta}_v, \tag{2.3}$$

and

$$\boldsymbol{\beta}_v = \boldsymbol{\Delta}^{-1} \sum_{j \in U^A} \pi_j^A \left(\frac{1}{\pi_j^A} - 1\right) \mathbf{d}_j y_{j,v} \tag{2.4}$$

with

$$\mathbf{\Delta} = \sum_{j \in U^A} \mathbf{d}_j \mathbf{d}_j' \pi_j^A \left(1 - \pi_j^A\right). \tag{2.5}$$

The variance (2.2) resembles the variance expression of the HT estimator under a Poisson sampling design, but it uses the residuals $\eta_{j,v}$, instead of the original value $y_{j,v}$. In practice, when $H = 1$ this is the variance approximation of the Conditional Poisson Sampling design (CPS, as introduced in Deville and Tillé, 2005). CPS selects samples by means of a Poisson sampling design without replacement until a given sample size is obtained.

To clarify the degree of approximation of (2.2), consider the SSRSWOR design. According to expression (2.2), we have

$$V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) = \left[M^A/(M^A - H)\right] \sum_{h=1}^{H} \sigma_{v,h}^2 M_h^A \left(\frac{M_h^A}{m_h^A} - 1\right),$$

where $\sigma_{v,h}^2$ is the design variance of the $y_{j,v}$ values in stratum U_h^A (see Appendix 4 of Falorsi and Righi, 2015). The above approximation works well when the number of domains H remains small compared to the overall population size M^A.

Let M^B, N^B, U_i^B and M_i^B be the number of units in U^B, the number of clusters in U^B, the i^{th} cluster of U^B with $\bigcup_{i=1}^{N^B} U_i^B = U^B$ and the number of units in the i^{th} cluster U_i^B, respectively. We denote by $y_{ik,r}$ the value of the r^{th} $(r = 1, \ldots, R)$ characteristic for the k^{th} unit of the i^{th} cluster of U^B and the population total of all $y_{ik,r}$'s by

$$Y_r^B = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik,r}.$$

Let $l_{j,ik}$ be an indicator variable of link existence: $l_{j,ik} = 1$ indicates that there is a link between j^{th} unit in U^A and k^{th} unit in U_i^B, while $l_{j,ik} = 0$ indicates otherwise.

Suppose that we carry out an indirect sampling process: if the unit $j \in U^A$ is included in s^A, then all the clusters U_i^B, for which $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik} > 0$, are observed (i.e., $y_{ik,r}$) in the indirect sample of population U^B. Let n^B be the size of the sample of clusters in population U^B obtained after the indirect sampling process. We estimate Y_r^B according to the estimator based on the theory of the Generalized Weight Share Method (GWSM) of Lavallée (2002, 2007):

$$\hat{Y}_r^B = \sum_{i=1}^{n^B} y_{i,r} w_i^B, \tag{2.6}$$

where

$$y_{i,r} = \sum_{k=1}^{M_i^B} y_{ik,r}$$

and

$$w_i^B = \sum_{j \in s^A} w_j^A \tilde{L}_{j,i}^B$$

with

$$\tilde{L}_{j,i}^B = \frac{L_{j,i}^B}{L_i^B}$$

and

$$L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B .$$

Theorem in Section 3 of Lavallée (2002, 2007) states that (2.6) provides an unbiased estimator for Y_r^B provided all links $l_{j,ik}$ can be correctly identified and $L_i^B > 0$ for all $i \in U^B$. By defining

$$z_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \, y_{i,r} , \tag{2.7}$$

the estimator (2.6) can be expressed as a usual Horvitz-Thompson (HT) estimator on the z values referring to the U^A population,

$$\hat{Y}_r^B = \sum_{s^A} z_{j,r} w_j^A . \tag{2.8}$$

Therefore, the variance $V\left(\hat{Y}_r^B\right)$ of $\hat{Y}_r^B$ may be expressed as the variance of the HT estimator on the U^A population. The approximate variance of $\hat{Y}_r^B$ for fixed size sampling designs is given by

$$V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \left[M^A/(M^A - H)\right] \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,r}^2 , \tag{2.9}$$

where $\hat{Y}_r^B \,\middle|\, \mathbf{m}^A$ is the HT estimator based on a general fixed sample size design with $\mathbf{m}^A$ units and the related vector $\boldsymbol{\pi}^A$,

$$\eta_{j,r} = z_{j,r} - \pi_j^A \mathbf{d}_j' \boldsymbol{\beta}_r$$

with

$$\boldsymbol{\beta}_r = \boldsymbol{\Delta}^{-1} \sum_{j \in U^A} \pi_j^A \left(\frac{1}{\pi_j^A} - 1\right) \mathbf{d}_j \, z_{j,r} .$$

Remark 2.1. An interesting extension of the above framework, useful in case of integrated studies, is the case of a total derived from a cross tabulation of a variable of the population U^A with a variable of the population U^B. In order to illustrate this extension, let y_v be a variable of U^A with C modalities and let $y_{j,v(c)}$ denote a dichotomous variable where $y_{j,v(c)} = 1$ if the unit j is characterized by the modality

$c\,(c = 1, \ldots, C)$ of y_v, and $y_{j,v(c)} = 0$ otherwise. Furthermore, let y_r be a variable of U^B with G modalities and let $y_{ik,r(g)}$ denote a dichotomous variable where $y_{ik,r(g)} = 1$ if the unit k of the cluster i is characterized by the modality $g\,(g = 1, \ldots, G)$ of y_r, and $y_{ik,r(g)} = 0$ otherwise. The total number of units of U^B characterized by the modality $g\,(g = 1, \ldots, G)$ of y_r and linked with units of population U^A characterized by the modality c of the variable y_v, can be defined as

$$Y^B_{(c,g)} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} l_{j,ik}\, y_{j,v(c)}\, y_{ik,r(g)} = \sum_{i=1}^{N^B} y_{i,(c,g)},$$

where $y_{i,(c,g)} = \sum_{j=1}^{N^A} \sum_{k=1}^{M_i^B} l_{j,ik}\, y_{j,v(c)} y_{ik,r(g)}$.

As an example, let us consider the case, illustrated in the introduction, of an integrated analysis examining the productivity of farms and the malnutrition of households, and suppose that $Y^B_{(c,g)}$ represents the total of persons with a malnutrition problem in the households of workers of farms characterized by an high productivity. In this case $y_{j,v(c)}$ has value 1 if the productivity of the farm j is high and $y_{ik,r(g)}$ has value 1 if the person k of the household i has a problem of malnutrition.

The GWSM estimator of $Y^B_{(c,g)}$ can be obtained directly from expression (2.8) using the transformed variable $z_{j,r} = \sum_{i=1}^{N^B} \tilde{L}^B_{j,i}\, y_{i,(c,g)}$.

3 Problem

Given the above framework, we are interested in finding the vector $\boldsymbol{\pi}^A = \left(\pi_1^A, \ldots, \pi_j^A, \ldots, \pi_{M^A}^A\right)'$ of inclusion probabilities that minimizes the expected survey cost bounding the sampling variances, $V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right)\,(v = 1, \ldots, V)$ and $V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)\,(r = 1, \ldots, R)$ under given variance constraints:

$$\begin{cases} \min \sum_{j \in U^A} c_j\, \pi_j^A \\ V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \le V_v^* \quad \forall v = 1, \ldots, V \\ V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \le V_r^* \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \le 1 \quad \forall j = 1, \ldots, M^A \end{cases} \tag{3.1}$$

where $\mathbf{m}^A$ is given by the $\boldsymbol{\pi}^A$ vector minimizing the cost function, $V_v^*\,(v = 1, \ldots, V)$ and V_r^* $(r = 1, \ldots, R)$ are the variance thresholds fixed by the sampling designer and c_j is the variable cost for observing the unit j in the population U^A and the $L_j^A = \sum_{i=1}^{N^B} L_{j,i}^B$ linked units in the population U^B. In other words, we want to obtain the optimal selection probabilities that will minimize the variance of estimates obtained for both U^A and U^B. For the agricultural example, this would translate to developing optimal selection probabilities that will lead to estimates for the population of farms, as well as the population of rural households, with specified precision.

A reasonable expression of c_j is

$$c_j = f_c\left(C^A, L_j^A; C^B\right), \tag{3.2}$$

where f_c is a known monotone non-decreasing function, C^A is the per unit cost for observing a unit in the population U^A and C^B is the cost for observing the elementary unit in the population U^B. Brewer and Gregoire (2009) propose an extensive analysis of different forms of costs functions.

The minimization problem (3.1) is a generalization of the univariate precision constrained optimization approach (Cochran, 1977). The problem (3.1) assumes that all the values $y_{j,v}$, $y_{i,r}$, L_j^A, $L_{j,i}^B$, L_i^B, $\boldsymbol{\beta}_v$ and $\boldsymbol{\beta}_r$ are known. In this case, problem (3.1) becomes a classical Linear Convex Separate Problem (LCSP) (Boyd and Vanderberg, 2004) and it can be solved by the algorithm proposed Chromy (1987), originally developed for multivariate optimal allocation in an SSRSWOR design and implemented in standard software tools. (See for example the Mauss-R software available at: http://www3.istat.it/strumenti/metodi/software/campione/mauss_r/.) Alternatively, the LCSP can be dealt with by the SAS procedure NLP as suggested by Choudhry et al. (2012). The vectors $\boldsymbol{\beta}_v$ and $\boldsymbol{\beta}_r$ depend on the vector $\boldsymbol{\pi}^A$. Falorsi and Righi (2015) define a new algorithm which finds the optimal solution taking into account the dependence between $\boldsymbol{\beta}_v$ and $\boldsymbol{\beta}_r$ with the optimal vector $\boldsymbol{\pi}^A$.

4 Informative contexts and optimization problem

Optimization problems as presented in (3.1) are quite theoretical since one needs to know the values of the variables of interest in both populations U^A and U^B, and the values of actual links among the units of the two populations. We now present three more concrete contexts involving various amount of information. We start from two contexts in which the information is very rich, whereas the third context considers a case in which the information is very poor. The latter context is the most common, although the growing availability of administrative registers and statistical software tools for data integration increases the plausibility of the first two contexts.

Context 1. The sampling frames for U^A and U^B are available. All the values L_j^A, $L_{j,i}^B$ and L_i^B are known and the values of $y_{j,v}$, $y_{i,r}$ are unknown but can be predicted by suitable superpopulation models.

This context may be realistic in countries, such as the Nordic ones, having well established register-based systems (Wallgren and Wallgren, 2014) in which the units of a given statistical register have unique identifiers of good quality, which allows identification of the same unit in the whole systems of registers. For the agricultural example, this means that one can link each farm to one or more rural households, and each rural household to one or more farms.

The *working* models that we study can be expressed under the following forms:

Unit level **Cluster level**

$$\begin{cases} y_{j,v} = \tilde{y}_{j,v} + u_{j,v} = f_v\left(\mathbf{x}_j; \boldsymbol{\varphi}_v\right) + u_{j,v} \\ E_{M_v}\left(u_{j,v}\right) = 0, \;\; E_{M_v}\left(u_{j,v}^2\right) = \sigma_{j,v}^2, \;\; \forall j \\ E_{M_v}\left(u_{j,v}, u_{l,v}\right) = 0, \;\; \forall j \neq l \end{cases} \quad , \quad \begin{cases} y_{i,r} = \tilde{y}_{i,r} + u_{i,r} = f_r\left(\mathbf{x}_i; \boldsymbol{\varphi}_r\right) + u_{i,r} \\ E_{M_r}\left(u_{i,r}\right) = 0, \;\; E_{M_r}\left(u_{i,r}^2\right) = \sigma_{i,r}^2, \;\; \forall i \\ E_{M_r}\left(u_{i,r}, u_{i',r}\right) = 0, \;\; \forall i \neq i' \end{cases} \qquad (4.1)$$

where, omitting the subscripts for sake of brevity, $\mathbf{x}$ are vectors of predictors (available in the two sampling frames), $\boldsymbol{\varphi}$ are the vectors of regression coefficients and $f(\mathbf{x}; \boldsymbol{\varphi})$ are known functions, u are the error terms, $\tilde{y}$ are the predicted values and $E_M(\cdot)$ denote the expectations under the models. The predictors $\mathbf{x}$ in the unit and cluster level models can be different. We assume that the parameters of the models are known, although in practice they are usually estimated.

Even if the model $f_r(\cdot)$ is not known, the model expectations at cluster level for the population U^B can be derived from a model defined at *elementary* unit level, indicated with $f_{re}(\cdot)$. The elementary unit level model can be stated as $y_{ik,r} = \tilde{y}_{ik,r} + u_{ik,r} = f_{re}(\mathbf{x}_{ik}; \boldsymbol{\varphi}_r) + u_{ik,r}$; $E_{M_{re}}(u_{ik,r}) = 0$; $E_{M_{re}}(u_{ik,r}^2) = \sigma_r^2$; $E_{M_{re}}(u_{ik,r}, u_{ik',r}) = \sigma_r^2 \ \rho_r \ \ \forall k \neq k'$; $E_{M_{re}}(u_{ik,r}, u_{i'k',r}) = 0 \ \ \forall i \neq i'$; where ρ_r is the intra-cluster correlation.

The model expectations at cluster level on the right-hand side of (4.1) can be easily derived as:

$$\tilde{y}_{i,r} = \sum_{k=1}^{M_i^B} \tilde{y}_{ik,r}; \ \ \sigma_{i,r}^2 = M_i^B \sigma_r^2 [1 + (M_i^B - 1)\rho_r]; \ \ E_{M_r}(u_{i,r}, u_{i',r}) = 0 \ \text{ for } \ i \neq i'.$$

Note that the *working* models (4.1) are variable specific. They are introduced as useful tools for developing the sampling design, but they are not necessarily representing exactly the real models generating the data.

According to (4.1), the model predictions and the variances of the z variables are given by

$$E_{M_r}(z_{j,r}) = \tilde{z}_{j,r} = \sum_{i=1}^{N^B} \tilde{L}_{j,i}^B \, \tilde{y}_{i,r} \ \text{ and } \ V_{M_r}(z_{j,r}) = \sigma_{j,zr}^2 = \sum_{i=1}^{N^B} \left(\tilde{L}_{j,i}^B\right)^2 \ \sigma_{i,r}^2. \tag{4.2}$$

Thus, in the optimization problem (3.1), the variance terms, $V(\hat{Y}_v^A \,|\, \mathbf{m}^A)$ and $V(\hat{Y}_r^B \,|\, \mathbf{m}^A)$, are replaced by the Anticipated Variances. Denoting with $E(\cdot)$ the expectation under the sampling design, the anticipated variance (AV) of $\hat{Y}_v^A$ may be reformulated as follows:

$$\text{AV}(\hat{Y}_v^A) = E_{M_v} E(\hat{Y}_v^A - Y_v^A)^2 = E_{M_v} V(\hat{Y}_v^A - Y_v^A) + V_{M_v} E(\hat{Y}_v^A - Y_v^A).$$

We have

$$E(\hat{Y}_v^A - Y_v^A) = 0,$$

and

$$V(\hat{Y}_v^A - Y_v^A) = V(\hat{Y}_v^A \,|\, \mathbf{m}^A) \ \cong \ \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) \eta_{j,v}^2.$$

The same result may be derived for the estimate $\hat{Y}_r^B$. Thus, we obtain the following expressions:

$$\text{AV}(\hat{Y}_v^A) = E_{M_v} V(\hat{Y}_v^A \,|\, \mathbf{m}^A) \ \cong \ \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_v}(\eta_{j,v}^2) \tag{4.3}$$

$$\mathrm{AV}\left(\hat{Y}_r^B\right) = E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_r}\left(\eta_{j,r}^2\right) \tag{4.4}$$

where $E_{M_v}\left(\eta_{j,v}^2\right)$ and $E_{M_r}\left(\eta_{j,r}^2\right)$ are given by expressions (A.2) and (B.2) of Appendices A and B.

The problem (3.1) for searching the optimal $\boldsymbol{\pi}^A$ vector is then reformulated as follows:

$$\begin{cases} \min \sum_{j \in U^A} c_j \ \pi_j^A \\ E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \le V_v^* \quad \forall v = 1, \ldots, V \\ E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \le V_r^* \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \le 1 \quad \forall j = 1, \ldots, M^A. \end{cases} \tag{4.5}$$

Remark 4.1. The anticipated variances in (4.5) have cumbersome formulae. A conservative simplified expression of $E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right)$ is given in Remark 4.1 of Falorsi and Righi (2015). More simplified conservative approximations of both $E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right)$ and $E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$ are obtained by approximating the sampling design variance with the Poisson sampling variance. We then have

$$E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \le \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_v}\left(y_{j,v}^2\right),\ E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \le \sum_{j \in U^A} \left(\frac{1}{\pi_j^A} - 1\right) E_{M_r}\left(z_{j,r}^2\right),$$

replacing $\eta_{j,\upsilon}$ and $\eta_{j,r}$ by $y_{j,\upsilon}$ and $z_{j,r}$, respectively, where $E_{M_v}\left(y_{j,v}^2\right) = \tilde{y}_{j,v}^2 + \sigma_{j,v}^2$ and $E_{M_r}(z_{j,r}^2) = \tilde{z}_{j,r}^2 + \sigma_{j,zr}^2$ (see Appendix B). Conservative approximations are a safe choice in this setting, since they eliminate the risk of defining an insufficient sample size for the expected accuracies.

Remark 4.2. Lavallée and Labelle-Blanchet (2013) deal with the problem of indirect sampling applied to skewed populations by suggesting eight alternative methods for modifying the links, $l_{j,ik}$, to reduce the variance of the estimates in the presence of skewed populations, while keeping estimation unbiased. Using the methods 2 and 3 proposed by these authors, the algorithm can run by simply replacing the links $l_{j,ik}$ by weighted links, $\theta_{j,ik}$, in $E_{M_r} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$.

Context 2. The links $l_{j,ik}$ are not known with certainty but the probabilities of links existing, $\Pr\left(l_{j,ik} = 1\right) = \lambda_{j,ik}$, are available.

To include the linkage uncertainty in the optimization, we assume the links follow a Bernoulli model M_l, $l_{j,ik} \sim \mathrm{B}\left(\lambda_{j,ik}\right)$, where $E_{M_l}\left(l_{j,ik}\right) = \lambda_{j,ik}$ and $V_{M_l}\left(l_{j,ik}\right) = \lambda_{j,ik}\left(1 - \lambda_{j,ik}\right)$. We assume the parameters $\lambda_{j,ik}$ to be known, although in practice they are usually estimated with probabilistic record linkage procedures (Lavallée and Caron, 2001). For the agricultural example, such a situation would occur when, for instance, the population of farms is linked to the population of rural households using probabilistic record linkage because no common identifier exists. In this framework, the anticipated variance must take into account both models M_l and M_r. Since

$$E_{M_l}E_{M_r}E\left(\hat{Y}_r^B - Y_r^B\right)^2 = E_{M_l}E_{M_r}V\left(\hat{Y}_r^B - Y_r^B\right) + E_{M_l}V_{M_r}E\left(\hat{Y}_r^B - Y_r^B\right) + V_{M_l}E_{M_r}E\left(\hat{Y}_r^B - Y_r^B\right)$$

and $E\left(\hat{Y}_r^B - Y_r^B\right) = 0,$ the problem (4.5) can be reformulated as follows:

$$\begin{cases} \min \sum_{j\in U^A} E_{M_l}\left(c_j\right)\ \pi_j^A \\ E_{M_v}V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \le V_v^* \quad \forall v = 1, \ldots, V \\ E_{M_l}E_{M_r}V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \le V_r^* \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \le 1 \quad \forall j = 1, \ldots, M^A \end{cases} \tag{4.6}$$

where

$$E_{M_l}E_{M_r}V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong \sum_{j\in U^A}\left(\frac{1}{\pi_j^A} - 1\right)E_{M_l}E_{M_r}\left(\eta_{j,r}^2\right), \tag{4.7}$$

$$E_{M_l}\left(c_j\right) = f_c\left(\Lambda_j^A; C^B\right),$$

with $\Lambda_j^A = \sum_{i=1}^{N^B}\Lambda_{j,i}^B$ and $\Lambda_{j,i}^B = \sum_{k=1}^{M_i^B}\lambda_{j,ik}.$

The main results for the derivation of the expression of $E_{M_l}E_{M_r}V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right)$ are given in Appendix C. These are derived using Taylor series approximation and postulating the independence of the process which generates the links $l_{j,ik}$ with the one that creates the variables of interest $y_{i,r}$. Under these approximations, the predicted values $\tilde{z}_{j,r}$ are obtained as

$$\tilde{z}_{j,r} \cong \sum_{i=1}^{N^B}\tilde{\Lambda}_{j,i}^B\tilde{y}_{i,r} \tag{4.8}$$

where

$$\tilde{\Lambda}_{j,i}^B = \frac{\Lambda_{j,i}^B}{\Lambda_i^B}$$

with

$$\Lambda_i^B = \sum_{j=1}^{M^A}\Lambda_{j,i}^B. \tag{4.9}$$

The uncertainty on total survey costs, which depends both on the selected sample and the model uncertainty on costs, obliges us to consider the expected costs $E_{M_l}\left(c_j\right)$ in the optimization problem. Steel and Clark (2014) show how the uncertainty on the expected costs can affect the accuracy of the sample design.

Context 3. Data integration is not possible because the record linkage process does not provide good linkages, or simply because the frame of population U^B does not exist.

This is the most common context in developing countries. It may also characterize specific survey contexts in developed countries, for instance in the case of hard-to-reach populations. Returning to the agricultural example, this would mean that one might have a list of farms, but not a list of rural households.

In this case, the problem of optimal integrated sampling can be solved by using all the available information, even if of poor quality. In the following, three options for dealing with the optimization problem are illustrated starting from the option which requires the minimum of information to those which need more information that could be expensive to obtain.

Option 3.1. *Building the predictions of the z variables and decreasing the variance thresholds V_r^* by a scale factor.* Suppose that from the frame of population U^A, it is possible to know the values of a size variable γ related to the total links L_j^A of the units j. For instance, if the population U^A is a population of farms and the population U^B is a population of households, then the number of workers in the farms (variable γ_j) can represent a good approximation of the total number of links, L_j^A, of the farm. Suppose further that the totals or the estimated totals, $\tilde{Y}_{r(q)}^B$, are available at certain domain level, $U_{(q)}^B$ $(q = 1, \ldots, Q)$, defined at geographic level, with $U^B = \bigcup_{q=1}^{Q} U_{(q)}^B$ and $U_{(q)}^B \cap U_{(q')}^B = \varnothing$ for $q \neq q'$. Then the predicted z variables can be defined as:

$$\tilde{z}_{j,r} = \frac{\gamma_j}{\sum_{l \in U_{(q)}^A} \gamma_l} \tilde{Y}_{r(q)}^B \quad \text{for} \quad j \in U_{(q)}^A, \tag{4.10}$$

where $U_{(q)}^A$ denotes the geographic domain q for the population U^A. In practice, the ratio approach in (4.10) assumes that unit j can be given a share of the total $\tilde{Y}_{r(q)}^B$ proportional to the size of the unit itself. Other examples of building the predictions of the z values are illustrated in Section 5.3.2 of *Guidelines on Integrated Survey Framework* (FAO, 2015).

Having determined the predictions, $\tilde{z}_{j,r}$, it may be reasonable to assume that the following relationship holds:

$$E_{M_{zr}}\left(z_{j,r}^2\right) = \tilde{z}_{j,r}^2 + \sigma_{j,zr}^2 \cong k_r \tilde{z}_{j,r}^2, \tag{4.11}$$

where $k_r > 1$. Under (4.11), it is straightforward to show that

$$E_{M_{zr}} V\left(\hat{Y}_r^B \,\middle|\, \mathbf{m}^A\right) \cong k_r V\left(\hat{\tilde{Y}}_r^B \,\middle|\, \mathbf{m}^A\right),$$

where $\hat{\tilde{Y}}_r^B = \sum_{j \in s^A} w_j^A \tilde{z}_{j,r}$. The sampling variance $V\left(\hat{\tilde{Y}}_r^B \,\middle|\, \mathbf{m}^A\right)$ may be computed using expressions (2.2), (2.3), (2.4) and (2.5) by substituting the variable $y_{j,v}$ the prediction $\tilde{z}_{j,r}$. The optimization problem for searching for the optimal $\boldsymbol{\pi}^A$ vector can then be reformulated as:

$$\begin{cases} \min \sum_{j \in U^A} E_{M_\Lambda}\left(c_j\right) \pi_j^A \\ E_{M_v} V\left(\hat{Y}_v^A \,\middle|\, \mathbf{m}^A\right) \leq V_v^* \quad \forall v = 1, \ldots, V \\ V\left(\hat{\tilde{Y}}_r^B \,\middle|\, \mathbf{m}^A\right) \leq V_r^* / k_r \quad \forall r = 1, \ldots, R \\ 0 < \pi_j^A \leq 1 \quad \forall j = 1, \ldots, M^A. \end{cases} \tag{4.12}$$

The sample designer may find the solution by running the optimization problem (4.12) with alternative reasonable choices of the k_r value (e.g., $k_r = 2$, 3 or 4), and studying the sensitivity of the different solutions. Note that $k_r \cong 1 + \left[\mathrm{CV}\left(z_{j,r}\right)\right]^2$, where $\left[\mathrm{CV}\left(z_{jr}\right)\right]^2 = \sigma^2_{j,zr} \big/ \tilde{z}^2_{j,r}$. Therefore (4.11) holds if the $\left[\mathrm{CV}\left(z_{j,r}\right)\right]^2$ values are approximately constant.

Option 3.2. *Extremal case of Context 2, with uniformity of links in specific domains.* If the number or estimated number of clusters and of elementary units $N^B_{(q)}$ and $M^B_{(q)}$ of the domains $U^B_{(q)}$ $(q = 1, \ldots, Q)$ are available, then in the absence of information on the links $l_{j,ik}$, it might be reasonable to assume that these are homogeneous over the domains; that is, $l_{j,ik} \sim B\left(\lambda_{j,ik}\right)$, where $\lambda_{j,ik} = \gamma_j \big/ M^B_{(q)}$.

Furthermore, suppose that, in this context, the predictions $\tilde{y}_{i,r}$ and the sampling variances $\sigma^2_{i,r}$ could be assumed to be homogeneous within the domains $U^B_{(q)}$, i.e., $\tilde{y}_{i,r} = \tilde{y}_{r(q)}$ and $\sigma^2_{i,r} = \sigma^2_{r(q)}$ for $i \in U^B_{(q)}$. Then, the optimization problem may be dealt with as an extremal case of Context 2, with uniformity of links in specific domains.

Remark 4.3. Note that with this option, the predictions $\tilde{z}_{j,r}$ are equivalent to those expressed in (4.10). Indeed, it is reasonable to consider that, in the absence of information, the size in terms of elementary unit of the cluster U^B_i can be set as equal to its mean defined at the domain level: $M^B_i \cong \bar{M}^B_{(q)} = M^B_{(q)} \big/ N^B_{(q)}$ for $U^B_i \in U^B_{(q)}$. Then, the following approximations hold

$$\Lambda^B_{j,i} = \sum_{k \in U^B_i} \lambda_{j,ik} \cong \bar{M}^B_{(q)} \frac{\gamma_j}{M^B_{(q)}} = \frac{\gamma_j}{N^B_{(q)}}; \quad \Lambda^B_i = \sum_{j \in U^A_{(q)}} \Lambda^A_{j,i} \cong \frac{1}{N^B_{(q)}} \sum_{j \in U^A_{(q)}} \gamma_j .$$

Therefore, setting $\tilde{Y}^B_{r(q)} = \tilde{y}^B_{r(q)} N^B_{(q)}$ and postulating the independence of the process which generates the links $l_{j,ik}$ with the one that creates the variables of interest $y_{i,r}$, we can obtain

$$\tilde{z}_{j,r} \cong \sum_{i \in U^B_{(q)}} \frac{\Lambda^B_{j,i}}{\Lambda^B_i} \tilde{y}_{r(q)} = \sum_{i \in U^B_{(q)}} \frac{\gamma_j \big/ N^B_{(q)}}{\sum_{j \in U^A_{(q)}} \gamma_j \big/ N^B_{(q)}} \tilde{y}_{r(q)} = \frac{\gamma_j}{\sum_{j \in U^A_{(q)}} \gamma_l} \tilde{y}^B_{r(q)} N^B_{(q)} \quad \text{for} \quad j \in U^A_{(q)}.$$

Option 3.3. *Modeling the $z_{j,r}$ values.* Another alternative may be carried out by trying to model directly the z-values and the total number of links L^A_j with models of the type:

$$\begin{cases} z_{j,r} = \tilde{z}_{j,r} + u_{j,zr} = f_{zr}\left(\mathbf{x}_j; \boldsymbol{\varphi}_r\right) + u_{j,zr} \\ E_{M_{zr}}\left(u_{j,zr}\right) = 0, \; E_{M_{zr}}\left(u^2_{j,zr}\right) = \sigma^2_{j,zr}, \; \forall j \\ E_{M_{zr}}\left(u_{j,zr}, \; u_{j',zr}\right) = 0, \; \forall j \neq j' \end{cases} , \quad \begin{cases} L^A_j = \Lambda^A_j + u_{j,\Lambda} = f_\Lambda\left(\boldsymbol{\theta}_j; \boldsymbol{\varphi}_\Lambda\right) + u_{j,\Lambda} \\ E_{M_\Lambda}\left(u_{j,\Lambda}\right) = 0, \; E_{M_\Lambda}\left(u^2_{j,\Lambda}\right) = \sigma^2_{j,\Lambda}, \; \forall j \\ E_{M_r}\left(u_{j,\Lambda}, \; u_{j',\Lambda}\right) = 0, \; \forall j \neq j' \end{cases} \quad (4.13)$$

where $\mathbf{x}_j$ and $\boldsymbol{\theta}_j$ are vectors of auxiliary variables. The predictions Λ^A_j need to be positive. A useful model is the log-linear one (Xu and Lavallée, 2009): $\log\left(\Lambda^A_j\right) = \boldsymbol{\theta}'_j \boldsymbol{\varphi}_\Lambda$. The model on the right hand side of (4.13) allows the prediction of the total number of links Λ^A_j of the unit j, thus defining the expected survey cost attached to it. The optimization problem could be carried out using the variances of the predictions of the models (4.13).

Remark 4.4. Option 3.1 requires the minimum of information for the construction of the predictions $\tilde{z}_{j,r}$ and needs us to define of plausible values for the constants k_r. Option 3.2 involves the same information as Option 3.1 for the construction of the predictions $\tilde{z}_{j,r}$ (see Remark 4.3) but requires an estimate of the parameters $\sigma^2_{r(q)}$. These estimates can be obtained from either pilot or previous surveys conducted directly on the population U^B. Option 3.3 is the most complex and expensive, since it involves carrying out indirect pilot surveys on the population U^A for building plausible predictions of the parameters $\tilde{z}_{j,r}$, Λ^A_j, $\sigma^2_{j,zr}$ and $\sigma^2_{j,\Lambda}$.

Remark 4.5. A good strategy that should be robust against model failure is to select a balanced sample with respect to the auxiliary variables $\mathbf{x}_j$. In this case, the auxiliary variables $\mathbf{d}_j$ of the balancing equations are replaced by the augmented variables $\mathbf{d}^*_j = \left(\mathbf{d}'_j, \mathbf{x}'_j \left/ \pi^A_j\right.\right)'$. For the calculation of the variances, the residuals $\eta_{j,v}$ are substituted by the modified residuals $\eta^*_{j,v} = y_{j,v} - \pi^A_j \left(\mathbf{d}^*_j\right)' \boldsymbol{\beta}^*_v$, where $\boldsymbol{\beta}^*_v = (\boldsymbol{\Delta}^*)^{-1} \sum_{j\in U^A} \pi^A_j \left(\frac{1}{\pi^A_j} - 1\right) \mathbf{d}^*_j y_{j,v}$ with $\boldsymbol{\Delta}^* = \sum_{j\in U^A} \mathbf{d}^*_j \left(\mathbf{d}^*_j\right)' \pi^A_j \left(1 - \pi^A_j\right)$. For the modified residuals $\eta^*_{j,r}$, similar expressions are used.

Remark 4.6. A proportional-to-population-size allocation may be a reasonable strategy for stratified sampling designs in which the total sample size m^A is fixed. In this case the stratum sample size, m^A_h, may be defined as $m^A_h = m^A \left(\sum_{j\in U^A_h} x_j \left/ \sum_{j\in U^A} x_j\right.\right)$, where x_j is the measure of the size.

5 Empirical results

The results herein illustrated are obtained using real data from Districts 7, 8, 9 of the Gaza Province, Mozambique. They summarize the empirical results from an evaluation study illustrated in FAO (2014). Other empirical results of the proposed strategies (FAO, 2015) have been conducted on the database of agricultural households from Burkina Faso's General Census of Agriculture and confirm the general results illustrated below.

In the analysis using Mozambique data, the population U^A refers to farms. The database used for the experimentation include environmental and economic variables and gathers the information from the 2007 census of large and medium farms and from a sample survey of small farms (for the same year). The overall number of records is about 36,890, of which 890 are large and medium farms.

The second population, U^B, is the 2007 household census. The database's records are the individuals involved in agricultural, fishing, or forestry activities. The database contains approximately 54,000 records and includes several socio-demographic environmental and economic variables. The databases of the two populations were merged, creating a Master Sampling Frame (MSF) with artificial links between individuals and farms. The merging procedure exploited the following variables: for individuals, the type of job and the district of residence; for farms, the sector, the district and the number of employed persons by type of job. Before merging, a cleaning step of U^B was carried out, discarding records that did not feature the job type variable (approximately 9,000 records). Subsequently, approximately 36,000 records of U^B declare to be a

farmholder without any employed persons. For these cases, a one-to-one farm-individual link was defined. The remaining individuals were linked with the 890 farms, according to the following hierarchical rules:

- Each farm was linked to a number of individuals equal to the number of workers, depending upon job type;
- Individuals and farms in the same district were linked;
- Individuals were linked with private/public governmental farms when the type of employment and the farm sector agree.

The links were generated randomly, according to the categories defined by the hierarchical rules. The exercise did not seek to predict the links that actually exist in the two populations, but rather to create a realistic dataset for the evaluation.

Although the datasets of the two populations include several variables, in this study we have decided to focus on two of these. For U^A, we consider the number of animals, while for U^B, we consider the number of trees. This is in order to better highlight the impact (in terms of both accuracy and sample size) that the different contexts, described in Section 4, have on the sample of the population U^B. Summary statistics on these variables are shown in Table 5.1.

Table 5.1
The variables used in the simulation with data from Mozambique

Population*	Number of records*	Variable	Mean value	%CV**
U^A: Farms	36,890***	Number of animals	11.1	681.6
U^B: Households	45,000	Number of trees	4.5	107.5

* Districts 7, 8, 9 of the Gaza Province, Mozambique.
** %CV = (population standard deviation/mean) × 100.
*** From these, 890 are large and medium farms.

For both populations, we have considered as domains of interest the districts (3 domains) and the province (1 domain). Therefore, in total we consider 8 target totals of interest (2 variables × 4 domains).

5.1 Optimal designs for the different contexts

In the following, we address four contexts:

Context 0. No control on the sample of the population U^B. The sample is planned, controlling only the accuracy of the estimates of the variables of the farms. Once the sample for U^A is selected, the units of U^B, linked to those selected for the sample of the U^A population are included in the sample via the indirect sampling mechanism. The expected percent CVs, %CV, of the estimates obtained from the indirect sample of households are then computed as $\%\mathrm{CV} = \left(\sqrt{\mathrm{AV}\left(\hat{Y}\right)}\Big/Y\right)\times 100$.

Context 1. Sampling frames exist for both populations. All links are known and an integrated sample design is used, finding an optimal solution considering both populations. Therefore, the multivariate allocation is

carried out, controlling the accuracy of estimates from both the direct sample of farms and the indirect sample of individuals.

Context 2. Sampling frames exist for both populations, but links are estimated probabilities and an integrated sample design is used.

Context 3. A frame exists only for the population U^A. An integrated sample design is studied considering Options 3.1 and 3.2, which represent the most feasible solutions in real contexts.

Contexts 1, 2 and 3 are those defined in Section 4. Context 0 is introduced because it represents a useful tool for the evaluation of the integrated strategy.

A stratified sampling mechanism is assumed for the first population U^A, where the strata U_h^A are defined as districts (7, 8 and 9) by size class (1, 2, 3-4, 5-9, 10-19, 20-49, 50-99, 100+) based on the number of farm workers, thus obtaining 21 strata. As regards models (4.1), we considered mean stratum models with $\tilde{y}_{j,v} = \tilde{y}_{v,(h)}$ and $\sigma^2_{j,v} = \sigma^2_{v,(h)}$ for $j \in U_h^A$. These specifications lead to a standard SSRSWOR design for the farms where the strata coincide with the planned domains (see Falorsi and Righi, 2015, Remark 4.2). For the evaluation we used the exact formula of the variance for a SSRSWOR instead of using the approximation of variance for a SSRSWOR given in Section 2; however the two expressions are substantially equivalent. For U^B, we also consider a mean model, defined at district level d, with $\tilde{y}_{i,r} = \tilde{y}_{r,(d)}$, and $\sigma^2_{i,r} = \sigma^2_{r,(d)}$ for $i \in U_d^B$.

The evaluation studies use software, developed in the R language, that implements the optimal sampling for the standard SSRSWOR designs as well as for more general sampling designs (e.g., balanced designs and incomplete stratification designs). It is available at http://www.istat.it/en/tools/methods-and-it-tools/design-tools/multiwaysampleallocation). Once installed, the software features a comprehensive user guide in English. Another software which considers only the SSRSWOR designs is MAUSS-R available at http://www.istat.it/it/strumenti/metodi-e-software/software/mauss-rdownload.

For each context, the variance constraints are expressed in terms of %CVs. The analyses presented in this section are focused on the contexts, and we use a symplified version of the cost functions. The cost c_j for observing the unit j in the population U^A with the linked units in the population U^B is fixed as equal to 1. More detailed analyses on costs are presented in Section 5.2.

Some further specifications for each context are herein illustrated (see Table 5.2).

Context 0. The variance constraints are fixed (only for the farm estimates: number of animals) at 6.5% at the province level and at 10% at the district level, resulting in a sample of 2,122 farms.

Context 1. The constraints for the farm and household estimates have been fixed so as to determine a sample roughly of 2,100 farms. In this way, the variance constraints are fixed for the farm estimates, animals at 10% at the province level and at 15% at the district level. Those for the household estimates are fixed at 2.5% at the province level and 5% at the district level. Note that this choice of constraints makes it possible to carry out the comparison between the two contexts with roughly the same sample size, even if in Context 1, the variance constraints on the estimates of the population U^A are larger than those fixed in Context 0.

Context 2. The CV constraints for the household and farm estimates are equal to those adopted in Context 1. The integrated observation is planned in the sample design phase by taking into account the uncertainty in the links. This has been carried out by considering a simplified model which assumes that, for each worker in a given farm, there is only one *strong* link (with value ψ) with an individual in the population of households and α weak links (with value τ) with other individuals in the same district, where ψ and τ are probabilities, where $\psi \gg \tau$. Let $l_{j\omega,ik}$ denote the link between the worker ω of the farm j and the individual k of the household i and suppose that these links follow a Bernoulli model M_l, where

$$E_{M_l}\left(l_{j\omega,ik}\right) = \lambda_{j\omega,ik} = \begin{cases} \psi \text{ for only one worker } j\omega \in U^A \text{ and one individual } ik \in U^B \\ \tau \text{ for only one worker } j\omega \in U^A \text{ and } \alpha \text{ individuals } ik \in U^B \end{cases}, \quad (5.1)$$

in which $\tau = \frac{1-\psi}{\alpha}$.

In the simulation we have considerered different combinations of values of the probabilities of strong links, ψ, of weak links, τ, and of the number of individuals, α, with a weak link. These combinations are illustrated in Table 5.3.

Context 3. The CV constrains for the households and farms estimates are equal to those adopted in Context 1. In Table 5.3, we derived the allocation considering the Option 3.2, proposed for Context 3. The results of Option 3.1 are presented at the end of this section.

Finally, note that for all the three contexts, the optimization problem has been set up in terms of π_j^A. With a SSRSWOR design, this may be seen as a problem of allocation for stratified sampling.

Table 5.2
Variance constraints in the different contexts

Contexts	Variance Constraints*			
	U^A: variable Animals		U^B: variable Trees	
	Province	District	Province	District
Context 0	6.5%	10%	No constraints	No constraints
Context 1	10%	15%	2.5%	5%
Context 2	10%	15%	2.5%	5%
Context 3	10%	15%	2.5%	5%

* Expressed in terms of %CV.

Table 5.3
Main results of the evaluation

Contexts		Sample size	Realized Coefficent of variations (%)						
			U^A: variable Animals				U^B: variable Trees		
			Province	District			Province	District	
				7	8	9		7	8	9
Context 0		2,122	6.5	10.0	10.0	10.0	1.5	6.8	12.7	1.4
Context 1		2,106	8.8	7.5	4.1	15.0	1.8	5.0	5.0	2.0
Context 2	$\psi = 0.90, \tau = 0.10, \alpha = 1$	2,146	8.8	7.2	4.1	15.0	2.2	5.0	5.0	2.4
	$\psi = 0.50, \tau = 0.10, \alpha = 5$	2,573	7.5	6.5	4.0	12.7	2.5	5.0	5.0	2.8
	$\psi = 0.30, \tau = 0.08, \alpha = 9$	2,767	7.0	6.4	4.0	11.9	2.5	5.0	5.0	2.8
	$\psi = 0.10, \tau = 0.09, \alpha = 9$	2,826	6.9	6.2	4.0	11.6	2.5	5.0	5.0	2.8
Context 3	Option 3.2	2,936	6.6	6.2	3.9	11.2	2.5	5.0	5.0	2.8

Looking at the main results of the evaluation, highlighted in Table 5.3, the following evidences emerge:

Context 0 vs Context 1. In the two contexts, the farm sample size is of about 2,100 farms.

- For Context 0, the expected %CVs of the farm estimates at the district level are exactly at the constraint level of 10%, defined for this context.
- In Context 1, we see that all the %CVs of the farm estimates at district level respect the constraints of 15% (defined for this context), being however considerably lower than 10% for the districts 7 and 8, showing that these districts are somewhat oversampled with respect to the target precisions. This is because in the second allocation, part of the farm sample is required to achieve the required indirect sample of households (FAO, 2014, studies this inefficiency issue in great detail).
- Considering now the precision of the estimates for the population U^B, we found that the expected sample sizes of households were approximately 5,300 records in both contexts. In Context 0, the %CVs are much higher than the desired level of 5%, being even larger than 12% in the District 8. With the sampling allocation resulting from Context 1, the desired precision of the estimates of population U^B are always respected, as well as those of population U^A, even if the constraints for these estimates have been defined larger than those adopted in Context 0.
- Thus, the integrated approach to the sampling allocation carried out in Context 1 enables control of the precision of the estimates for both populations of interest, however paying some loss in precision for the estimates for population U^A.

Context 1 vs Context 2. For the comparison between the Contexts 1 and 2, the analysis focuses upon the overall sample sizes, since the %CVs are under the constraint levels in both contexts.

- In the presence of strong links for Context 2 ($\psi = 0.90$, $\tau = 0.10$, $\alpha = 1$), there is only a small increase in the sample sizes (40 farms), while the CVs remain under the desired level of precision, altough being slightly increased for the household estimates.
- As the links become weaker, the sample sizes increase significantly. This is due to the achievement of the expected %CVs for the household estimates.
- Conversely in Context 2, the expected CVs for the farm estimates are lower than the targeted levels, suggesting that the the farms are somewhat oversampled with respect to the target levels of precision.

Context 3 vs other contexts. Having considered the Option 3.2 in Table 5.3, Context 3 may be considered as an extremal case of Context 2. Even in this case, the analysis focuses on the overall sample sizes, since all the %CVs are under the constraint levels:

- The maximization of the links uncertainty, represented by Option 3.2, causes an increase in the sample size of about 30%: from the sample size of 2,106 to that of 2,936.
- Examining Context 2, we note that we obtain results similar to those of Context 3 when the level, ψ, of the strong link is around 10%.

- Even in this case, the farms are somewhat oversampled with respect to the target levels of precision.

More detailed analysis of Context 3. Below, some more detailed analyzes are illustrated, aimed at better clarifying some aspects of the problem of sampling allocation for the integrated observation of two related populations. We explore Option 3.1 and the proportional allocation proposed in Remark 4.6 because of their practical importance. For the proportional allocation, we considered as measure of size (see Remark 4.6) the total number of employed people. The $\tilde{z}_{j,r}$ are obtained by expression (4.10). In this context, we have to define the k_r value. In order to identify a single k_r value, we exploited the data of Context 1 and first computed for each stratum the coefficent of variation of $z_{j,r}$, $\mathrm{CV}\left(z_{h,r}\right)$. Then, specific k_r values were computed at stratum level, as $k_{hr} = 1 + \left[\mathrm{CV}\left(z_{h,r}\right)\right]^2$ and finally the k_r value considered in this evaluation was obtained as a weighted mean of the k_{hr} values: $k_r = \sum_h k_{h,r} w_h$. We computed the weights w_h with two different alternatives, resulting in the two values: $k_r = 2.75$ and $k_r = 2.16$. With the first alternative, the w_h were defined proportional to the sum of the weights L_j^A at stratum level; while in the second alternative, the w_h were defined proportional to the quantity $\sqrt{\mathrm{CV}\left(z_{h,r}\right)}\,\overline{Y}_{r,h}^{B}\,N_h^A$, where $\overline{Y}_{r,h}^{B}$ and N_h^A are the mean value of variable y_r and the number of units in the stratum, respectively. For each alternative, we ran the problem (4.12), with the constraints defined in Table 5.2 for Context 1, obtaining an overall sample size, n^A, equal respectively to 1,639 and 1,517. The main results of the experiment are illustrated in Table 5.4, in which for both k_r values we show: (i) the expected %CVs, obtained as solution of problem (4.12) under the hypotesis that relation (4.11) holds; (ii) the *true* expected %CVs, that is, those obtained under Context 1 on the basis of the stratum sample sizes defined by the solution of the problem (4.12); and (iii) the true %CVs obtained, under Context 1, with the proportional allocation proposed in Remark 4.6.

Table 5.4
Expected and realized %CVs of the domain estimates of total number of trees with the sampling allocation obtained as solution of problem (4.12) and proportional allocation

Estimation Domains	$k_r = 2.75$, $n^A = 1{,}639$			$k_r = 2.16$, $n^A = 1{,}517$		
	Expected %CV, obtained as solution of problem (4.12), assuming that (4.11) holds	True expected %CV, under Context 1, with allocation defined by (4.12)	True expected %CV under Context 1, with proportional allocation	Expected %CV, obtained as solution of problem (4.12), assuming that (4.11) holds	True expected %CV, under Context 1, with allocation defined by (4.12)	True expected %CV under Context 1, with proportional allocation
Province	2.11	1.94	1.76	2.11	2.04	1.83
District 7	4.95	6.80	6.10	4.95	8.20	6.34
District 8	4.99	6.45	13.23	4.99	6.45	13.79
District 9	2.36	2.0	1.81	2.36	2.0	1.88

The main findings of this evaluation are the following:

- The strategy proposed by Option 3.1 seems to be effective, since it allows control of the sampling errors, avoiding the situation where these exceed by a large amount the desired accuracy for the different estimation domains.

- With the use of a unique k_r, the true expected %CVs (columns 3 and 7 of the Table 5.4) for some estimation domains are larger than the defined benchmarks and, in some others, the estimates are much more accurate than required.
- The choice of a larger value of the k_r parameter seems to be a safe choice, if the main objective of the sampling allocation is to avoid sampling errors in specific estimation domains that are too large.
- Even if it seems effective for the accuracy of the overall estimate at province level, the proportional allocation (columns 4 and 8 of the Table 5.4) does not allow control of extremal discrepancies from the expected accuracy in some estimation domains (see district 8).

5.2 Evaluation on costs

This evaluation considers Context 1 in which the sampling frames for both populations are available, and in which it is possible to build an integrated observation of the two populations. We focus on two observational strategies: the first considers two independent samples, one for farms and one for individuals. Therefore, a truly integrated analysis cannot be performed. The second observational strategy applies an integrated sampling design that selects a direct sample of farms and an indirect sample of the households of the workers of the sampled farms.

We adopted the variance constraints established for the Context 1 (see Table 5.5).

Table 5.5
Variance Constraints in the evaluation on costs

Variance Constraints *			
U^A: variable Animals		U^B: variable Trees	
Province	**District**	**Province**	**District**
10%	15%	2.5%	5%

* Expressed in terms of %CV.

For the direct sampling designs, we adopted a SSRSWOR design, where the population U^A was stratified by crossclassfying the districts and the size classes of the farms, and the population U^B was stratified by district. The cost for interviewing the farms varies ($C^A = 1, 2, 5$ and 10), which leads to performing four different evaluations. The cost C^B for interviewing an individual is set equal to 1.

For indirect sampling designs, we define the overall cost of interviewing the farm and the farms workers together by two different specifications of equation (3.2):

$$c_j = C^A + L_j^A C^B, \tag{5.2}$$

$$c_j = C^A + \sqrt{L_j^A}\, C^B. \tag{5.3}$$

The increase of the cost function (5.3) is lower than the increase of the cost function (5.2) when L_j^A increases.

We perform a precision-constrained optimal allocation for both independent sampling designs. The different C^A values (1, 2, 5 and 10) do not affect the farm sample size while the costs increase proportionally. Given the variance constraints in Table 5.5 with the independent strategy, the sample sizes of farms and individual are respectively 1,010 and 3,388. The total cost is then 4,398 when setting $C^A = 1$. In the integrated sample strategy, the costs do affect the allocation, essentially because if the farm interview costs increases, the number of sampled farms decreases and the allocation increases sample sizes of strata with the largest farms.

Table 5.6 below shows the sample sizes of farms and the expected sample sizes of individuals when cost model (5.2) is used to calculate the costs of individual interviews in the integrated allocation. We see that the farm sample is more than double the sample size, considering farms alone (1,101). The increase in size is due to precision constraints on the household estimates.

Table 5.6
Sample sizes for the integrated sample allocation, when the overall individual costs are given by (5.2)

Cost per farm interview (C^A)	1	2	5	10
Farms	2,388	2,289	2,190	2,137
Individuals	4,504	4,491	4,862	4,905

Table 5.7 below shows the allocation when equation (5.3) is used for the cost of individual interviews in the integrated allocation.

Table 5.7
Sample sizes for the integrated sample allocation, when the overall individual costs are given by (5.3)

Cost per farm interview (C^A)	1	2	5	10
Farms	2,135	2,121	2,111	2,108
Individuals	4,834	4,874	5,283	5,360

Tables (5.6) and (5.7) show that the integrated sample size of farms is roughly twice that of the independent allocation of farms. Thus the expected variance of the estimates will be much lower than the desidered variance constraints, suggesting that integrated sample allocation mainly depends on the variance constraints related to the individual parameters to be estimated.

Figures 5.1 and 5.2 show the cost for independent and integrated sampling. The integrated observational strategy is generally more expensive, except when the cost per farm interview is equal to 1 and the cost function given by (5.3). In this evaluation, the integrated nature of the sample is not needed as no cross tabulation of population U^A variables with population U^B variables are examined; then, the independent allocation will be more efficient in term of precision. Another cost function could however partially rebalance the two observational strategies in term of costs.

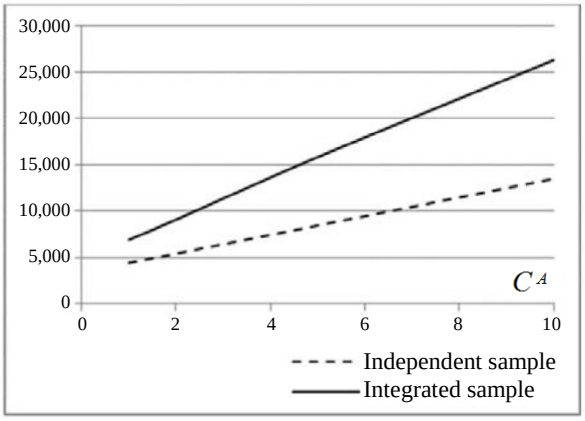


Figure 5.1 Overall costs integrated vs two independent allocations using (5.2).

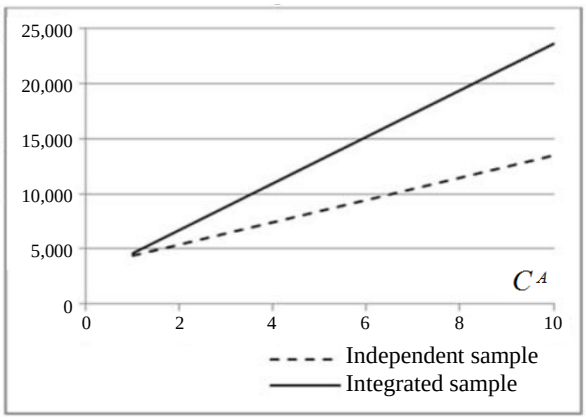


Figure 5.2 Overall costs – integrated vs two independent allocations using (5.3).

6 Conclusions

In this paper, we studied the problem of the definition of optimal sampling designs for survey strategies aiming at observing in an integrated way different statistical populations related to each other. This is particularly relevant in the agricultural sector where the integrated observation allows measurement of global phenomena that affect different statistical populations such as farms and households. The integrated observation is realized by directly sampling the first population and indirectly observing the second population, exploiting the links existing among the units of the two populations. We studied the problem considering three different contexts concerning information about the links. These range from two contexts in which the information is very rich, to the third context considering a case in which the information is very poor. The uncertainty on variables of the two populations, on links and on the z variables (built by the indirect sampling mechanism) is treated by introducing suitable superpopulation models for which expected values (of first and second order) are considered as known when launching the algorithm for the optimal sampling. Empirical studies were performed on real data of a developing country: Mozambique.

The main conclusions are summarized as follows.

Integrated vs independent observation. The integrated observation is essential to measure thoroughly global phenomena which impact on different populations. The main advantage is that it allows the cross tabulation of population U^A variables with population U^B variables. Furthermore, the integrated observation is necessary when the frame for the population U^B does not exist and an indirect sampling mechanism is needed. This is the case examined in Context 3. However, for Contexts 1 and 2, if only aggregates are examined independently from each other in the two populations, the independent allocation will be more efficient.

Cost issues. The loss in efficiency of the integrated observation can be reduced if, as assumed with cost function (5.3), the average cost of observing the elementary unit of U^B decreases when the size of the indirectly observed clusters increase. In this case, the performance of the integrated sample allocation and of the two independent allocations could be closer or similar as in the evaluation study. Nevertheless, it is complex to establish which relationship between C^A and C^B leads to two strategies with similar costs, since the allocations depend on not only on the cost of interview but also on the variability of the target parameters in the two populations and on the set of variance constraints.

Controlling the errors in the design phase. The integrated approach to allocation enables the CVs of the estimates for integrated populations to be controlled. If this is not done, the CVs of the indirectly observed population might be very high.

The impact on the uncertainty on the sample sizes. An increase in the model variances (on the variables or on the links) causes a significant increase in the sample sizes. This stresses the need of having good models for predicting the unknown variables or the links.

Appendix A

To obtain the model expectation $E_{M_v}\left(\eta_{j,v}^2\right)$, let $\boldsymbol{\eta}_v = \{\eta_{j,v}\}$ be the M^A vector of residuals, where

$$\boldsymbol{\eta}_v = \mathbf{Y}_v - \boldsymbol{\Pi}\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{Y}_v, \tag{A.1}$$

where $\mathbf{Y}_v = \{y_{j,v}\}$ denotes the M^A vector with the values of v^{th} variable of interest and $\boldsymbol{\Pi} = \text{diag}\{\pi_j^A\}$ indicates the diagonal matrix with the M^A inclusion probabilities. According to model (4.1), the vector $\mathbf{Y}_v$ may be expressed as $\mathbf{Y}_v = \tilde{\mathbf{Y}}_v + \mathbf{u}_v$, where $\tilde{\mathbf{Y}}_v = \{\tilde{y}_{i,v}\}$ and $\mathbf{u}_v = \{u_{i,v}\}$ denotes the M^A vectors of predictions and model residuals. Adopting the above matrix notation, the specific residuals $\eta_{j,v}$ can be expressed as $\eta_{j,v} = \left(\tilde{y}_{j,v} + u_{j,v}\right) - \pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\left(\tilde{\mathbf{Y}}_v + \mathbf{u}_v\right)$. Therefore, the model expected values of the squared terms are given by:

$$\begin{aligned} E_{M_v}\left(\eta_{j,v}^2\right) = \tilde{y}_{j,v}^2 + \sigma_{j,v}^2 \quad & - 2\pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\tilde{\mathbf{Y}}_v - 2\pi_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{0}_{j\sigma_v} \\ & + \pi_j^2\tilde{\mathbf{Y}}_v'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\tilde{\mathbf{Y}}_v \\ & + \pi_j^2\boldsymbol{\sigma}_v'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\boldsymbol{\sigma}_v \end{aligned} \tag{A.2}$$

where $\boldsymbol{\sigma}_v = \{\sigma_{j,v}\}$ is the M^A column vector of model standard errors of the V variables and $\mathbf{0}_{j\sigma_v} = (0, \ldots, \sigma^2_{j,v}, \ldots, 0)'$ is a vector in which the j^{th} element is equal to $\sigma^2_{j,v}$ and all other elements are zeroes. Using the above matrix notation and according to Falorsi and Righi (2015), the anticipated variance can be approximated by the following expression:

$$E_{M_v}\left[V\left(\hat{\bar{Y}}_v^A \,\middle|\, \mathbf{m}^A\right)\right] = \left[M^A/(M^A - H)\right]\left[\tilde{\mathbf{Y}}_v'\boldsymbol{\Pi}^{-1}\tilde{\mathbf{Y}}_v + \boldsymbol{\sigma}_v'\boldsymbol{\Pi}^{-1}\boldsymbol{\sigma}_v - \tilde{\mathbf{Y}}_v'\tilde{\mathbf{Y}}_v + \boldsymbol{\sigma}_v'\boldsymbol{\sigma}_v + \text{AAV}_{3,v}\right].$$

Letting $\mathbf{a}_v = \mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\tilde{\mathbf{Y}}_v$, $\mathbf{b}_v = \boldsymbol{\sigma}_v'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\boldsymbol{\sigma}_v$ and $\mathbf{c}_v = \boldsymbol{\sigma}_v'\text{diag}\left[\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'\right]\boldsymbol{\sigma}_v$, we then have $\text{AAV}_{3,v} = \mathbf{a}_v'(\mathbf{I} - \boldsymbol{\Pi})\left(2\tilde{\mathbf{Y}}_v - \boldsymbol{\Pi}\mathbf{a}_v\right) + \mathbf{1}'(\mathbf{I} - \boldsymbol{\Pi})\left(2\mathbf{b}_v - \boldsymbol{\Pi}\mathbf{c}_v\right)$, where the scalars defined as (A.1.4), (A.1.7) and (A.1.8) in Falorsi and Righi (2015) are respectively the elements of the vectors $\mathbf{a}_v$, $\mathbf{b}_v$ and $\mathbf{c}_v$.

Appendix B

Adopting the matrix notation, the residuals $\eta_{j,r}$ can be expressed as

$$\eta_{j,r} = \mathbf{l}_j'\left(\tilde{\mathbf{Y}}_r + \mathbf{u}_r\right) - \pi_j\boldsymbol{\delta}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\left(\tilde{\mathbf{Y}}_r + \mathbf{u}_r\right), \tag{B.1}$$

where $\mathbf{L} = \{\tilde{L}^B_{j,i}\}$ is the $M^A \times N^B$ matrix of standardized links, and $\tilde{\mathbf{Y}}_r = \{y_{i,r}\}$ and $\mathbf{u}_r = \{u_{i,r}\}$ denote respectively the N^B vectors with the values of the predictions and of the residuals of the r^{th} variable of interest being $\mathbf{l}_j'$ is the j^{th} row of the matrix $\mathbf{L}$. Therefore, the model expected values of the squared terms is given by:

$$\begin{aligned} E_{M_r}\left(\eta^2_{j,r}\right) &= \tilde{\mathbf{Y}}_r'\mathbf{l}_j\mathbf{l}_j'\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{l}_j\mathbf{l}_j'\boldsymbol{\sigma}_r \\ &\quad - 2\pi_j\tilde{\mathbf{Y}}_r'\mathbf{d}_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r - 2\pi_j\boldsymbol{\sigma}_r'\mathbf{d}_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\boldsymbol{\sigma}_r \\ &\quad + \pi_j^2\tilde{\mathbf{Y}}_r'\mathbf{L}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r \\ &\quad + \pi_j^2\boldsymbol{\sigma}_r'\mathbf{L}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}_j'\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\boldsymbol{\sigma}_r \end{aligned} \tag{B.2}$$

where $\boldsymbol{\sigma}_r = \{\sigma_{j,r}\}$ is the N^B column vector of model standard errors of the y_r variables. Following the above notation, we have:

$$E_{M_v}\left[V\left(\hat{\bar{Y}}_r^A \,\middle|\, \mathbf{m}^A\right)\right] = \left[M^A/(M^A - H)\right]\left[\tilde{\mathbf{Y}}_r'\mathbf{L}'\boldsymbol{\Pi}^{-1}\mathbf{L}\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{L}'\boldsymbol{\Pi}^{-1}\mathbf{L}\boldsymbol{\sigma}_r - \tilde{\mathbf{Y}}_r'\mathbf{L}'\mathbf{L}\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}_r'\mathbf{L}'\mathbf{L}\boldsymbol{\sigma}_r + \text{AAV}_{3,r}\right].$$

Letting $\mathbf{a}_r = \mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{L}\tilde{\mathbf{Y}}_r$, $\mathbf{b}_r = \boldsymbol{\sigma}_r'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})\,\boldsymbol{\sigma}_r$, $\mathbf{c}_r = \boldsymbol{\sigma}_r'\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I} - \boldsymbol{\Pi})'(\mathbf{I} - \boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{D}'\boldsymbol{\sigma}_r$, we have $\text{AAV}_{3,r} = \mathbf{a}_r'(\mathbf{I} - \boldsymbol{\Pi})\left(2\mathbf{L}\tilde{\mathbf{Y}}_r - \boldsymbol{\Pi}\mathbf{a}_r\right) + \mathbf{1}'(\mathbf{I} - \boldsymbol{\Pi})\left(2\mathbf{b}_r - \boldsymbol{\Pi}\mathbf{c}_r\right)$.

Finally, we note the following

$$\begin{aligned} E_{M_r}\left(z^2_{j,r}\right) &= \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\left(\tilde{y}^2_{i,r} + \sigma^2_{i,r}\right) + \sum_{i=1}^{N^B}\tilde{L}^B_{j,i}\,\tilde{y}_{i,r}\sum_{i'\neq i}\tilde{L}^B_{j,i'}\,\tilde{y}_{i',r} \\ &= \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\left(\tilde{y}^2_{i,r} + \sigma^2_{i,r}\right) + \sum_{i=1}^{N^B}\tilde{L}^B_{j,i}\,\tilde{y}_{i,r}\left(\tilde{z}_{j,r} - \tilde{L}^B_{j,i}\,\tilde{y}_{i,r}\right) \\ &= \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\left(\tilde{y}^2_{i,r} + \sigma^2_{i,r}\right) + \tilde{z}^2_{j,r} - \sum_{i=1}^{N^B}\left(\tilde{L}^B_{j,i}\right)^2\tilde{y}^2_{i,r} \\ &= \tilde{z}^2_{j,r} + \sigma^2_{j,zr}. \end{aligned}$$

Appendix C

Starting from (B.2), we have:

$$\begin{aligned} E_{M_l}E_{M_r}\left(\eta^2_{j,r}\right) &= \tilde{\mathbf{Y}}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{l}'_j\right)\tilde{\mathbf{Y}}_r + \boldsymbol{\sigma}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{l}'_j\right)\boldsymbol{\sigma}_r \\ &\quad - 2\pi_j \tilde{\mathbf{Y}}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\tilde{\mathbf{Y}}_r \\ &\quad - 2\pi_j \boldsymbol{\sigma}'_r E_{M_l}\left(\mathbf{l}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\boldsymbol{\sigma}_r \\ &\quad + \pi_j^2 \tilde{\mathbf{Y}}'_r E_{M_l}\left(\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\tilde{\mathbf{Y}}_r \\ &\quad + \pi_j^2 \boldsymbol{\sigma}'_r E_{M_l}\left(\mathbf{L}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{D}\boldsymbol{\Delta}^{-1}\mathbf{d}_j\mathbf{d}'_j\boldsymbol{\Delta}^{-1}\mathbf{D}'(\mathbf{I}-\boldsymbol{\Pi})\,\mathbf{L}\right)\boldsymbol{\sigma}_r. \end{aligned} \tag{C.1}$$

The above expected value can be easily derived based on the following general result. Let $\mathbf{A} = \left\{a_{j,j'}\right\}$ be a generic $M^A \times M^A$ matrix. The generic element $g_{i,i'}$ in the position i, i' of the squared $N^B \times N^B$ matrix $\mathbf{L}'\mathbf{A}\mathbf{L} = \left\{g_{i,i'}\right\}$ is given by $g_{i,i'} = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \tilde{L}^B_{j,i}\,\tilde{L}^B_{j',i'}\,a_{j,j'}$.

We have $E_{M_l}\left(g_{i,i'}\right) = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'}\,a_{j,j'} + \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \mathrm{Cov}_{M_l}(\tilde{L}^B_{j,i}, \tilde{L}^B_{j',i'})\;a_{j,j'}$.

Taylor's series first order approximation of $\tilde{\Lambda}_{j,i}$ is given by $\tilde{L}^B_{j,i} \cong \frac{1}{\Lambda_i}\left(L^B_{j,i} - \tilde{\Lambda}^B_{j,i}\,L^B_i\right)$. Therefore, we have

$$\begin{aligned} \mathrm{Cov}_{M_l}(\tilde{L}^B_{j,i}, \tilde{L}^B_{j',i'}) &\cong \frac{1}{\Lambda^B_i}\frac{1}{\Lambda^B_{i'}}\mathrm{Cov}_{M_l}\left[\left(L^B_{j,i} - \tilde{\Lambda}^B_{j,i}L^B_i\right), \left(L^B_{j',i'} - \tilde{\Lambda}^B_{j',i'}L^B_{i'}\right)\right] \\ &= \begin{cases} \left[V_{M_l}(L^B_{j,i})(1-2\tilde{\Lambda}^B_{j,i}) + (\tilde{\Lambda}^B_{j,i})^2 V_{M_l}(L^B_i)\right]\Big/(\Lambda^B_i)^2 & \text{for } j=j' \text{ and } i=i' \\ \left[\tilde{\Lambda}^B_{j,i}\tilde{\Lambda}^B_{j,i'}V_{M_l}(L^B_i) - \tilde{\Lambda}^B_{j,i}V_{M_l}(L^B_{j,i}) - \tilde{\Lambda}^B_{j,i'}V_{M_l}(L^B_{j',i})\right]\Big/(\Lambda^B_i)^2 & \text{for } j\neq j' \text{ and } i=i' \\ 0 & \text{for } j=j' \text{ and } i\neq i' \\ 0 & \text{for } j\neq j' \text{ and } i\neq i' \end{cases} \end{aligned} \tag{C.2}$$

where

$$V_{M_l}\left(L^B_{j,i}\right) = \sum_{k=1}^{M^b_i}\lambda_{j,ik}\left(1-\lambda_{j,ik}\right),\; V_{M_l}(L^B_i) = \sum_{j=1}^{M^A} V_{M_l}\left(L^B_{j,i}\right). \tag{C.3}$$

Equation (C.2) is derived from the following result. For $i = i'$ and $j = j'$, we obtain

$$\begin{aligned} \mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i}, \tilde{L}^B_{j,i}\right) &= V_{M_l}\left(\tilde{L}^B_{j,i}\right) \\ &\cong \frac{1}{\Lambda^2_i}\left[V_{M_l}\left(L^B_{j,i}\right) + \left(\tilde{\Lambda}^B_{j,i}\right)^2 V_{M_l}(L^B_i) - 2\tilde{\Lambda}^B_{j,i}\,\mathrm{Cov}_{M_l}\left(L^B_{j,i}, L^B_i\right)\right] \\ &= \frac{1}{\Lambda^2_i}\left[V_{M_l}\left(L^B_{j,i}\right)\left(1-2\tilde{\Lambda}^B_{j,i}\right) + \left(\tilde{\Lambda}^B_{j,i}\right)^2 V_{M_l}(L^B_i)\right] \end{aligned} \tag{C.4}$$

where $\mathrm{Cov}_{M_l}\left(L^B_{j,i}, L^B_i\right) = V_{M_l}\left(L^B_{j,i}\right)$. For $i = i'$ and $j \neq j'$, we obtain

$$\begin{aligned}\mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\, \tilde{L}^B_{j',i}\right) &\cong \frac{1}{\Lambda_i^2}\mathrm{Cov}_{M_l}\left[\left(L^B_{j,i} - \tilde{\Lambda}^B_{j,i}\, L_i\right), \left(L^B_{j',i} - \tilde{\Lambda}^B_{j',i} L_i\right)\right] \\ &= \frac{1}{\Lambda_i^2} V_{M_l}\left[L^B_{j,i}L^B_{j',i} - \tilde{\Lambda}^B_{j',i}L^B_{j,i}L_i - \tilde{\Lambda}^B_{j,i}L^B_{j',i}L_i + \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i}\,L_i^2\right] \\ &= \frac{1}{\Lambda_i^2}\left[\tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i} V_{M_l}\left(L^B_i\right) - \tilde{\Lambda}^B_{j,i}\, V_{M_l}\left(L^B_{j,i}\right) - \tilde{\Lambda}^B_{j',i}\, V_{M_l}\left(L^B_{j',i}\right)\right]. \qquad \text{(C.5)}\end{aligned}$$

Let $\mathbf{a} = \{a_j\}$ be a generic M^A vector. The generic element ${}^j g_{i,i'}$ in the position i, i' of the squared $N^B \times N^B$ matrix $\mathbf{l}_j\mathbf{a}'\mathbf{L} = \left\{{}^j g_{i,i'}\right\}$ is given by ${}^j g_{i,i'} = \sum_{j'=1}^{M^A} \tilde{L}^B_{j,i}\,\tilde{L}^B_{j',i'}\,a_{j'}$, where $E_{M_l}\left({}^j g_{i,i'}\right) = \sum_{j'=1}^{M^A} \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'}\,a_{j'} + \sum_{j'=1}^{M^A} \mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\, \tilde{L}^B_{j',i'}\right)\, a_{j'}$.

Finally, denote by $\left\{{}^{jj} g_{i,i'}\right\}$ the generic element in the i, i'^{th} position of the matrix $\mathbf{l}_j\mathbf{l}'_j$. Its generic expected value is given by $E_{M_l}\left({}^{jj} g_{i,i'}\right) = \tilde{\Lambda}^B_{j,i}\,\tilde{\Lambda}^B_{j',i'} + \mathrm{Cov}_{M_l}\left(\tilde{L}^B_{j,i},\, \tilde{L}^B_{j',i'}\right)$.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1989001/article/14578-eng.pdf.

Boyd, S., and Vanderberg, L. (2004). *Convex Optimization*. Cambridge University Press.

Brewer, K.R.W., and Gregoire, T.G. (2009). Introduction to survey sampling. In *Handbook of Statistics – Sample Surveys: Design, Methods and Applications*, (Eds., D. Pfeffermann and C.R. Rao), Elsevier B.V. 29A, 9-37.

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 1, 23-29. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012001/article/11682-eng.pdf.

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. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Falorsi, P.D., and Righi, P. (2015). Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. *Survey Methodology*, 41, 1, 215-236. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015001/article/14149-eng.pdf.

FAO (2014). *Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02-2014. http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf.

FAO (2015). *Guidelines on Integrated Survey Framework.* GUIDELINES & HANDBOOKS http://gsars.org/en/guidelines-for-the-integrated-survey-framework/. Accessed on August 2016.

Khan, M.G.M., Mati, T. and Ahsan, M.J. (2010). An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics*, 26, 695-708.

Kokan, A., and Khan, S. (1967). Optimum allocation in multivariate surveys: An analytical solution. *Journal of the Royal Statistical Society, Series B*, 29, 115-125.

Lavallée, P. (2002). Le sondage indirect, ou la méthode du partage des poids. Éditions de l'Université de Bruxelles (Belgique) et Éditions Ellipses (France), 215 pages.

Lavallée, P. (2007). *Indirect Sampling.* New York: Springer.

Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 2, 155-169. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001002/article/6092-eng.pdf.

Lavallée, P., and Labelle-Blanchet, S. (2013). Indirect sampling applied to skewed populations. *Survey Methodology*, 39, 1, 183-215. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013001/article/11829-eng.pdf.

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

Steel, D.G., and Clark, R.G. (2014). Potential gains from using unit level cost information in a model-assisted framework. *Survey Methodology*, 40, 2, 231-242. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14110-eng.pdf.

Wallgren, A., and Wallgren, B. (2014). *Register-Based Statistics: Administrative Data for Statistical Purposes*. New York: John Wiley & Sons, Inc. Chichester, UK. ISBN: ISBN 978-1-119-94213-9.

Winkler, W.E. (2001). *Multi-Way Survey Stratification and Sampling*. Research Report Series, Statistics #2001-01. Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233.

Xu, X., and Lavallée, P. (2009). Treatments for link nonresponse in indirect sampling. *Survey Methodology*, 35, 2, 153-164. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11038-eng.pdf.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 513-531
Statistics Canada, Catalogue No. 12-001-X

A grouping genetic algorithm for joint stratification and sample allocation designs

Mervyn O'Luing, Steven Prestwich and S. Armagan Tarim[1]

Abstract

Finding the optimal stratification and sample size in univariate and multivariate sample design is hard when the population frame is large. There are alternative ways of modelling and solving this problem, and one of the most natural uses genetic algorithms (GA) combined with the Bethel-Chromy evaluation algorithm. The GA iteratively searches for the minimum sample size necessary to meet precision constraints in partitionings of atomic strata created by the Cartesian product of auxiliary variables. We point out a drawback with classical GAs when applied to the grouping problem, and propose a new GA approach using "grouping" genetic operators instead of traditional operators. Experiments show a significant improvement in solution quality for similar computational effort.

Key Words: Grouping genetic algorithm; Optimal stratification; Sample allocation; R software.

1 Introduction

In this paper we address the optimization problem of jointly determining stratification and sample allocation for univariate and mulitivariate scenarios. To serve this purpose, we refer to (Ballin and Barcaroli, 2013). In principle the optimal stratification (i.e., that which yields the smallest sample size) can be found by testing all possible partitionings of *atomic strata*, but the number of possible partitionings grows exponentially with the number of atomic strata.

An efficient search algorithm is necessary to avoid evaluating each possible partitioning. Genetic algorithms (GAs) often converge quickly to optimal or near optimal solutions, and are particularly good at navigating rugged search spaces containing many local minima. The Bethel-Chromy algorithm combines similar algorithms from (Bethel, 1985, 1989) and (Chromy, 1987) and is suitable for univariate and mulitivariate cases. It uses lagrangian multipliers to find the minimum sample size that meets precision constraints for a given stratification. (Ballin and Barcaroli, 2013) combine a GA with this algorithm to search for the minimum sample size. It is used to evaluate each partitioning created by the GA. A full description of the methodology and problem statement is found in (Ballin and Barcaroli, 2013). However, they use a classical GA which is known to be unsuitable for partitioning problems.

In this paper we propose to apply genetic operators to the GA that are better suited to this application. It is an example of the class of evolutionary algorithms called *Grouping Genetic Algorithms* (GGAs). The GA has been updated following this work (Barcaroli, 2019). Section 2 motivates the work and introduces GGAs. Section 2.3 describes our GGA for the problem. Section 3 compares the original GA with our GGA on publicly-available test data. Section 4 describes a version of our GGA with enhanced performance, using a fast C++ implementation of the *bethel.r* function which we integrated into R using the Rcpp package. Section 5 concludes the paper.

1. Mervyn O'Luing and Steven Prestwich, Insight Centre for Data Analytics, Department of Computer Science, University College Cork, Ireland. E-mail: mervyn.oluing@insight-centre.org and steven.prestwich@insight-centre.org; S. Armagan Tarim, Cork University Business School, University College Cork, Ireland. E-mail: armagan.tarim@ucc.ie.

2 Classical vs grouping genetic algorithms

In this section we discuss "classical" and "grouping" GAs, and explain why the latter are more appropriate for our problem.

2.1 Classical genetic algorithms

GAs are a nature-inspired class of optimisation algorithms, modelled on the ability of organisms to solve the complex problem of adaptation to life on Earth. The variables of an optimisation problem are called *genes* and their values *alleles*. A candidate solution is a list of alleles called a *chromosome*. A set of chromosomes is usually called a *population*, so to avoid confusion with the target population we shall use *chromosome population* when referring to GAs. The objective function (which is maximised by convention) is called the chromosome's *fitness*. The search for fit chromosomes (solutions with high objective) uses two *genetic operators*: small random changes called *mutation*, equivalent to small local moves in a hill-climbing algorithm; and large changes called *crossover* in which the genes of two *parent chromosomes* are *recombined*. One well-known recombination operator is *single-point crossover*: choose two *parent* chromosomes with alleles

$$a_1, \ldots, a_N \qquad b_1, \ldots, b_N,$$

select a random integer i (the *crossover point*) such that $1 \leq i < N$, and generate two new *offspring* chromosomes

$$a_1, \ldots, a_i, b_{i+1}, \ldots, b_N \qquad b_1, \ldots, b_i, a_{i+1}, \ldots, a_N.$$

These might be further subjected to random *mutation*, in which a few alleles are changed, before placing them back into the chromosome population. There are a variety of methods for selecting parents and replacing existing chromosomes. In *generational* GAs the entire chromosome population is replaced by offspring, and parents are often selected randomly but with a bias toward fitter chromosomes; while in *steady-state* GAs only one offspring is generated in each GA iteration, and usually replaces the least-fit chromosome in the chromosome population. GAs often give more robust results than search algorithms based on hill-climbing, because of their use of recombination. They have found many applications since their introduction in 1975 by John Holland.

The original GA which is represented in the R (R Core Team, 2015) package *SamplingStrata* (Barcaroli, 2014), is an elitist generational GA in which the atomic strata L are considered to be elements of a set (or genes) for a standard crossover strategy. In each iteration the best solutions (the *elite*) are carried over to the next generation. Each gene represents a variable in the problem. We refer to this as a classical GA because a classical problem representation and genetic operators are used, as described below.

Dividing atomic strata into disjoint groups is an example of a *grouping* problem, related to *cutting*, *packing* and *partitioning* problems. The motivation for our work is that classical GAs are known to perform

poorly on grouping problems. The reason is that the chromosomal representation of a grouping contains a great deal of *symmetry* (or *redundancy*): permuting the group names yields an equivalent grouping, so each grouping has multiple representations. Symmetry has a damaging effect on GAs because recombining similar parent groupings might yield a very different offspring grouping, violating the basic GA principle that parents should tend to produce offspring with similar fitness. In extreme cases, a classical GA might perform even worse than a completely random search. We provide two examples to illustrate the problem.

To illustrate the problem with symmetry in our first example the parents represent the same grouping in different ways. Note that to increase readability, letters A - F are used as alleles instead of integers in the presentation here. Consider the following two chromosomes:

	groups represented					
chromosome	A	B	C	D	E	F
ABCDEF	{1}	{2}	{3}	{4}	{5}	{6}
FEDCBA	{6}	{5}	{4}	{3}	{2}	{1}

which both represent the grouping $\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$. Now suppose we apply single-point crossover to obtain two new offspring chromosomes from these parents. Arbitrarily choosing the center of the chromosomes as the crossover point, we obtain offspring:

	groups represented					
chromosome	A	B	C	D	E	F
ABCCBA	{1, 6}	{2, 5}	{3, 4}	∅	∅	∅
FEDDEF	∅	∅	∅	{3, 4}	{2, 5}	{1, 6}

which both represent the completely unrelated grouping $\{\{1, 6\}, \{2, 5\}, \{3, 4\}\}$: no groups at all are passed from the parents to the offspring. Hence the offspring and parent fitnesses can be completely unrelated to each other, which reduces the GA to near-random search. As another example, consider the following two classical chromosomes:

	groups represented					
chromosome	A	B	C	D	E	F
AECFEC	{1}	∅	{3, 6}	∅	{2, 5}	{4}
DFFDAA	{5, 6}	∅	∅	{1, 4}	∅	{2, 3}

which in turn represent the different groupings $\{\{1\}, \{3, 6\}, \{2, 5\}, \{4\}\}$ and $\{\{5, 6\}, \{1, 4\}, \{2, 3\}\}$. Using the same crossover strategy we obtain offspring:

	groups represented					
chromosome	A	B	C	D	E	F
AECDAA	{1, 5, 6}	∅	{3}	{4}	{2}	∅
DFFFEC	∅	∅	{6}	{1}	{5}	{2, 3, 4}

representing the groupings $\{\{1, 5, 6\}, \{3\}, \{4\}, \{2\}\}$ and $\{\{6\}, \{1\}, \{5\}, \{2, 3, 4\}\}$. Note that these offspring have very little in common with their parents, as the only preserved groups are $\{1\}$ and $\{4\}$.

2.2 Grouping genetic algorithms

The symmetry problem can be tackled by designing more complex genetic representations and operators (Galinier and Hao, 1999) or by clustering techniques (Pelikan and Goldberg, 2000). The risk of clustering is that genetic diversity may be lost if the clusters are too tight, leading to search stagnation (Prügel-Bennett, 2004). Instead we follow the former approach by designing a GGA (Falkenauer, 1998), which have been shown to perform far better than classical GAs on grouping problems.

GGAs are designed specifically to solve grouping problems and have found many applications, including WiFi network deployment (Agustín-Blas, Salcedo-Sanz, Vidales, Urueta and Portilla-Figueras, 2011), wireless network design (Brown and Vroblefski, 2004), steel plate cutting (Hung, Sumichrast and Brown, 2003), production plant layout (De Lit, Falkenauer and Delchambre, 2000) and social network analysis (James, Brown and Ragsdale, 2010). They may use the same heuristics as other GAs (parent selection, offspring replacement, etc) but they use different genetic encoding and operators: that is, how they map a problem to chromosomes and how they perform recombination and mutation. We shall illustrate these differences on the above examples.

GGAs represent a grouping as an ordered list of subsets, omitting empty sets. The parents in the second example of Section 2.1 might be represented in this way:

$$\langle\{1\}, \{3, 6\}, \{2, 5\}, \{4\}\rangle \quad \langle\{5, 6\}, \{1, 4\}, \{2, 3\}\rangle.$$

GGA mutation is simple: an item is moved from one group to another. However, the GGA recombination operator is more complicated. Choose a *crossing section* in each parent, for example $\langle\{1\}, \{3, 6\}\rangle$ from the 1^{st} parent and $\langle\{1, 4\}\rangle$ from the 2^{nd} parent. Then *inject* the 1^{st} crossing section into the 2^{nd} parent at a random point, and vice-versa:

$$\langle\{1\}, \{3, 6\}, \underline{\{1, 4\}}, \{2, 5\}, \{4\}\rangle \quad \langle\{5, 6\}, \{1, 4\}, \underline{\{1\}}, \underline{\{3, 6\}}, \{2, 3\}\rangle.$$

Next remove any repeated objects that were already in the receiving parent:

$$\langle\varnothing, \{3, 6\}, \{1, 4\}, \{2, 5\}, \varnothing\rangle \quad \langle\{5\}, \{4\}, \{1\}, \{3, 6\}, \{2\}\rangle.$$

Finally remove any empty sets:

$$\langle\{3, 6\}, \{1, 4\}, \{2, 5\}\rangle \quad \langle\{5\}, \{4\}, \{1\}, \{3, 6\}, \{2\}\rangle.$$

These are the offspring. Clearly, both offspring have much in common with both parents, as 5 of the 7 parent groups survive in the offspring: $\{1\}$, $\{4\}$, $\{1, 4\}$, $\{2, 5\}$ and $\{3, 6\}$. In the first example of Section 2.1 it is easily verified that both offspring represent the same grouping as the parents, as one would expect. This property of the GGA injection-based recombination makes it much more likely that offspring have similar fitness to parents, which in turn helps the GGA to iteratively improve the chromosome population.

It might be noticed that the GGA problem representation still contains symmetry: any grouping still has multiple representations, obtained by permuting the subsets in the ordered list. But the genetic operators are almost independent of this ordering so it is almost irrelevant. The only effect of the ordering is to limit the set of possible injections: in the second example of Section 2.1 we cannot inject a non-existent crossing section for example such as $\langle\{1\}, \{4\}\rangle$ from parent 1 because those two groups are not adjacent. This limit is removed by an additional genetic operator called *inversion* which selects a section of the chromosome and reverses it. For example

$$\langle\{1\}, \{2\}, \underline{\{3, 6\}}, \underline{\{4\}}, \underline{\{5\}}\rangle \rightarrow \langle\{1\}, \{2\}, \underline{\{5\}}, \underline{\{4\}}, \underline{\{3, 6\}}\rangle.$$

This does not change the grouping represented by the chromosome, but reordering the groups in the chromosome makes all injections possible.

Injection, mutation and inversion are the common operators used in GGAs, but there is no canonical algorithm. Instead GGAs tend to be tailored for specific applications, and in principle any GA can be adapted to grouping problems by using grouping operators. In Section 2.3 we design a GGA for our problem.

2.2.1 Note on implementation

For the sake of clarity the descriptions in Section 2.2 omit implementation details, for example the fact that GGA chromosomes are usually implemented in two parts (or sometimes more). The first part uses a classical representation as above, while the second part lists the nonempty groups as a permutation. Injection occurs on the second parts of parent chromosomes and some renaming of groups is necessary.

Typically we decide in advance the number of iterations which we wish to run the algorithm for. This should be enough to give the GGA a chance to converge on the optimum solution after the mutation and inversion probabilities have been applied. If, however, the optimum solution is known beforehand the algorithm can be set to stop at this point.

The number of iterations is usually decided with experience of using the GGA on similar target and auxiliary variables for similar datasets, or with the existing dataset and target and auxiliary variables. It may require a number of experiments using the GGA (or GA) before the number of iterations needed to reach convergence can be estimated. In fact there is a possibility that either the GGA or GA would appear to have reached convergence after a set number of iterations, but instead have become trapped in a local minimum. It may be useful to increase the number of iterations and try alternative mutation probabilities in order to be certain that it has converged on a global minimum.

This implies a number of trial runs before finally deciding the parameters under which to run the algorithms. Therefore the fact the GGA has been shown to attain convergence quicker than the GA is likely to compound the improvement in total processing time. In the experiments described below we keep the number of iterations small as we want to demonstrate the ability of the GGA to converge on a solution within that number of iterations.

We use either the mutation settings specified in the examples provided by (Ballin and Barcaroli, 2013) or the default mutation settings in (Barcaroli, 2014). We apply grouping genetic operators and inversion to the GA designed by (Ballin and Barcaroli, 2013): it is the grouping genetic operators that make it a GGA. Thus we compare the performance between the different GA and GGA genetic operators rather than experiment with parameters such as varying the number of iterations, chromosome population size, mutation probability, or elitism rate.

The mutation probability can be selected in advance by the user. Typically, the probability of mutation should be such that it increases the chance of the GGA leaving a local minimum, but not disrupt the natural evolution of chromosomes from one generation to the next. On the other hand we have fixed the inversion probability at 0.01, because this is enough to maintain diversity.

The size of the chromosome population can be decided by trial and error. It is advisable to consider the evaluation time of each chromosome when setting the size: if there are too many chromosomes in the set, it might take an extra long time to move from one iteration to the next, and we found that the *bethel.r* algorithm (i.e., the Bethel-Chromy evaluation algorithm in (Barcaroli, 2014)) takes several seconds to evaluate even one chromosome for the larger datasets we used in this paper (we discuss this further in Section 4).

For further details on the implementation of GGAs (e.g., elitism rate) we refer the reader to papers such as (Falkenauer, 1998).

2.3 Application to the joint stratification and sample allocation problem

As mentioned above our GGA is based on the GA described in (Ballin and Barcaroli, 2013) and represented in *R* in the *SamplingStrata* package (Barcaroli, 2014), but with grouping operators and chromosomes instead of the classical versions. This change is the only novelty of our algorithm (except for the optimisation described in Section 4) but its effect on performance is large. We inserted the GGA into a modified version of the function called *rbga.r* from the *genalg* R package (Willighagen, 2005). It is designed to work with the other functions in *SamplingStrata*, and is applied to the joint stratification and optimum sample size problem. The GGA is summarised in Figure 2.1.

Following the problem statement in (Ballin and Barcaroli, 2013) we summarise the cost function as follows:

$$C(n_1, \ldots, n_H) = C_0 + \sum_{h=1}^{H} C_h n_h,$$

where C_0 is the fixed cost and C_h is the average cost of interviewing one unit in stratum h and n_h is the number of units, or sample, allocated to stratum h. In our analysis C_0 is set to 0, and C_h is set to 1. The expectation of the estimator of the "g^{th}" population total is:

$$E\left(\hat{T}_g\right) = \sum_{h=1}^{H} N_h \bar{Y}_{h,g} \quad (g = 1, \ldots, G),$$

where $\bar{Y}_{h,g}$ is the mean of the G different target variables Y in each stratum h. The variance of the estimator is given by:

$$\text{VAR}\left(\hat{T}_g\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{h,g}^2}{n_h} \quad (g = 1, \ldots, G). \tag{2.1}$$

The upper limit of variance or precision U_g is expressed as a coefficient of variation CV for each $\hat{T}_g$:

$$\text{CV}\left(\hat{T}_g\right) = \frac{\sqrt{\text{VAR}\left(\hat{T}_g\right)}}{E\left(\hat{T}_g\right)} \leq U_g. \tag{2.2}$$

The problem can be summarised as follows:

$$\min n = \sum_{h=1}^{H} n_h$$
$$\text{CV}\left(\hat{T}_g\right) \leq U_g.$$

Grouping Genetic Algorithm (GGA)

Step 1: Initialization

(a) Randomly generate a chromosome population of size N_P.

Step 2: Selection part 1

(a) Rank chromosomes based on sample size.

(b) Save best E chromosomes for the next generation.

Step 3: Inversion

With probability 0.01 invert groups in the N_P chromosomes.

Step 4: Selection part 2

For each of the remaining $N_P - E$ chromosomes in the new generation:

(a) Draw parents 1 and 2 from the aforementioned N_P chromosomes (higher ranked chromosomes have a higher probability of being selected).

(b) Perform crossover as explained in Section 2.2.

(c) Remove empty groups.

(d) Renumber groups.

Step 5: Mutation

Mutate integers in $N_P - E$ chromosomes at a selected probability.

Step 6: if #iterations<maximum

(optional: and sample size $>$ desired value) go to step 2.

Figure 2.1 Pseudocode for our GGA.

3 Comparing the genetic algorithms

We now run a number of comparisons between the original GA and our GGA using publicly available datasets. Unless otherwise stated, for all the cases presented below, we adopt the following parameter setting for both genetic algorithms, where $N_P = 20, U_g \equiv 0.05,$ the elitism rate is 0.2, and the mutation probability is 0.05.

3.1 A comparison for the iris dataset

(Ballin and Barcaroli, 2013) use the iris dataset (Anderson, 1935; Fisher, 1936; R Core Team, 2015) to demonstrate that the GA they propose can find the optimum stratification i.e., the stratification or grouping of atomic strata which supplies the minimum sample size. The iris dataset is small and is widely available. It has 150 observations for 5 variables Sepal Length, Sepal Width, Petal Length, Petal Width and Species.

Species is a categorical variable which has three levels, setosa, versicolor and virginica, each of which have 50 observations. The remaining four variables are continuous measurements for length and width in centimetres. (Ballin and Barcaroli, 2013) select Petal Length and Petal Width as variables of interest, i.e., target variables. They select Sepal Length and Species as two auxiliary variables.

They convert Sepal Length to a categorical variable using a k-means algorithm (Hartigan and Wong, 1979) to define three clusters (i.e., 4.3 to less than 5.5, 5.5 to less than 6.5, 6.5 to 7.9). The cross product of the categorical version of Sepal Length with Species creates 9 atomic strata. However, one atomic stratum is empty because there are no corresponding values in Petal Length and Petal Width. Therefore there are 8 usable atomic strata for this example.

Table 3.1
Reproduction of table of atomic strata for estimating the minimum sample size for the target variables of iris dataset as found in (Ballin and Barcaroli, 2013), page 379

Stratum	N	M1	M2	S1	S2	X1	X2	DOMAIN
[4.3; 5.5] (1)*setosa	45	1.466667	0.244444	0.17127	0.106574	[4.3; 5.5] (1)	setosa	1
[4.3; 5.5] (1)*versicolor	6	3.583333	1.166667	0.491313	0.205481	[4.3; 5.5] (1)	versicolor	1
[4.3; 5.5] (1)*virginica	1	4.5	1.7	0	0	[4.3; 5.5] (1)	virginica	1
[5.5; 6.5] (2)*setosa	5	1.42	0.26	0.172047	0.08	[5.5; 6.5] (2)	setosa	1
[5.5; 6.5] (2)*versicolor	35	4.268571	1.32	0.367051	0.189435	[5.5; 6.5] (2)	versicolor	1
[5.5; 6.5] (2)*virginica	23	5.230435	1.947826	0.318194	0.28873	[5.5; 6.5] (2)	virginica	1
[6.5; 7.9] (3)*versicolor	9	4.677778	1.455556	0.193091	0.106574	[6.5; 7.9] (3)	versicolor	1
[6.5; 7.9] (3)*virginica	26	5.876923	2.107692	0.494825	0.228579	[6.5; 7.9] (3)	virginica	1

The initial atomic strata are reproduced in Table 3.1 where M_g refers to the means for the corresponding Y_g values in each atomic stratum l_k; S_g refers to the corresponding stratum population standard deviations. There are 4,140 possible partitionings of the 8 atomic strata. Consequently, it is possible to test within a reasonable amount of time the sample size for the entire search space using the *bethel.r* function. This has already been done (Ballin and Barcaroli, 2013) and the minimum sample size is known to be 11.

This test can be used to determine whether the new GA correctly finds the minimum sample size without exploring the entire search space. We use $N_p = 10$ in this case. For this test the *bethel.r* function will search for the minimum sample size, in integers rather than real numbers. The chromosomes will then be ranked by sample size in ascending order. Accordingly the elite chromosomes are taken into the next iteration and the remaining chromosomes are generated using the recombination method for each algorithm.

We will compare the number of chromosomes generated to find the optimal stratification in the two algorithms as well as the number of iterations. Our anticipation is that the GGA should be more efficient, and thus typically find the optimal solution in fewer iterations than the GA.

The maximum number of iterations is set to 200, because using (Ballin and Barcaroli, 2013) as a guide we anticipate that both algorithms will find the correct solution in fewer iterations than this. Thus we have added a piece of code to both algorithms such that they stop when the optimal sample size, $n = 11$, has been reached and supply the number of iterations taken to reach that point. This approach is different to that of (Ballin and Barcaroli, 2013) who report the number of times in 10 experiments the GA finds the correct solution for a given number of iterations ranging incrementally from 25 to 200. However, we feel this approach would better demonstrate that the GGA can find the correct solution in less iterations even on the small iris dataset experiment.

Table 3.2
Iris dataset experiment results for GA and GGA

	(a) GA			(b) GGA		
Number of	**Experiment**	**Iterations**	**Chromosomes**	**Experiment**	**Iterations**	**Chromosomes**
	1	14	228	1	11	180
	2	8	132	2	7	116
	3	17	276	3	6	100
	4	40	644	4	22	356
	5	31	500	5	9	148
	6	13	212	6	11	180
	7	15	244	7	8	132
	8	9	148	8	7	116
	9	15	244	9	9	148
	10	15	244	10	11	180
	11	14	228	11	3	52
	12	8	132	12	9	148
	13	17	276	13	27	436
	14	40	644	14	12	196
	15	31	500	15	16	260
	16	13	212	16	6	100
	17	15	244	17	20	324
	18	9	148	18	6	100
	19	15	244	19	7	116
	20	15	244	20	6	100
	21	16	260	21	11	180
	22	67	1,076	22	7	116
	23	19	308	23	8	132
	24	9	148	24	5	84
	25	11	180	25	7	116
	26	20	324	26	5	84
	27	32	516	27	6	100
	28	10	164	28	6	100
	29	37	596	29	9	148
	30	9	148	30	6	100

Table 3.2 provides the number of iterations (and chromosomes generated) taken to find $n = 11$ over 30 experiments for both GAs.

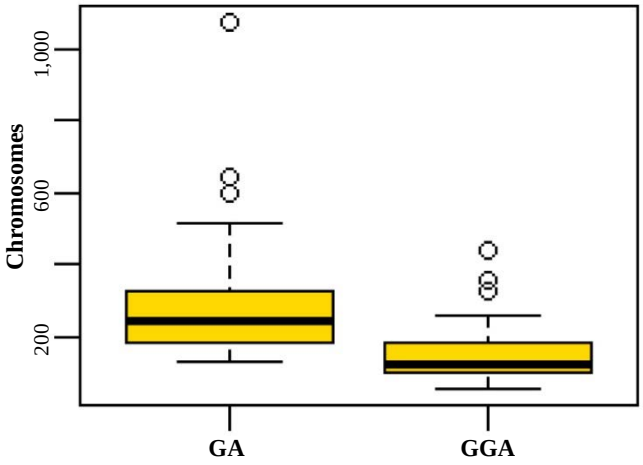


Figure 3.1 Boxplot distribution of number of Chromomomes generated to find $n = 11$ for GA and GGA after 30 experiments.

Figure 3.1 provides the distribution of the number of chromosomes generated to find the optimal solution for the GA and the GGA. The boxplots indicate that the GGA typically needs to generate fewer chromosomes to find the optimum solution.

Table 3.3
Example stratifications for the GA and GGA on the iris dataset for $n = 11$

		Y1			Y2		
	Stratum	N	Mean	SD	Mean	SD	Sample Size
GA	1	50	1.462	0.1685	0.246	0.1026	2
	2	50	4.26	0.4562	1.326	0.1911	3
	3	1	4.5	0	1.7	0	1
	4	23	5.2304	0.3112	1.9478	0.2824	3
	5	26	5.8769	0.4852	2.1077	0.2241	2
Total		150					11
GGA	1	23	5.2304	0.3112	1.9478	0.2824	3
	2	50	1.462	0.1685	0.246	0.1026	2
	3	26	5.8769	0.4852	2.1077	0.2241	2
	4	51	4.2647	0.4529	1.3333	0.1962	4
Total		150					11

Table 3.3 provides example stratifications for the GA and GGA that both provide the optimal sample size necessary to meet precision constraints. (Ballin and Barcaroli, 2013) indicate that a number of partitionings from the total of 4,140 possible partitionings provide the minimum sample size. These range in size from 3 to 5 strata. It is seen that the GGA results in fewer, less fragmented design strata. The same tendency can be observed in the latter cases.

3.2 Swiss municipality dataset

The swissminucipalities dataset provided by (Barcaroli, 2014) refers to the Swiss municipalities in 2003. Each municipality belongs to one of seven regions which are at the NUTS-2 level, i.e., equivalent to provinces. Each region contains a number of cantons, which are administrative subdivisions. There are 26 cantons in Switzerland. The data, which was sourced from the Swiss Federal Statistical Office and is included in the *sampling* and *SamplingStrata* packages, contains 2,896 observations (each observation refers to a Swiss municipality in 2003). They comprise 22 variables, details of which can be examined in (Barcaroli, 2014).

The target estimates are the totals of the population by age class in each Swiss region. In this case, the G target variables will be:

Y1: number of men and women aged between 0 and 19,

Y2: number of men and women aged between 20 and 39,

Y3: number of men and women aged between 40 and 64,

Y4: number of men and women aged 65 and over.

We consider 6 auxiliary variables, formed using the same k-means clustering method as the iris dataset example:

X1: classes of total population in the municipality. 18 categories,

X2: classes of wood area in the municipality. 3 categories,

X3: classes of area under cultivation in the municipality. 3 categories,

X4: classes of mountain pasture area in the municipality. 3 categories,

X5: classes of area with buildings in the municipality. 3 categories,

X6: classes of industrial area in the municipality. 3 categories.

There are 7 regions, which we treat as population domains of design to distinguish them from the design strata, replicating the experiment outlined in (Barcaroli, 2014). The number of non-empty atomic strata is 641 in the population. We set the minimum population size of stratum to be 2, and the maximum number of iterations to be 400. The results for Sample Size and Strata after 30 experiments each with 400 iterations are summarised in Figure 3.2 below.

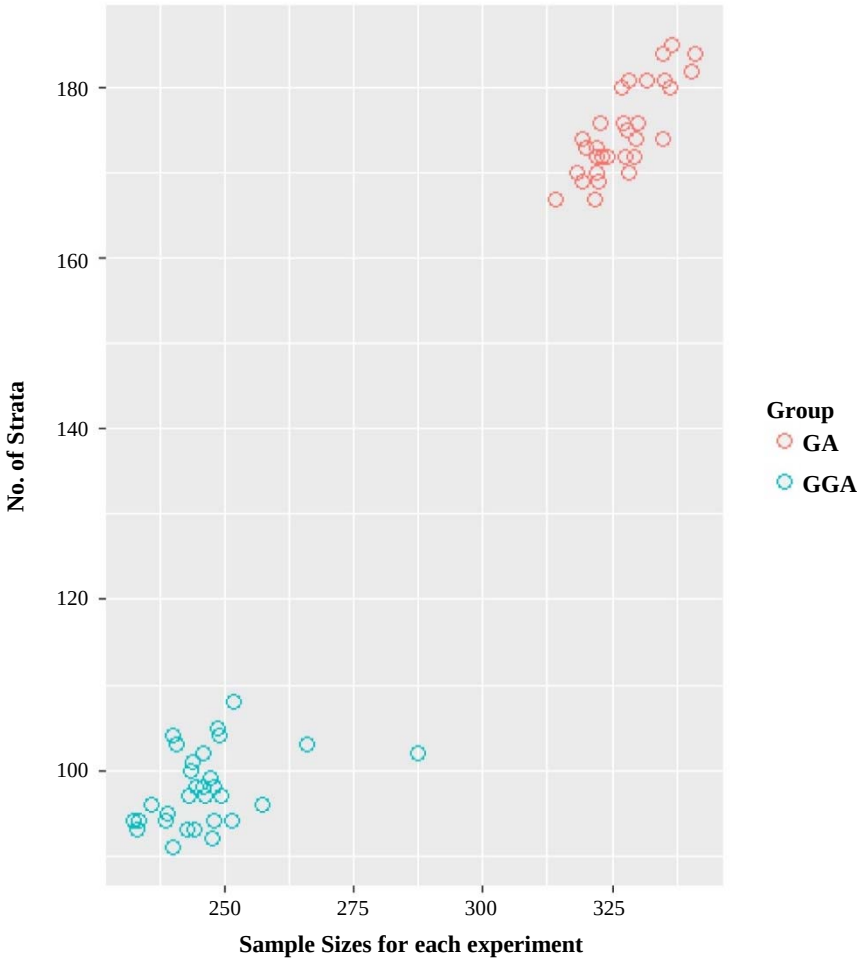


Figure 3.2 Scatterplot of Results for Strata v Sample size for GA and GGA after 30 experiments.

Figure 3.2 clearly shows that the GGA returns a smaller sample size to the GA for these settings. The median for the GGA, 246, is 25% lower than that for the GA, 328.

3.3 2015 American Community Survey Public Use Microdata

The United States has been conducting a decennial census since 1790. In the 20^{th} century censuses were split into long and short form versions. A subset of the population was required to answer the longer version of the census, with the remainder answering the shorter version. After the 2000 census the longer questionnaire became the annual American Community Survey (ACS) (US Census Bureau, 2013). The 2015 ACS Public Use Microdata Sample (PUMS) file (US Census Bureau, 2016) is a sample of actual responses to the ACS representing 1% of the US population. The PUMS file contains 1,496,678 records each of which represents a unique housing unit or group quarters. There are 235 variables. The full data dictionary is available in (US Census Bureau, 2016). We selected the following to be target variables:

1. household income (past 12 months),
2. property value,
3. selected monthly owner costs,
4. fire/hazard/flood insurance (yearly amount),

and the following auxiliary variables:

1. units in structure,
2. tenure,
3. work experience of householder and spouse,
4. work status of householder or spouse in family households,
5. house heating fuel,
6. when structure first built.

The PUMS data for which all the values are present contains 619,747 records. We use the 51 states (based on census definitions) as domains.

In the convergence plots of Figure 3.3, the black line represents the best or lowest sample size for the chromosome population in each iteration, whereas the red line represents the mean sample size for the chromosome population in each iteration.

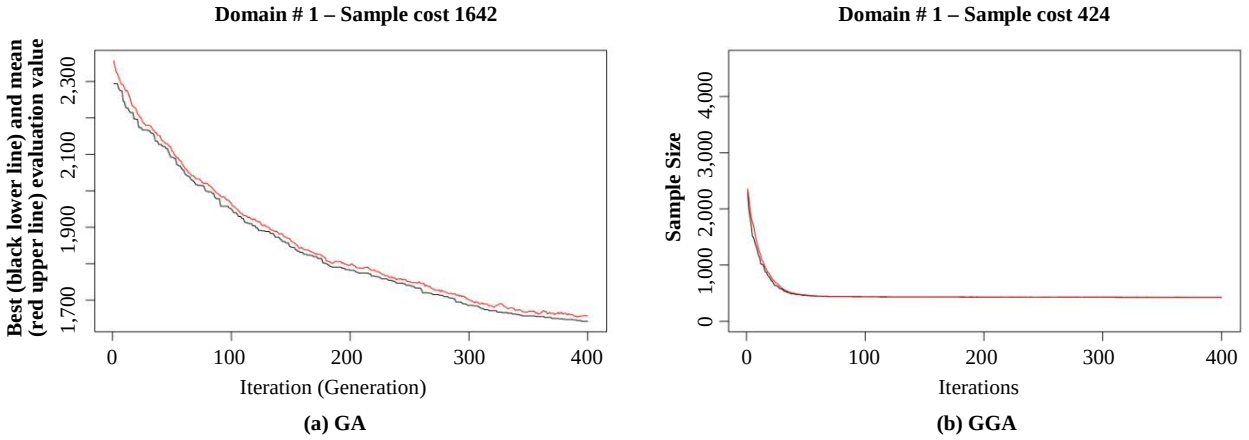


Figure 3.3 Convergence plots for Sample Size after the 1st experiment for GA and GGA. Note the different scales on the vertical axes.

The GA appears to be reducing the sample size steadily but does not appear to have reached a local minimum after 400 iterations. The GGA appears to have reached a local or global minimum very quickly.

3.4 Kaggle Data Science for Good challenge Kiva Loans data

The online crowdfunding platform kiva.org provided a dataset of loans issued to people living in poor and financially excluded circumstances around the world over a two year period for a Kaggle Data Science for Good challenge. The dataset has 671,205 unique records. We selected these target variables:

1. term in months,
2. lender count,
3. loan amount,

and the following auxiliary variables:

1. sector,
2. currency,
3. activity,
4. region,
5. partner id,

to create atomic strata. For these variables we removed any records with missing values. We then proceeded to remove any countries with less than 10 records from the sampling frame. This resulted in a sampling frame with 614,361 records. The variable country-code defines the 73 design domains in this experiment.

Table 3.4
Sample size and strata for the Kiva Loans data from the GA and the GGA after 100 iterations

GA		GGA		Reduction	
Sample size	**Strata**	**Sample size**	**Strata**	**Sample size**	**strata**
78,018	43,030	11,963	1,793	84.67%	95.83%

Table 3.4 shows an 84.67% reduction in sample size and a 95.83% reduction in the number of strata after 100 iterations. Figure 3.4 shows that for the same starting chromosome population size for Domain 1 of the Kiva Loans dataset, the GGA attained a good sample size in less than 100 iterations, but after 10,000 iterations the GA had not converged and the sample size was still much higher than the GGA.

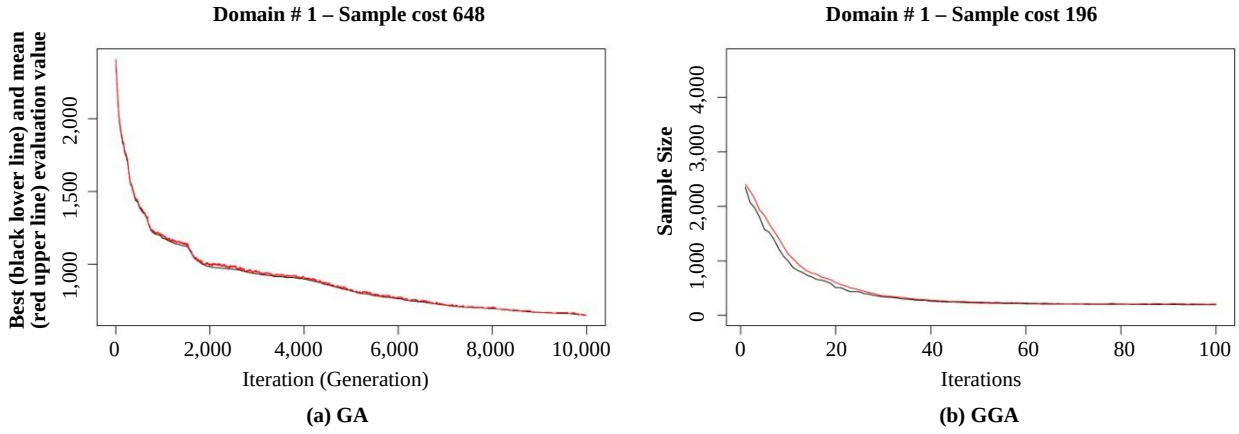


Figure 3.4 Convergence plots for Sample Size for the 1st Domain for GA (10,000 iterations) and GGA (100 iterations) in the Kiva Loans dataset experiment. Note the different scales on the vertical and horizontal axes.

3.5 UN Commodity Trade Statistics data

Kaggle also hosts a copy of the UN Statistical Division Commodity Trade Statistics data. Trade records are available from 1962. We took a subset of data for the year 2011 and removed records with missing observations. This resulted in a data set with 351,057 records. We selected the following target variable:

1. trade_usd

which refers to the value of trade in USD (US dollars), and the following auxiliary variables:

1. commodity,
2. flow,
3. category.

The variable commodity is a categorical description of the type of commodity, e.g., Horses, live except pure-bred breeding. The variable flow describes whether the commodity was an import, export, re-import or re-export. The variable category describes the category of commodity, e.g., silk or fertilisers. The 171 categories of country or area were selected as domains.

Table 3.5
Sample size and strata for the UN Commodity Trade Statistics data from the GA and the GGA after 100 iterations

GA		GGA		Reduction	
Sample size	**Strata**	**Sample size**	**Strata**	**Sample size**	**strata**
288,638	191,000	84,181	16,555	70.84%	91.33%

3.6 2000 US census data

The Integrated Public Use Microdata Series extract is a 5% sample of the 2000 US census data (Ruggles, Genadek, Goeken, Grover and Sobek, 2017). The file contains 6,184,483 records. The US Census Data will be very similar to the ACS data as the latter is an annual version of the former. But for this experiment we selected different target and auxiliary variable combinations. The single target variable in this test is usually a key focus of household surveys:

1. total household income.

We used the following information as auxiliary variables (note these are variables which are likely available in administrative data):

1. annual property insurance cost,
2. annual home heating fuel cost,
3. annual electricity cost,
4. house value.

The house value variable (VALUEH) reports the midpoint of house value intervals (e.g., 5,000 is the midpoint of the interval of less than 10,000), so we have treated it as a categorical variable. As with the 2015 ACS PUMS dataset we have taken a subset for which all values are present. This has resulted in a subset with 627,611 records. The domain for this experiment was Census region and division.

Table 3.6
Sample size and strata for the 2000 US census data by Census region and division from the GA and the GGA after 100 iterations

	Sampling frame		GA solution		GGA solution	
Division	**Sampling Units**	**Atomic Strata**	**Sample sizes**	**Strata**	**Sample sizes**	**Strata**
New England	116,045	87,084	81,012	52,628	376	58
Middle Atlantic	183,543	138,470	130,862	86,002	416	75
East North Central	65,480	58,055	53,075	35,794	327	42
West North Central	31,408	29,413	26,525	18,248	324	38
South Atlantic	97,189	83,357	76,716	51,457	440	49
East South Central	21,631	20,429	18,256	12,500	451	62
West South Central	22,582	20,919	18,750	12,730	407	39
Mountain	26,765	25,041	22,161	14,791	351	30
Pacific	62,968	54,864	50,136	33,653	358	49
Total	627,611	517,632	477,493	317,803	3,446	442

The results show a sample size of 3,446 for the GGA and a sample size of 477,493 for the GA after 100 iterations.

4 An improved Bethel implementation

Our GGA was proposed and developed so that it would work with the rest of the functions in *SamplingStrata*. Therefore the rest of the functions in the package remained unchanged. This includes the *bethel.r* function which evaluates the fitness of chromosomes in every iteration and is computationally expensive. For instance, for the PUMS dataset the experiment took approximately 30 days for either GA or GGA with 100 iterations.

We searched for performance bottlenecks in *bethel.r* using the R *lineprof* package. Our analysis of results suggested that the function within *bethel.r* called *chromy* appears to take the bulk of computational time. A further examination reveals that *chromy* contains a while loop with a default setting of 200 iterations. Furthermore *bethel.r* itself can be run on each chromosome in any chromosome population on a dataset of any functional size (which we have the computation power to process) for any number of iterations. Bigger datasets will take longer to process. We expected that performance would be improved by converting the *bethel.r* algorithm into C++ then integrating that into R using the *Rcpp* package (Eddelbuettel, 2013).

Table 4.1
Performance comparison for the above datasets using the R and Rcpp versions of the Bethel-Chromy algorithm

Dataset	Records	Domains	Atomic Strata	Bethel μs	BethelRcpp μs	Speed-up Factor
iris	150	1	8	2,684.77	143.13	18.76
swissmunicipalities	2,896	7	641	99,916	10,749.51	9.29
American Community Survey 2015	619,747	51	123,007	565,278,500	47,858,200	11.81
Kiva Loans Data	614,361	73	84,897	826,297,710	82,894,480	9.97
UN Commodity Trade Data 2011	351,057	171	350,895	139,749,810	87,555,870	1.6
US Census Data 2000	627,611	9	517,632	2,686,771	1,303,667	2.06

Table 4.1 shows the median time taken to run the Bethel algorithm one hundred times for the datasets we used to conduct our analysis. Our results confirm that the C++ version of Bethel is faster than the R version. The speed up could make a practical difference in the number of iterations that can be run in *SamplingStrata* due to the processing times required for *bethel.r*. However, performance will vary according to the size and complexity of the problem. The speed up is achieved because C++ enables communication at a lower level with the computer than R. However, it is also due to the complexity of the analysis conducted in each for loop as well as the fact that larger data will restrict the available memory. It should also be noted that the C++ version of Bethel was compared with the R version as two stand alone functions. The performance of the C++ version of Bethel within the GGA is not compared with that of the R version in the GA. This would be part of a larger project to create a C++ version of the *SamplingStrata* package and integrating it into R.

5 Conclusion and further work

We created a GGA as an alternative to the existing *SamplingStrata* GA in R. We then compared the two algorithms using a number of datasets. The GGA compares favourably with the GA at finding the correct solution and meeting constraints on smaller datasets, but significantly outperforms the GA on larger datasets where the number of iterations was restricted. This is useful for datasets where the number of iterations has to be constrained owing to computational burden. We have also reported faster processing times by integrating the *bethel.r* function with C++ using the Rcpp package.

This work can be developed in several ways. Alternative evaluation techniques to speed up the algorithm could be considered. Further research could also be undertaken into other machine learning techniques for solving this problem.

The GGA could be applied to other problems which tackle more general sampling designs with modifications required only for the algorithm evaluating the fitness of chromosomes (i.e., the Bethel-Chromy algorithm). For example instead of searching for a stratified simple random sample to meet precision constraints based on population totals or means, the GGA could consider stratified probability proportional to size sampling with an evaluation algorithm that uses more general estimators (e.g., regression or ratio estimators) or more general parameters (e.g., a correlation coefficient).

The evaluation algorithm might also be modified to look at scenarios in which the population variances are not known. In these cases, data from previous censuses, administrative records, or proxy surveys can be used to estimate the population variance. However, estimation of the population variance in a large number of atomic strata requires more careful research.

Finally, the groupings of atomic strata by the GGA can be difficult to interpret. For instance, an ordinal auxiliary variable taking values 1 to 4 may be unnaturally separated, where the atomic strata corresponding to values 1 and 3 are grouped in one design stratum and those with values 2 and 4 are grouped in another design stratum. It might be interesting to explore less-than-optimal sample sizes for stratifications that are easier to interpret. For instance, one may impose constraints on the admissible groupings. This would require research into the formulation of appropriate admissibility constraints and their effective implementation in the GGA.

Acknowledgements

We wish to acknowledge Steven Riesz of the Economic Statistical Methods Division of the U.S. Census Bureau and Brian J. McElroy of the Economic Reimbursable Survey Division of the U.S Census Bureau, both of whom answered questions which were of assistance in choosing which U.S. Census Bureau data to use. We would also like to thank Giulio Barcaroli and Marco Ballin, the co-authors of (Ballin and Barcaroli, 2013), for independently testing our GGA. Last but not least we are extremely grateful to the editorial staff and reviewers of *Survey Methodology* for their constructive suggestions in the review process for this journal submission, especially their suggestions for future work.

References

Agustín-Blas, L.E., Salcedo-Sanz, S., Vidales, P., Urueta, G. and Portilla-Figueras, J.A. (2011). Near optimal citywide WiFi network deployment using a hybrid grouping genetic algorithm. *Expert Systems with Applications*, 38(8), 9543-9556.

Anderson, E. (1935). The irises of the gaspe peninsula. *Bulletin of the American Iris society*, 59, 2-5.

Ballin, M., and Barcaroli, G. (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, 39, 2, 369-393. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11884-eng.pdf.

Barcaroli, G. (2014). SamplingStrata: An R package for the optimization of stratified sampling. *Journal of Statistical Software*, 61(4), 1-24.

Barcaroli, G. (2019). Optimization of sampling strata with the SamplingStrata package. https://cran.r-project.org/web/packages/SamplingStrata/vignettes/SamplingStrata.html, accessed April 29, 2019.

Bethel, J.W. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Section*, American Statistical Association, 209-212. https://www.overleaf.com/project/5ae8997d310d9a2939f40335.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey methodology*, 15, 1, 47-57. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1989001/article/14578-eng.pdf.

Brown, E.C., and Vroblefski, M. (2004). A grouping genetic algorithm for the microcell sectorization problem. *Engineering Applications of Artificial Intelligence*, 17(6), 589-598.

Chromy, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Section*, American Statistical Association.

De Lit, P., Falkenauer, E. and Delchambre, A. (2000). Grouping genetic algorithms: An efficient method to solve the cell formation problem.

Eddelbuettel, E. (2013). Seamless R and C++ Integration with Rcpp, *ISBN 978-1-4614-6867-7* 10.1007/978-1-4614-6868-4.

Falkenauer, E. (1998). *Genetic Algorithms and Grouping Problems*. New York: John Wiley & Sons, Inc.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188.

Galinier, P., and Hao, J.K. (1999). Hybrid evolutionary algorithms for graph coloring. *Journal of Combinatorial Optimization*, 3(4), 379-397.

Hartigan, J.A., and Wong, M.A. (1979). Hybrid evolutionary algorithms for graph coloring.algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C* (Applied Statistics), 28(1), 100-108.

Hung, C., Sumichrast, R.T. and Brown, E.C. (2003). CPGEA: A grouping genetic algorithm for material cutting plan generation. *Computers & Industrial Engineering*, 44(4), 651-672.

James, T., Brown, E. and Ragsdale, C.T. (2010). Grouping genetic algorithm for the blockmodel problem. *IEEE Transactions on Evolutionary Computation*, 14(1), 103-111.

Pelikan, M., and Goldberg, D.E. (2000). Genetic algorithms, clustering, and the breaking of symmetry. *Proceedings of the Sixth International Conference on Parallel Problem Solving from Nature*.

Prügel-Bennett, A. (2004). Symmetry breaking in population-based optimization. *IEEE Transactions on Evolutionary Computation*, 8(1), 63-79.

R Core Team (2015). *R A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruggles, S., Genadek, K., Goeken, R., Grover, J. and Sobek, M. (2017). Integrated public use microdata series: Version 7.0 [dataset]. Minneapolis: University of minnesota.

U.S. Census Bureau (2013). *American Community Survey Information Guide*. http://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf, accessed February 15, 2017.

U.S. Census Bureau (2016). *2015 ACS PUMS DATA DICTIONARY*. http://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMSDataDict15.pdf, accessed February 15, 2017.

U.S. Census Bureau (2016). *2015 ACS Public Use Microdata Sample (PUMS)*. Washington, D.C. https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t#.

Willighagen, E. (2005). Genalg: R based genetic algorithm. *R Package Version 1*.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 533-542
Statistics Canada, Catalogue No. 12-001-X

"Optimal" calibration weights under unit nonresponse in survey sampling

Per Gösta Andersson[1]

Abstract

High nonresponse is a very common problem in sample surveys today. In statistical terms we are worried about increased bias and variance of estimators for population quantities such as totals or means. Different methods have been suggested in order to compensate for this phenomenon. We can roughly divide them into imputation and calibration and it is the latter approach we will focus on here. A wide spectrum of possibilities is included in the class of calibration estimators. We explore linear calibration, where we suggest using a nonresponse version of the design-based optimal regression estimator. Comparisons are made between this estimator and a GREG type estimator. Distance measures play a very important part in the construction of calibration estimators. We show that an estimator of the average response propensity (probability) can be included in the "optimal" distance measure under nonresponse, which will help to reduce the bias of the resulting estimator. To illustrate empirically the theoretically derived results for the suggested estimators, a simulation study has been carried out. The population is called KYBOK and consists of clerical municipalities in Sweden, where the variables include financial as well as size measurements. The results are encouraging for the "optimal" estimator in combination with the estimated average response propensity, where the bias was reduced for most of the Poisson sampling cases in the study.

Key Words: Unit nonresponse; Calibration weights; Poisson sampling.

1 Introduction

In a survey the response (nonresponse) mechanism for units is in reality unknown. To avoid defining a proper probability measure which might not be meaningful or realistic, one usually discusses the nonresponse situation in terms of a propensity for a unit to participate. To be able to take into account the possible nonresponse effect on estimators, it is however the practice to treat the propensities as probabilities to be estimated (e.g., propensity scores). This can be done for individual units, for groups of units or as an "average" over the whole response set.

For example, in Haziza and Lesage (2016) two main approaches are discussed: calibration weighting with and without foregoing propensity score weighting, the former case involving model-based estimation. The authors warn against potential negative effects on the bias and variance for the resulting estimators when not taking into account the propensities. (These two options of weighting are referred to by the authors as two-step and one-step procedures, respectively not to be mistaken for the two- and single-step calibrations as defined by Särndal and Lundström (2005).) However, in the simulation study by Haziza and Lesage (2016) the sampling design plays no role, since there $n = N$ and the focus is solely on how the auxiliary information relates to the study variable and the nonresponse mechanism.

In this paper we propose to use a nonresponse version of what in the full response case is called the (design-based) optimal regression estimator. The underlying distance measure is a quadratic form with a more complex structure (see Andersson and Thorburn (2015)) than the one leading to the GREG estimator

1. Per Gösta Andersson, Associate Professor, Department of Statistics, Stockholm University, Stockholm. E-mail: per.gosta.andersson@stat.su.se.

(see Deville and Särndal (1992)). As it turns out there is also room for refinement in terms of the average response propensity (probability) when constructing the distance measure under nonresponse, which leads to a modified “optimal” estimator.

1.1 Outline of the paper

Section 2 starts with an introduction to the calibration idea under full response before dealing with the nonresponse situation. Three estimators of a population total are mainly considered: the GREG related estimator and two versions of the “optimal” estimator. Some theoretical results for the resulting bias follows. Section 3 contains a simulation study where simple random sampling and Poisson sampling are used for illustration. The Poisson design enables us to construct and investigate a situation where the auxiliary information is involved in the design as well as in the nonresponse mechanism. We also illustrate the risks of using an incorrect model when estimating individual propensities. We end with concluding remarks in Section 4.

1.2 Notation and setup

We will start with a population U of size N from which we take a probability sample s of size n_s with inclusion probabilities $\pi_1, \ldots, \pi_N$. Nonresponse means that we only observe the response set r of size n_r. Our aim is to estimate the study variable total $t_y = \sum_U y_k$. We assume access to an auxiliary variable vector $\mathbf{x}$ of dimension J, where either $\mathbf{x} = \mathbf{x}^*$ and $\left(\mathbf{x}_k^*\right)_{k\in U}$ are known (the population level) or $\mathbf{x} = \mathbf{x}^o$ and $\left(\mathbf{x}_k^o\right)_{k\in s}$ are known (the sample level) or possibly a mixture of these cases: $\mathbf{x} = \left(\mathbf{x}^{*\prime},\ \mathbf{x}^{o\prime}\right)'$.

2 Calibration estimation

2.1 Calibration estimators under full response

Starting with the full response situation $(r = s)$ and following the procedure as established by Deville and Särndal (1992), the calibration estimator is defined as

$$\hat{t}_{y\text{cal}} = \sum_s w_{ks} y_k,$$

where the sample dependent weights w_{ks} are chosen so that

$$\sum_s w_{ks}\mathbf{x}_k = \mathbf{t}_x, \text{ (the calibration equation)} \tag{2.1}$$

while also minimizing the quadratic distance measure

$$\left(\mathbf{w}_s - \mathbf{w}_{0s}\right)' \mathbf{R} \left(\mathbf{w}_s - \mathbf{w}_{0s}\right),$$

where $\mathbf{w}_s = (w_{ks})_{k\in s}$, $\mathbf{w}_{0s} = (1/\pi_k)_{k\in s} = (d_k)_{k\in s}$ and $\mathbf{R}$ is diagonal. (Alternative distance measures are considered in both Deville and Särndal (1992) and Haziza and Lesage (2016).)

In other words, given the constraint (2.1) the w_{ks} should be "as close as possible" to the design weights d_k, which is desirable since $\sum_s d_k y_k$ is an unbiased estimator of t_y.

The resulting weights are

$$\mathbf{w}_s = \mathbf{w}_{0s} + \mathbf{R}^{-1}\mathbf{x}'\left(\mathbf{X}\mathbf{R}^{-1}\mathbf{X}'\right)^{-1}\left(\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_\mathbf{x}\right).$$

It turns out that the model assisted homoskedastic GREG estimator $\hat{t}_{yr}$ (Särndal, Swensson and Wretman (1992)) is a calibration estimator for which

$$\mathbf{R} = \left(\mathbf{w}_{0s}\mathbf{I}_{n_s}\right)^{-1},$$

where I_{n_s} is the unit diagonal matrix of size n_s.

Another calibration estimator is the optimal regression estimator $\hat{t}_{y\text{opt}}$ (see e.g., Rao (1994) and Montanari (1998)), for which

$$\mathbf{R} = \left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}\pi_k\pi_l}\right)^{-1}_{k,l\in s},$$

as shown by Andersson and Thorburn (2005).

Asymptotically, this estimator has (in a design-based sense) minimum variance among linear regression type estimators.

2.2 Calibration estimators under nonresponse

In the nonresponse case, a possible calibration estimator is

$$\sum_r w_{kr} y_k,$$

where it should hold that

$$\sum_r w_{kr}\mathbf{x}_k = \mathbf{X}, \tag{2.2}$$

where $\mathbf{X} = \sum_U \mathbf{x}_k^*$, if the auxiliary information is known up to the population level. Otherwise, $\mathbf{X} = \sum_s d_k \mathbf{x}_k^o$, the unbiased estimator of $\mathbf{t}_\mathbf{x}$. (We can also combine the two types of information in the constraint $\mathbf{X}$.)

For a variety of cases weights fulfilling the requirement (2.2) are presented by e.g., Särndal and Lundström (2005). Using the direct approach, where all information is used in one single calibration, we get

$$w_{kr} = d_k \left(1 + \mathbf{x}_k' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right)\right). \tag{2.3}$$

The resulting estimator will henceforth be denoted $\hat{t}_{y\text{cal}}$. (Other approaches, including two-step procedures, are presented and investigated by e.g., Andersson and Särndal (2016).)

An evident question to ask is: What is the underlying distance measure generating these weights? Särndal and Lundström (2005) do not comment on this particular issue, but according to Lundström and Särndal (1999), we should choose “w_k ‘as close as possible’ to the d_k”, which does not seem quite adequate under nonresponse. Going back to Lundström (1997) we will find that the corresponding distance measure *is* actually

$$(\mathbf{w}_r - \mathbf{w}_{0r})' \left(\mathbf{w}_{0r} \mathbf{I}_{n_r}\right)^{-1} (\mathbf{w}_r - \mathbf{w}_{0r}),$$

where $\mathbf{w}_r = (w_{kr})_{k \in r}$ and $\mathbf{w}_{0r} = (d_k)_{k \in r}$.

If we have a random mechanism generating the response set r from the sample s with probabilities θ_k of inclusion, we can view the nonresponse situation as a two-phase design and this is the assumption we will make in the following. Then we should minimize the distance between w_{kr} and $d_k \cdot (1/\theta_k)$. Using some modelling θ_k can be estimated by $\hat{\theta}_k$, to be put to use for the distance minimization. But in this paper we will not go in the direction of model-based inference. In order to reduce the bias effect under nonresponse one could instead in the distance measure think of comparing w_{kr} not with d_k, but with $d_{k,\text{alt}} = d_k \cdot c$, where c is a constant larger than 1, aiming to compensate for the “average” nonresponse effect.

However, Lundström (1997) shows that in many important cases, namely when one can find a vector $\boldsymbol{\mu}$ for which $\boldsymbol{\mu}'\mathbf{x}_k = 1$, for all k, the multiplicative increase in $d_{k,\text{alt}}$ implies the same resulting calibration weights w_{kr}. This follows from the result that if $\boldsymbol{\mu}'\mathbf{x}_k = 1$, for all $k \in U$, we can simplify the expression (2.3) of w_{kr} as

$$w_{kr} = d_k \mathbf{x}_k' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{X}.$$

Thus, we have an invariance property for the weights. The result holds also when the population is partitioned into groups and the initial weights are inflated with a constant within each group. Note that if we include a constant, e.g., “1”, as a first component of the auxiliary vector $\mathbf{x}_k$, we can simply let $\boldsymbol{\mu}' = (1, 0, \ldots, 0)$ to achieve $\boldsymbol{\mu}'\mathbf{x}_k = 1$.

With this as a background we propose to use alternative “optimal” weights resulting from the distance measure

$$(\mathbf{w}_r - \mathbf{w}_{0r})' \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l}\right)_{k,l \in r}^{-1} (\mathbf{w}_r - \mathbf{w}_{0r}),$$

leading to $\hat{t}_{y\text{opt}}$. (π_{kl} denotes the inclusion probability for the pair (k, l)).

It is to be observed that as for the full response situation, there are cases for which the "optimal" weights are identical to (2.3), as e.g., under simple random sampling.

Using quotation marks around *optimal* is deliberate, but under full response *optimal* has a very clear meaning. As mentioned earlier, the optimal regression estimator has asymptotically minimum variance among linear regression estimators. Adding nonresponse where the nonresponse mechanism is at least partially unknown, makes it difficult to define optimality criteria in a proper way.

For this "optimal" measure it might be fruitful to replace d_k with $d_{k,\text{alt}}$, where we include in $d_{k,\text{alt}}$ the reciprocal of an estimate of the average response probability $\bar{\theta}_U = \sum_U \theta_k \big/ N$. One simple candidate is

$$\hat{\bar{\theta}}_U = n_r / n_s ,$$

thus yielding $d_{k,\text{alt}} = d_k \cdot (n_s / n_r)$. Another natural choice is

$$\hat{\bar{\theta}}_U = \sum_r d_k \Big/ \sum_s d_k , \tag{2.4}$$

since $E\left(\sum_s d_k\right) = N$ and $E\left(\sum_r d_k\right) = \sum_U \theta_k = N\bar{\theta}$, which lead to $E\left(\sum_r d_k \Big/ \sum_s d_k\right) \approx \bar{\theta}_U$. The resulting modified estimator is denoted by $\hat{t}_{y\text{optm}}$. (Also observe that $E(n_r/n_s) \approx \sum_U (\theta_k/d_k) \Big/ \sum_U (1/d_k)$.

In the following simulation study we will focus on a sampling design where generally $\hat{t}_{y\text{cal}} \neq \hat{t}_{y\text{opt}}$, namely Poisson sampling. The independence of drawings simplifies the "optimal" distance measure:

$$\sum_r \frac{\pi_k^2}{1-\pi_k} (w_{kr} - d_k)^2 = \sum_r \frac{(w_{kr} - d_k)^2}{d_k (d_k - 1)}$$

and minimization yields

$$w_{kr} = d_k \left(1 + (d_k - 1)\, \mathbf{x}_k' \left(\sum_r d_k (1 - d_k)\, \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right)\right).$$

For the modified "optimal" estimator d_k is replaced by $d_{k\text{alt}} = d_k \cdot \left(1 \big/ \hat{\bar{\theta}}_U\right)$, with $\hat{\bar{\theta}}_U$ as in (2.4).

2.2.1 Bias for calibration estimators under nonresponse

We can write $\hat{t}_{y\text{cal}}$ as

$$\hat{t}_{y\text{cal}} = \sum_r d_k y_k + \hat{\mathbf{B}}_{U;\theta} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right), \tag{2.5}$$

where $\hat{\mathbf{B}}_{U;\theta} = \left(\sum_r d_k \mathbf{x}_k' y_k\right)\left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$. In order to arrive at an approximate expression for the bias of $\hat{t}_{y\text{cal}}$ and subsequently $\hat{t}_{y\text{opt}}$ and $\hat{t}_{y\text{optm}}$, we follow the derivation in Särndal and Lundström (2005) and first note that $\hat{t}_{y\text{cal}}$ can be rewritten as

$$\hat{t}_{y\text{cal}} = \sum_r d_k y_k + \mathbf{B}_{U;\theta} \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right) + \left(\hat{\mathbf{B}}_{U;\theta} - \mathbf{B}_{U;\theta}\right)\left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right),$$

where $\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}_k' y_k\right)\left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$.

If we let $\hat{t}_{y\text{cal}} - t_y = A_1 + A_2$, where $A_1 = \sum_r d_k y_k - t_y + \mathbf{B}_{U;\theta}\left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right)$ and $A_2 = \left(\hat{\mathbf{B}}_{U;\theta} - \mathbf{B}_{U;\theta}\right)\left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right)$, it can further be shown that

$$A_1 = \sum_r d_k e_{\theta k} - \sum_U e_{\theta k} + \mathbf{B}^o_{U;\theta}\left(\sum_s d_k \mathbf{x}^o_k - \sum_U \mathbf{x}^o_k\right),$$

where $e_{\theta k} = y_k - \mathbf{B}_{U;\theta}\mathbf{x}_k$ and $\mathbf{B}^o_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}^o_k \mathbf{x}^{o\prime}_k\right)^{-1} \sum_U \theta_k \mathbf{x}^o_k y_k$.

Then

$$E\left(\hat{t}_{y\text{cal}}\right) - t_y \approx E\left(A_1\right) = \sum_U \theta_k e_{\theta k} - \sum_U e_{\theta k} = -\sum_U \left(1 - \theta_k\right) e_{\theta k},$$

since it can be argued that $\hat{\mathbf{B}}_{U;\theta}$ is a consistent estimator of $\mathbf{B}_{U;\theta}$ and therefore $E\left(A_2\right) \approx 0$.

The approximation for the bias of $\hat{t}_{y\text{cal}}$ is called the nearbias:

$$\text{nearbias}\left(\hat{t}_{y\text{cal}}\right) = -\sum_U \left(1 - \theta_k\right) e_{\theta k}.$$

The nearbias of $\hat{t}_{y\text{cal}}$ is zero if $\theta_k = 1$, for all $k \in U$ and/or $y_k = \mathbf{B}_{U;\theta}\mathbf{x}_k$, for all $k \in U$.

Then, if we consider $\hat{t}_{y\text{opt}}$, we have that

$$\hat{t}_{y\text{opt}} = \sum_r d_k y_k + \left(\mathbf{X} - \sum_r d_k \mathbf{x}_k\right)\hat{\mathbf{C}}_{U;\theta}, \tag{2.6}$$

where

$$\hat{\mathbf{C}}_{U;\theta} = \left(\sum_{k\in r}\sum_{l\in r} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \frac{\mathbf{x}'_k}{\pi_k} \frac{y_l}{\pi_l}\right)\left(\sum_{k\in r}\sum_{l\in r} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \frac{\mathbf{x}_k}{\pi_k} \frac{\mathbf{x}'_l}{\pi_l}\right)^{-1}.$$

Since $\hat{t}_{y\text{opt}}$ can be written as (2.6), which is of the same form as for $\hat{t}_{y\text{cal}}$ in (2.5), we will again arrive at the nearbias expression

$$\text{nearbias}\left(\hat{t}_{y\text{opt}}\right) = -\sum_U \left(1 - \theta_k\right) e_{\theta k}, \tag{2.7}$$

where $e_{\theta k} = y_k - \mathbf{C}_{U;\theta}\mathbf{x}_k$ and with θ_{kl} denoting the response probability for the pair (k, l):

$$\mathbf{C}_{U;\theta} = \left(\sum_{k\in U}\sum_{l\in U} \theta_{kl}\left(\pi_{kl} - \pi_k\pi_l\right)\frac{\mathbf{x}'_k}{\pi_k}\frac{y_l}{\pi_l}\right)\left(\sum_{k\in U}\sum_{l\in U} \theta_{kl}\left(\pi_{kl} - \pi_k\pi_l\right)\frac{\mathbf{x}_k}{\pi_k}\frac{\mathbf{x}'_l}{\pi_l}\right)^{-1}.$$

If we use the alternative weighting $d_{k,\text{alt}} = d_k \cdot \left(1/\hat{\bar{\theta}}\right) = d_k \cdot \left(\sum_s d_k / \sum_r d_k\right)$, we get that

$$\text{nearbias}\ \left(\hat{t}_{y\text{optm}}\right) = E\left(\sum_r d_{k,\text{alt}} e_{\theta k} - \sum_U e_{\theta k}\right) \approx \sum_U \frac{\theta_k}{\bar{\theta}_U} e_{\theta k} - \sum_U e_{\theta k} = -\sum_U \left(1 - \frac{\theta_k}{\bar{\theta}_U}\right) e_{\theta k},$$

where $\sum_U \left(1 - \left(\theta_k / \bar{\theta}_U\right)\right) = 0$, to be compared with (2.7), where $\sum_U \left(1 - \theta_k\right) = N\left(1 - \bar{\theta}_U\right)$.

Unless $\boldsymbol{\mu}'\mathbf{x}_k = 1$, for all $k \in U$, an equivalent expression can be obtained for $\hat{t}_{y\text{cal}}$. On the other hand, if the restriction $\boldsymbol{\mu}'\mathbf{x}_k = 1$, for all $k \in U$ *does* hold, it can be shown (Särndal and Lundström (2005)) that

$$\text{nearbias}\left(\hat{t}_{y\text{cal}}\right) = -\sum_{U} e_{\theta k},$$

which holds independently of the sampling design and which is a result completely in line with the aforementioned invariance property of the calibration weights.

3 A simulation study

Properties of the estimators were studied by means of a Monte Carlo simulation. We used an authentic population called KYBOK, which consists of $N = 832$ clerical municipalities in Sweden in 1992. (This population was also used for simulation purposes in Särndal and Lundström (2005) and Andersson and Särndal (2016).)

The study variable y_k is "Expenditure on administration and maintenance" ($t_y = 1{,}023{,}983$). The population is divided into four groups with respect to size, from the smallest to the largest. The group sizes are $N_1 = 218$, $N_2 = 272$, $N_3 = 290$ and $N_4 = 52$. The moon vector is $\mathbf{x}_k^o = \left(x_{1k}^o, \ldots, x_{4k}^o\right)'$, where $x_{ik}^o = 1$ if the unit k belongs to population group i and otherwise 0, $i = 1, \ldots, 4$. The quantitative star variable x_k^* is the square root of "Revenue advances", which is highly positively correlated with y_k.

The sample size/expected sample size was 300 and we used the exponential response probability

$$\theta_k = 1 - \exp\left(-c \cdot x_k^*\right),\ k \in U, \tag{3.1}$$

where c is chosen according to the desired average response probability; in this study varying between 0.60 and 0.86 (the latter value being the chosen response probability in e.g., Särndal and Lundström (2005)). Two sampling designs have been considered separately: simple random sampling and Poisson sampling. In the latter case $\pi_k \propto x_k^*$. For each combination of design, sample size/expected sample size and average response probability, 10,000 samples were generated. For each such sample s, a response set r was created by performing independent Bernoulli trials with probability θ_k of success, $k \in s$.

The estimators of main interest are $\hat{t}_{y\text{cal}}$, $\hat{t}_{y\text{opt}}$ and $\hat{t}_{y\text{optm}}$, but in this simulation study we will also include an example of parametric propensity modelling based on $\hat{t}_{y\text{cal}}$. A simple choice is the logistic (logit) model

$$\theta_k = \frac{\exp\left(\mathbf{x}'\mathbf{B}\right)}{1 + \exp\left(\mathbf{x}'\mathbf{B}\right)}, \tag{3.2}$$

where we let $\mathbf{x}' = \left(1\ x^*\right)$. For each sample with its observed nonresponse maximum likelihood estimation was used to obtain $\hat{\mathbf{B}}$, yielding estimates $\left(\hat{\theta}\right)_{k \in r}$. To obtain $\hat{t}_{y\,\text{cal logit}}$ the design weights d_k are then replaced by $d_k\left(1/\hat{\theta}_k\right)$ before calibration. The logit model (3.2) is misspecified since the true response probability is determined by (3.1).

An arbitrary estimator $\hat{t}_y$ is assessed by the empirical (simulation estimated) bias $\left(\hat{B}\right)$, variance $\left(\hat{V}\right)$ and mean squared error $\left(\widehat{\text{MSE}}\right)$:

$$\hat{B} = \hat{E}\left(\hat{t}_y\right) - t_y = \frac{1}{K}\sum_{j=1}^{K}\hat{t}_{yj} - t_y$$

$$\hat{V} = \frac{1}{K}\sum_{j=1}^{K}\left(\hat{t}_{yj} - \hat{E}\left(\hat{t}_y\right)\right)^2$$

$$\widehat{\mathrm{MSE}} = \hat{B}^2 + \hat{V},$$

where $K = 10{,}000$.

Observe that expressions such as "the bias has increased" should be interpreted in the following as an increase of the bias in absolute value.

3.1 Results

As a benchmark for the study where auxiliary information is not used at the design stage, let us first consider the results for simple random sampling in Table 3.1. This is a case where $\hat{t}_{y\mathrm{cal}} = \hat{t}_{y\mathrm{opt}}$. (Actually, to get equality the "star" information is $\mathbf{x}_k^* = \left(1,\, x_k^*\right)'$ for $\hat{t}_{y\mathrm{cal}}$.) As will hold throughout this study the bias of $\hat{t}_{y\,\mathrm{cal\,logit}}$ is considerably larger than the bias of $\hat{t}_{y\mathrm{cal}}$, which is a natural effect from the construction of $\hat{t}_{y\,\mathrm{cal\,logit}}$ based on a misspecified nonresponse model. Furthermore, of these two estimators $\hat{t}_{y\,\mathrm{cal\,logit}}$ has always the largest variance.

Looking instead at the results in Table 3.2 for Poisson sampling, we can first observe that for $\hat{t}_{y\mathrm{cal}}$ both the bias and the variance are larger than under simple random sampling. $\hat{t}_{y\mathrm{opt}}$, on the other hand, has highly reduced bias under Poisson sampling compared with simple random sampling, whereas there is a slight increase in the variance. Then, turning to the proposed modified estimator $\hat{t}_{y\mathrm{optm}}$ we observe a further reduction in bias, except for $\bar{\theta}_U = 0.86$. Actually, the bias has a monotonic behaviour and changes sign from positive to negative for $\bar{\theta}_U \approx 0.64$. However, compared with $\hat{t}_{y\mathrm{opt}}$ the variance of $\hat{t}_{y\mathrm{optm}}$ is increased due to the inclusion of $\hat{\bar{\theta}}_U$ in (2.4), thus leading to a trade-off between the bias and the variance. We also note that of these two estimators $\hat{t}_{y\mathrm{optm}}$ displays the largest MSE values, since the dominating part of the MSE is the variance for these low levels of bias.

Table 3.1
Empirical bias ($\hat{B}$), variance ($\hat{V}$) and mean squared error ($\widehat{\mathrm{MSE}}$) for $\hat{t}_{y\mathrm{cal}}$ (Cal), $\hat{t}_{y\,\mathrm{cal\,logit}}$ (Cal logit) and $\hat{t}_{y\mathrm{opt}}$ (Opt) under simple random sampling (n = 300) with average response probabilities 0.86, 0.70, 0.65 and 0.60

		Simple random sampling (Cal = Opt)			
$\hat{B}\left(*10^{-4}\right)$	$\bar{\theta}_U$	0.86	0.70	0.65	0.60
	Cal	-2.44	-4.00	-4.47	-4.89
	Cal logit	4.81	19.4	26.4	35.5
$\hat{V}\left(*10^{-8}\right)$	$\bar{\theta}_U$	0.86	0.70	0.65	0.60
	Cal	8.40	9.59	10.2	11.3
	Cal logit	10.7	10.9	13.2	16.1
$\widehat{\mathrm{MSE}}\left(*10^{-9}\right)$	$\bar{\theta}_U$	0.86	0.70	0.65	0.60
	Cal	1.44	2.57	3.01	3.52
	Cal logit	3.38	38.9	71.9	127

Table 3.2
Empirical bias ($\hat{B}$), variance ($\hat{V}$) and mean squared error ($\widehat{\mathrm{MSE}}$) for $\hat{t}_{y\mathrm{cal}}$ (Cal), $\hat{t}_{y\mathrm{cal\,logit}}$ (Cal logit), $\hat{t}_{y\mathrm{opt}}$ (Opt) and $\hat{t}_{y\mathrm{optm}}$ (Optm) under Poisson sampling ($E(n) = 300$) with average response probabilities 0.86, 0.70, 0.65 and 0.60

		Poisson sampling			
$\hat{B}\,(*10^{-4})$	$\bar{\theta}_U$	0.86	0.70	0.65	0.60
	Cal	-2.88	-4.71	-5.17	-5.69
	Cal logit	-12.1	-27.5	-32.9	-38.8
	Opt	-0.0732	-0.329	-0.516	-0.810
	Optm	0.690	0.274	0.0536	-0.277
$\hat{V}\,(*10^{-9})$	$\bar{\theta}_U$	0.86	0.70	0.65	0.60
	Cal	4.46	5.25	5.56	5.81
	Cal logit	5.17	6.60	7.17	7.57
	Opt	1.39	1.63	1.75	1.84
	Optm	2.05	2.89	3.22	3.51
$\widehat{\mathrm{MSE}}\,(*10^{-9})$	$\bar{\theta}_U$	0.86	0.70	0.65	0.60
	Cal	5.29	7.47	8.23	9.05
	Cal logit	19.8	82.2	115	127
	Opt	1.39	1.64	1.78	1.91
	Optm	2.10	2.90	3.22	3.52

4 Concluding remarks

The family of linear calibration techniques in survey sampling contains a variety of alternative weightings under full response, including GREG estimators and the optimal regression estimator. The nonresponse situation offers still more options and challenges and we have studied the "optimal" estimator while also taking into account average response propensities (probabilities). The approach has been design-based since the modified "optimal" estimator can be motivated by asymptotic argumentation and we have furthermore not used any modelling for the response propensities. The results are encouraging, especially concerning reduction of the bias for the suggested estimator. Further work will include the construction of a variance estimator, which should be valid conditionally on the size of the response set.

Acknowledgements

The author is grateful to Carl-Erik Särndal for valuable discussions and suggestions during the research that eventually led to this paper. The author also thanks an anonymous reviewer and an associate editor for helpful comments on the manuscript.

References

Andersson, P.G., and Särndal, C.-E. (2016). Calibration for nonresponse treatment: In one or two steps? *Statistical Journal of the IAOS*, 32, 375-381.

Andersson, P.G., and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31, 1, 95-99. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005001/article/8092-eng.pdf.

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

Haziza, D., and Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1), 129-145.

Lundström, S. (1997). *Calibration as a Standard Method for Treatment of Nonresponse*, Ph.D. thesis, Department of Statistics, Stockholm University.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 13, 305-327.

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 1, 69-77. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1998001/article/3911-eng.pdf.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester, UK: Wiley.

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

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 543-565
Statistics Canada, Catalogue No. 12-001-X

A method to correct for frame membership error in dual frame estimators

Dong Lin, Zhaoce Liu and Lynne Stokes[1]

Abstract

Dual frame surveys are useful when no single frame with adequate coverage exists. However estimators from dual frame designs require knowledge of the frame memberships of each sampled unit. When this information is not available from the frame itself, it is often collected from the respondent. When respondents provide incorrect membership information, the resulting estimators of means or totals can be biased. A method for reducing this bias, using accurate membership information obtained about a subsample of respondents, is proposed. The properties of the new estimator are examined and compared to alternative estimators. The proposed estimator is applied to the data from the motivating example, which was a recreational angler survey, using an address frame and an incomplete fishing license frame.

Key Words: Hartley estimator; Bias-adjustment; Misclassification.

1 Introduction

In single frame surveys, all sample units are selected from the same frame. This is the preferred method when the frame covers the population of interest. But no single frame with adequate coverage exists for some applications, or it may be too expensive to sample from a complete frame that also includes units that are not part of the target population. In such cases, a multiple frame design can be an attractive alternative.

The dual frame design, in which two overlapping frames are used for sampling, was introduced by Hartley (1962). The frames, labeled A and B, collectively cover the population of interest U, which is made up of three mutually exclusive subsets known as domains. Domain a contains units in frame A but not in B; domain b contains units in frame B but not in A; and domain ab contains the units in the intersection of A and B. In dual frame surveys, samples are selected independently from frames A and B. The sampler typically does not know in advance which frame units are in which domains, but will ascertain this for the units that are sampled as part of the data collection process.

Several estimators are available for combining the information from the two samples to estimate parameters of U. All require adjusting the weights of the sampled units based on domain membership. Thus, knowing the domain of sampled units is critical for estimation from dual frame sample designs. When domain membership can be determined accurately, unbiased estimators of means and totals can be constructed. However, when estimators are based on inaccurate domain information, they can be biased. Hartley (1962) developed a family of dual frame estimators. In this paper we propose a bias correction method for Hartley's estimator that is available if either the misclassification probabilities are known, or more realistically, can be estimated from a random subsample of respondents. This estimator is compared

1. Dong Lin is Manager, PenFed Credit Union, Mclean, Virginia. E-mail: schewiz1984@msn.com; Zhaoce Liu is Ph.D. Candidate in Statistics, Southern Methodist University, Dallas, TX. E-mail: zhaocel@smu.edu; Lynne Stokes is Professor and Chair, Department of Statistical Sciences, Southern Methodist University, Dallas, TX. E-mail: slstokes@mail.smu.edu.

to one proposed by Lohr (2011) for the same purpose, and the different assumptions under which each is justified are clarified.

The motivating application for this work was the Marine Recreational Information Program (MRIP), a data collection system carried out by the National Oceanic and Atmospheric Administration (NOAA), to provide bi-monthly estimates of fish catch by species in U.S. marine waters. One part of the program was the Coastal Household Telephone Survey (CHTS), which provided an estimate of the number of fishing trips made by recreational anglers in each coastal state in each time period. A National Research Council report (NRC, 2006) identified deficiencies of the original survey design, and pilot surveys tested improvements. Several of these pilots had a dual frame sample design, which supplemented a residential frame (either a telephone or address frame of all households in the state) with the state's angler license registry. The telephone frame is inefficient; for example, only about 5% of CHTS sample contacts in urban areas reached an angler (NRC, 2006). The license frame is efficient but incomplete, due to omission of anglers for a variety of exemptions from licensing, as well as to illegal angling. Domain membership information for each angler sampled from the residential frame was requested from the survey respondent by asking if he or she was licensed. Andrews, Brick, Mathiowetz and Stokes (2010) demonstrated that the domain membership data obtained in this way is unreliable, as respondents both overreport and underreport license ownership. The goal of this research was to determine the effect of this error on estimation of number of trips and find a method to mitigate the damage.

The pilot study we used for this research was conducted by mail. The two frames were an address frame for all residents of the state, which was provided by US Postal Service, and the license frame listing all anglers licensed by the state, provided by the state natural resource agency. The address frame obviously did not contain information about whether the household members had fishing licenses, so that question was asked of all respondents sampled from the address frame in this study. The license frame did of course have the addresses so that it could be determined whether or not the anglers sampled from this frame resided in the state. Because we had address information, it was possible to link the households on the address frame with the addresses on the license frame fairly easily. However, determining whether the persons responding to the address frame survey were the licensed individuals in the household was laborious, since no names were requested. For this reason, it was difficult to determine whether a household was on both frames or not.

In Section 2, we review estimation for dual frame designs. In Section 3, we present examples to show how domain misclassication error affects estimation. In Section 4, we derive a bias correction method for Hartley's dual frame estimator and compare it to an alternative (Lohr, 2011). In Section 5, we present results of a simulation designed to examine the accuracy of inference using the bias-corrected estimators. The method is illustrated in Section 6 by an application to data from one of NOAA's pilot angler surveys. A discussion follows in Section 7.

2 Dual frame estimation

Since the introduction of dual frame sample designs by Hartley (1962), many estimators have been proposed (Fuller and Burmeister, 1972; Kalton and Anderson, 1986; Bankier, 1986; Skinner and Rao, 1996). In this section, we focus on Hartley's estimator, since it was used in our MRIP pilot study application. Using Hartley's original notation, we denote by N, N_A, N_B, N_a, N_b and N_{ab} the number of elements in U, A, B, a, b and ab, respectively. Then the following relationships hold: $N = N_a + N_{ab} + N_b$, $N_A = N_a + N_{ab}$, and $N_B = N_{ab} + N_b$. Denote the samples from frames A and B as s_A and s_B, and unit i's inclusion probability in the two samples as π_i^A and π_i^B.

The population total Y can be written as the sum of totals of the three mutually exclusive domains.

$$Y = Y_a + Y_{ab} + Y_b, \tag{2.1}$$

where $Y_a = \sum_{i \in a} y_i$, $Y_b = \sum_{i \in b} y_i$ and $Y_{ab} = \sum_{i \in ab} y_i$. Estimators of the total can be written as the sum of total estimators in the three different domains, which is

$$\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b. \tag{2.2}$$

Hartley's estimator (1962) is

$$\hat{Y}_H = \hat{Y}_a + \theta \hat{Y}_{ab}^A + (1 - \theta)\, \hat{Y}_{ab}^B + \hat{Y}_b, \tag{2.3}$$

where $\hat{Y}_{ab}^A = \sum_{i \in s_{ab}^A} w_i^A y_i$ denotes the estimator of Y_{ab} using information from frame A, and $\hat{Y}_{ab}^B = \sum_{i \in s_{ab}^B} w_i^B y_i$ is the corresponding estimator from frame B. s_{ab}^A (s_{ab}^B) is the subset of s_A (s_B) consisting of items that fall in domain ab and $w_i^A = 1/\pi_i^A$ and $w_i^B = 1/\pi_i^B$ are the sampling weights based on the inclusion probabilities in frames A and B. In practice, those weights are typically adjusted for non-response and possibly undercoverage as well. Y_a and Y_b are estimated similarly as $\hat{Y}_a = \sum_{i \in s_a} w_i^A y_i$ and $\hat{Y}_b = \sum_{i \in s_b} w_i^B y_i$. θ is a number between 0 and 1 that adjusts the weights of items from frames A and B. When θ denotes a constant, theory provides an optimal value for it that will minimize the estimator's variance. However it will depend on unknown parameters, and thus must be estimated. Another approach that may be used is to select a value of θ that is proportional to the reciprocal of the sample sizes from the two frames. This will be a constant, and will be near optimal if the sample designs have similar design effects and a small overlap domain.

If θ is a constant, $\hat{Y}_H$ is linear in the data, and its properties are easy to calculate. If $\theta = 0$ or 1, the estimator for the overlap domain depends on data from only one of the two frames. The optimal value of θ for minimizing the variance of $\hat{Y}_H$ (Hartley, 1962) is

$$\theta_H = \frac{V\left(\hat{Y}_{ab}^B\right) + \mathrm{Cov}\left(\hat{Y}_b^B, \hat{Y}_{ab}^B\right) - \mathrm{Cov}\left(\hat{Y}_a^A, \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^A\right) + V\left(\hat{Y}_{ab}^B\right)}. \tag{2.4}$$

The variances and covariances in (2.4) are unknown and must be estimated from the sample if the optimal form for θ is to be used in (2.3). In that case, the resulting weights are random and contribute to the variability of the estimator. Another disadvantage of using the optimal value of θ in $\hat{Y}_H$ is that θ_H is different for different response variables, which results in inconsistency among estimators of related quantities. For example, the optimal estimators of the number of shore and boat fishing trips made by anglers do not necessarily sum to the optimal estimator of total number of fishing trips. For that reason, in practice, a constant value of θ is frequently used; for example, a value of $\theta = 1/2$ is sometimes recommended when sample sizes are similar in the two frames (Lohr, 2011).

3 Misclassification in dual frame surveys

The domain of every sampled unit must be known to calculate any dual frame estimator. For example, (2.3) shows that in $\hat{Y}_H$, the weight of a unit sampled from the residential frame should be adjusted by θ or $1 - \theta$ if the respondent is a license holder, but not otherwise. When this information is unavailable from the frames themselves, it must be collected from the respondent. In our application, we found that some respondents could not provide accurate information about whether they owned a fishing license or not. Inaccurate information about domain membership causes domain misclassfication error, which affects the properties of estimators of means and totals. In this section, we examine the effect of misclassification error on the bias and variance of the Hartley estimator.

Frame A domain misclassification error can occur in two ways. The first is that a respondent sampled from frame A who is in domain ab identifies himself or herself as being in domain a. It also can occur in the other direction; i.e., a respondent sampled from frame A who is in domain a reports that he or she is in ab. We refer to the domain reported by the respondent as being the perceived or reported domain to distinguish it from their true domain. Thus each respondent belongs to one of four subgroups based on membership in the intersection of true and perceived domain identities. In this paper, we use a superscripted asterisk (*) on the domain name to indicate the perceived domain, so the four subgroups are denoted as $a \cap a^*$, $ab \cap ab^*$, $a \cap ab^*$, and $ab \cap a^*$. These labels are also used as subscripts for parameters to indicate the subset of the population to which it applies. For example, let frame A and frame B denote the CHTS and angler registry frames, respectively. Then a person living in a household on the CHTS frame and who has no fishing license, but reports that he does, belongs to subgroup $a \cap ab^*$. The universe of all such persons on frame A would be denoted as $U_{a \cap ab^*}$, and the size of the population subgroup, and the total and mean number of fishing trips for this subgroup would be denoted as $N_{a \cap ab^*}$, $Y_{a \cap ab^*}$, and $\bar{Y}_{a \cap ab^*}$, respectively. If such a person actually did have a fishing license, then he would be in subgroup $ab \cap ab^*$. In our study, the analyst would not be able to distinguish between the domains of these two respondents, and would place them both in the perceived overlap domain ab^*. The sampling weight appropriate to the actual but unobserved overlap domain (ab) would be applied to units belonging to $U_{ab^*} = U_{a \cap ab^*} \cup U_{ab \cap ab^*}$

instead of to $U_{ab} = U_{ab \cap a^*} \cup U_{ab \cap ab^*}$, and estimated totals for units in U_{ab^*} would replace those from U_{ab} in (2.2), causing the bias we investigate in this paper.

The domain misclassification error just described for frame A can occur in either frame in some studies. In our study this was not a problem because the license frame included addresses, so we knew whether or not the registered angler would appear on the state residential frame. However, we allow for the more general case, so the notation defined for frame A is extended to frame B. A complication this causes is that we must distinguish the cases in the perceived overlap domain (ab^*) according to the frame from which the unit originates. This is because the units in the perceived overlap may differ depending on the originating frame. For example, in our study, a person sampled from the CHTS can report that they do not have a license when they actually do, so would not be in ab^*. However, if the same person were to be sampled from the license frame, they would be correctly perceived to be in the overlap ab^*. Therefore, when confusion can occur, we extend the notation of the perceived overlap domain to indicate the frame from which the unit originates, as $ab^*(A)$ and $ab^*(B)$. For example, in our study, $\bar{Y}_{ab^*(A)}$ is the mean number of trips for all those in the state (frame A) who would report that they have a license when asked, which includes some who do not. However $\bar{Y}_{ab^*(B)}$ would be the mean number of trips for only those in the true domain ab, because our frame B did not suffer from domain misclassification errors.

Now we rewrite $\hat{Y}_H$ in new notation to make it easier to derive its properties when misclassification occurs. First we consider the case when no misclassificaton error occurs. For this we define the *true* domain indicator $\delta_i^A(a)$ which take a value of 1 when unit i from frame A is in domain a and 0 otherwise. Since every unit is in one true domain or the other, the indicator that unit i is in domain ab is $1 - \delta_i^A(a)$. $\delta_i^B(b)$ is similarly defined for frame B. When there is no misclassification error, $\hat{Y}_H$ can be written as:

$$\begin{aligned}\hat{Y}_H &= \sum_{i=1}^{N_A} I_i^A\, w_i^A\, \delta_i^A(a)\, y_i + \theta \sum_{i=1}^{N_A} I_i^A\, w_i^A \left(1 - \delta_i^A(a)\right) y_i \\ &\quad + \sum_{i=1}^{N_B} I_i^B\, w_i^B\, \delta_i^B(b)\, y_i + (1-\theta) \sum_{i=1}^{N_B} I_i^B\, w_i^B \left(1 - \delta_i^B(b)\right) y_i \\ &= \sum_{i=1}^{N_A} I_i^A\, w_i^A\, x_{Ai} + \sum_{i=1}^{N_B} I_i^B\, w_i^B\, x_{Bi}, \end{aligned} \tag{3.1}$$

where I_i^A and I_i^B are indicators that unit i from frame A or B belongs to sample s_A or s_B, $x_{Ai} = \delta_i^A(a)\, y_i + \theta\left(1 - \delta_i^A(a)\right) y_i$ and $x_{Bi} = \delta_i^B(b)\, y_i + (1-\theta)\left(1 - \delta_i^B(b)\right) y_i$.

For comparing the effect of domain misclassification on estimator properties, we restrict attention to the special case of a simple random sample design in each frame, with sample sizes n_A and n_B. $\hat{Y}_H$ is unbiased when there is no misclassification error. When the fpc is negligible, the variance of $\hat{Y}_H$ in this case is

$$V\left(\hat{Y}_H\right) = N_A^2\, S_{X_A}^2 / n_A + N_B^2\, S_{X_B}^2 / n_B, \tag{3.2}$$

where

$$S^2_{X_A} = \frac{N_a}{N_A}\left(S^2_a + \bar{Y}^2_a\right) + \theta^2 \frac{N_{ab}}{N_A}\left(S^2_{ab} + \bar{Y}^2_{ab}\right) - \left(\frac{N_a}{N_A}\bar{Y}_a + \theta \frac{N_{ab}}{N_A}\bar{Y}_{ab}\right)^2,$$

$$S^2_{X_B} = \frac{N_b}{N_B}\left(S^2_b + \bar{Y}^2_b\right) + (1-\theta)^2 \frac{N_{ab}}{N_B}\left(S^2_{ab} + \bar{Y}^2_{ab}\right) - \left[\frac{N_b}{N_B}\bar{Y}_b + (1-\theta)\frac{N_{ab}}{N_B}\bar{Y}_{ab}\right]^2,$$

are the population variances of x_{Ai} and x_{Bi}. Here and subsequently S^2_d and $\bar{Y}_d$ denote the population variance and mean of y for domain d, whether d denotes the true domain ($d = a, b,$ or ab) or the perceived one ($d = a^*, b^*, ab^*(A),$ or $ab^*(B)$).

When domain misclassification occurs, we require notation for perceived domain membership indicators. For these, we define $\eta^A_i(a)$ to be 1 when unit i from frame A is in domain a^* and 0 otherwise. Because each unit is in one perceived domain or the other, the indicator that unit i is in ab^* is $1 - \eta^A_i(a)$. $\eta^B_i(b)$ is similarly defined for frame B. Then Hartley's estimator becomes

$$\begin{aligned}
\hat{Y}^*_H &= \hat{Y}_{a^*} + \theta\hat{Y}^A_{ab^*} + (1-\theta)\hat{Y}^B_{ab^*} + \hat{Y}_{b^*} \\
&= \sum_{i=1}^{N_A} I^A_i\, w^A_i\, \eta^A_i(a)\, y_i + \theta \sum_{i=1}^{N_A} I^A_i\, w^A_i \left(1 - \eta^A_i(a)\right) y_i \\
&\quad + \sum_{i=1}^{N_B} I^B_i\, w^B_i\, \eta^B_i(b)\, y_i + (1-\theta)\sum_{i=1}^{N_B} I^B_i\, w^B_i \left(1 - \eta^B_i(b)\right) y_i \\
&= \sum_{i=1}^{N_A} I^A_i\, w^A_i\, x^*_{Ai} + \sum_{i=1}^{N_B} I^B_i\, w^B_i\, x^*_{Bi}, \qquad (3.3)
\end{aligned}$$

where $x^*_{Ai} = \eta^A_i(a)\, y_i + \theta\left(1 - \eta^A_i(a)\right) y_i$ and $x^*_{Bi} = \eta^B_i(b)\, y_i + (1-\theta)\left(1 - \eta^B_i(b)\right) y_i$. The bias of $\hat{Y}^*_H$ is then

$$\begin{aligned}
\text{Bias} &= E\left(\hat{Y}^*_H\right) - Y \\
&= (1-\theta)\sum_{i=1}^{N_A}\left(\eta^A_i(a) - \delta^A_i(a)\right) y_i + \theta\sum_{i=1}^{N_B}\left(\eta^B_i(b) - \delta^B_i(b)\right) y_i \\
&= (1-\theta)\left(N_{a^*}\bar{Y}_{a^*} - N_a\bar{Y}_a\right) + \theta\left(N_{b^*}\bar{Y}_{b^*} - N_b\bar{Y}_b\right). \qquad (3.4)
\end{aligned}$$

Note that $N_{a^*}\bar{Y}_{a^*} - N_a\bar{Y}_a = Y_{a^*} - Y_a = \left(Y_{a\cap a^*} + Y_{ab\cap a^*}\right) - \left(Y_{a\cap a^*} + Y_{a\cap ab^*(A)}\right) = Y_{ab\cap a^*} - Y_{a\cap ab^*(A)}$. So the first term in the bias expression can be positive or negative and large or small, depending on the relative number in the population who wrongly perceive that they are or are not in the frame overlap, and their response means. The same is true of the second term. In theory the two could cancel each other out even if errors occurred in both directions, but of course that is unlikely.

The variance expression for $\hat{Y}^*_H$ is similar to that of $V\left(\hat{Y}_H\right)$:

$$V\left(\hat{Y}^*_H\right) = N^2_A\, S^2_{x^*_A} / n_A + N^2_B\, S^2_{x^*_B} / n_B, \qquad (3.5)$$

where

$$S^2_{x^*_A} = \frac{N_{a^*}}{N_A}\left(S^2_{a^*} + \bar{Y}^2_{a^*}\right) + \theta^2 \frac{N_{ab^*(A)}}{N_A}\left(S^2_{ab^*(A)} + \bar{Y}^2_{ab^*(A)}\right) - \left(\frac{N_{a^*}}{N_A}\bar{Y}_{a^*} + \theta\frac{N_{ab^*(A)}}{N_A}\bar{Y}_{ab^*(A)}\right)^2,$$

and

$$S^2_{x^*_B} = \frac{N_{b^*}}{N_B}\left(S^2_{b^*} + \bar{Y}^2_{b^*}\right) + (1-\theta)^2 \frac{N_{ab^*(B)}}{N_B}\left(S^2_{ab^*(B)} + \bar{Y}^2_{ab^*(B)}\right) - \left(\frac{N_{b^*}}{N_B}\bar{Y}_{b^*} + (1-\theta)\frac{N_{ab^*(B)}}{N_B}\bar{Y}_{ab^*(B)}\right)^2$$

denote the variances of x^*_{Ai} and x^*_{Bi}. The perceived overlap domain ab^* is further denoted by (A) or (B), because the same unit in ab may be perceived to belong to different domains if reached from different frames. The impact of misclassification on the mean square error (MSE) of Hartley's estimator will be discussed further in Section 3.3. Obiously, the bias of $\hat{Y}^*_H$ is 0 and $V\left(\hat{Y}^*_H\right) = V\left(\hat{Y}_H\right)$ if the actual and perceived domains are identical.

4 Bias correction for misclassification error

Lohr (2011, Section 6) addressed the same problem we are considering; i.e., how to adjust for domain misclassification in dual frame survey designs. She proposed an adjustment to Hartley's estimator to mitigate the bias induced from misclassification. We review Lohr's method and then compare it to our proposed alternative. We observe that in order for Lohr's method to produce an unbiased estimator of mean or total for Y, the units in the same domain must have equal means, regardless of their perceived domain membership. This assumption was not valid in our angler survey; rather, we found that fishing avidity was related to the angler's perceived, rather than actual, license status. Our bias correction method was developed for this different perspective of accounting for the misclassification error. It assumes that the items in the same perceived domain must have equal means, regardless of their true domain membership. We show that the proposed method has smaller MSE than Lohr's method when both assumptions are true. When they are not, the choice should be made based on the appropriate assumption.

4.1 Lohr's misclassification bias correction method

Lohr's multinomial misclassification model assumes that units from each of the domains a, ab and b have their own known probabilities of being perceived to be in one of the three domains. These probabilities are used to adjust the estimators to remove the bias arising from misclassification. In practice, these probabilities are usually unknown and must be estimated from a phase 2 sample, on which an expensive (and accurate) method is used to obtain domain membership. This approach will be discussed in Section 4.3.1. For now, these probabilities are assumed known.

We describe Lohr's method for a dual frame design, though she described her approach for multiple frames. We define random vectors composed of random variables for the three domains: $\mathbf{Y} = (Y_a, Y_{ab}, Y_b)'$, $\boldsymbol{\delta}_i^A = \left(\delta_i^A(a), \delta_i^A(ab), \delta_i^A(b)\right)'$ and $\boldsymbol{\delta}_i^B = \left(\delta_i^B(a), \delta_i^B(ab), \delta_i^B(b)\right)'$. If there is no misclassification error, $\hat{\mathbf{Y}}^A = \sum_{i \in s_A} \boldsymbol{\delta}_i^A w_i^A y_i$ and $\hat{\mathbf{Y}}^B = \sum_{i \in s_B} \boldsymbol{\delta}_i^B w_i^B y_i$ are vectors of unbiased estimators of domain totals from frames A and B. Then Hartley's estimator with fixed θ can be written as $\hat{Y}_H = \left(\mathbf{m}^A\right)' \hat{\mathbf{Y}}^A + \left(\mathbf{m}^B\right)' \hat{\mathbf{Y}}^B$ where $\mathbf{m}^A = (1, \theta, 0)'$ and $\mathbf{m}^B = (0, 1-\theta, 1)'$.

Now suppose misclassification error occurs. Let $\boldsymbol{\eta}_i^A = \left(\eta_i^A(a), \eta_i^A(ab), \eta_i^A(b)\right)'$ denote the vector of perceived domain membership indicators for unit i from frame A. This vector can be written as $\boldsymbol{\eta}_i^A = \left(\mathbf{M}_i^A\right)' \boldsymbol{\delta}_i^A$, where $\mathbf{M}_i^A$ is a 3×3 matrix containing a 1 in position (d_2, d_1) if unit i in domain d_2 was perceived to be in domain d_1, and 0 elsewhere. When the perceived rather than actual domain membership information is used, $\hat{Y}^A$ becomes

$$\hat{\mathbf{Y}}^{A^*} = \sum_{i \in s_A} \boldsymbol{\eta}_i^A w_i^A y_i = \sum_{i \in s_A} \left(\mathbf{M}_i^A\right)' \boldsymbol{\delta}_i^A w_i^A y_i.$$

Define the misclassification probability matrix for frame A as:

$$\boldsymbol{\Phi}^A = \begin{pmatrix} p_{a^*|a} & p_{ab^*(A)|a} & 0 \\ p_{a^*|ab} & p_{ab^*(A)|ab} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{4.1}$$

where $p_{d_1^*|d_2}$ is the probability that a unit in domain d_2 is perceived to be in domain d_1. $\hat{Y}^{A^*}$ will typically be biased for the domain totals of frame A when the off-diagonal elements are non-zero since $E\left(\hat{\mathbf{Y}}^{A^*}\right) = \left(\boldsymbol{\Phi}^A\right)' \mathbf{Y} \neq E\left(\hat{Y}^A\right)$.

To correct the bias, Lohr (2011) proposed the weight adjustment vector, $\tilde{\mathbf{m}}^A = \left(\boldsymbol{\Phi}^A\right)^+ \mathbf{m}^A$ where $\left(\boldsymbol{\Phi}^A\right)^+$ is the Moore-Penrose inverse of $\boldsymbol{\Phi}^A$. Although not explicitly stated, this method requires for unbiasedness the assumption that the true domain membership determines the mean; i.e., that

$$\bar{Y}_{a \cap ab^*} = \bar{Y}_{a \cap a^*}, \bar{Y}_{ab \cap ab^*(A)} = \bar{Y}_{ab \cap a^*} \tag{4.2}$$

Then $E\left(\left(\tilde{\mathbf{m}}^A\right)' \hat{\mathbf{Y}}^{A^*}\right) = \left(\left(\boldsymbol{\Phi}^A\right)^+ \mathbf{m}^A\right)' \left(\boldsymbol{\Phi}^A\right)' \mathbf{Y} = \left(\mathbf{m}^A\right)' \mathbf{Y}$. After implementing a similar equal means assumption, misclassification matrix and adjustment for frame B, we have Lohr's bias-adjusted estimator:

$$\begin{aligned} \hat{Y}_L &= \left(\tilde{\mathbf{m}}^A\right)' \hat{\mathbf{Y}}^{A^*} + \left(\tilde{\mathbf{m}}^B\right)' \hat{\mathbf{Y}}^{B^*} \\ &= \hat{Y}_{a^*} \phi_{11}^{A+} + \hat{Y}_{ab^*(A)} \phi_{21}^{A+} + \theta\left(\hat{Y}_{a^*} \phi_{12}^{A+} + \hat{Y}_{ab^*(A)} \phi_{22}^{A+}\right) \\ &\quad + (1-\theta)\left(\hat{Y}_{b^*} \phi_{32}^{B+} + \hat{Y}_{ab^*(B)} \phi_{22}^{B+}\right) + \hat{Y}_{b^*} \phi_{33}^{B+} + \hat{Y}_{ab^*(B)} \phi_{23}^{B+}, \end{aligned} \tag{4.3}$$

where ϕ_{ij}^{A+} is the (i, j) element of $\left(\boldsymbol{\Phi}^A\right)^+$, and ϕ_{ij}^{B+} is similarly defined. In the special case of dual frame designs that we are considering, the components of the pseudoinverse matrices can easily be written explicitly if $p_{ab^*(A)|a} + p_{a^*|ab} \neq 1$ and $p_{ab^*(B)|b} + p_{b^*|ab} \neq 1$ as:

$$\phi_{11}^{A+} = \frac{p_{ab^*(A)|ab}}{D_A},\ \phi_{21}^{A+} = \frac{-p_{a^*|ab}}{D_A},\ \phi_{12}^{A+} = \frac{-p_{ab^*(A)|a}}{D_A},$$

$$\phi_{22}^{A+} = \frac{p_{a^*|a}}{D_A},\ \phi_{33}^{B+} = \frac{p_{ab^*(B)|ab}}{D_B},\ \phi_{23}^{B+} = \frac{-p_{b^*|ab}}{D_B},$$

$$\phi_{32}^{B+} = \frac{-p_{ab^*(B)|b}}{D_B},\ \phi_{22}^{B+} = \frac{p_{b^*|b}}{D_B}, \tag{4.4}$$

where $D_A = 1 - \left(p_{ab^*(A)|a} + p_{a^*|ab}\right)$ and $D_B = 1 - \left(p_{ab^*(B)|b} + p_{b^*|ab}\right)$. If the misclassification probabilities were known, this estimator could be computed and would be unbiased, as long as (4.2) and similar assumptions for frame B means hold. Its variance can be derived by the same technique used to obtain equation (3.2). See Lin (2014), pages 28-29 for the variance expression.

4.2 An alternative misclassification bias correction procedure

Lohr's estimator eliminates the bias of Hartley's estimator under the assumption that the true domain membership determines the mean of the variable of interest. Because this assumption appeared not to hold for our angler survey data, we developed an alternative bias correction method requiring a more suitable set of assumptions for our application.

Recall that $Y_{a^*}, Y_{ab^*(A)}, Y_{ab^*(B)}, Y_{b^*}$ denote the population totals of the perceived domains, where the notation reflects that perceived domain membership in the frame overlap may depend on the frame from which the unit was sampled. The perceived domain totals and sizes can be decomposed by actual domain labels: e.g., $Y_{a^*} = Y_{a\cap a^*} + Y_{ab\cap a^*}$ and $N_{a^*} = N_{a\cap a^*} + N_{ab\cap a^*}$. We parameterize our model by defining the misclassification probabilities in the reverse form from Lohr's. That is, let $p_{d_1|d_2^*}$ denote the probability that an observation perceived to be in domain d_2 is actually in domain d_1 and define the matrix of misclassification probabilities from frame A as

$$\mathbf{\Lambda}^A = \begin{pmatrix} p_{a|a^*} & p_{ab|a^*} & 0 \\ p_{a|ab^*(A)} & p_{ab|ab^*(A)} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In our development, the equal means assumptions (4.2) are replaced by assumptions that the means are determined by perceived domain membership;

$$\bar{Y}_{a\cap a^*} = \bar{Y}_{ab\cap a^*},\ \bar{Y}_{a\cap ab^*(A)} = \bar{Y}_{ab\cap ab^*(A)},\ \bar{Y}_{b\cap b^*} = \bar{Y}_{ab\cap b^*},\ \bar{Y}_{b\cap ab^*(B)} = \bar{Y}_{ab\cap ab^*(B)} \tag{4.5}$$

When these assumptions hold, we have

$$Y_a = Y_{a^*}\, p_{a|a^*} + Y_{ab^*(A)}\, p_{a|ab^*(A)},\ Y_{ab} = Y_{a^*}\left(1 - p_{a|a^*}\right) + Y_{ab^*(A)}\left(1 - p_{a|ab^*(A)}\right), \tag{4.6}$$

with similar expressions for domain totals for frame B. Then a bias-corrected estimator is

$$\begin{aligned}\hat{Y}_{\mathrm{BC}} &= \left(\boldsymbol{\Lambda}^{A}\,\mathbf{m}^{A}\right)'\hat{Y}^{A^*} + \left(\boldsymbol{\Lambda}^{B}\,\mathbf{m}^{B}\right)'\hat{Y}^{B^*} \\ &= \hat{Y}_{a^*}\,p_{a|a^*} + \hat{Y}_{ab^*(A)}\,p_{a|ab^*(A)} + \theta\left(\hat{Y}_{a^*}\left(1-p_{a|a^*}\right) + \hat{Y}_{ab^*(A)}\left(1-p_{a|ab^*(A)}\right)\right) \\ &\quad + (1-\theta)\left(\hat{Y}_{b^*}\left(1-p_{b|b^*}\right) + \hat{Y}_{ab^*(B)}\left(1-p_{b|ab^*(B)}\right)\right) + \hat{Y}_{b^*}\,p_{b|b^*} + \hat{Y}_{ab^*(B)}\,p_{b|ab^*(B).}\end{aligned} \tag{4.7}$$

$\hat{Y}_{\mathrm{BC}}$ is unbiased when the equal means assumptions (4.5) hold. Its variance can be calculated using the same technique as that used to obtain equation (3.2). See Lin (2014), pages 33-34 for the variance expression.

4.3 Comparison of bias correction methods

In this section, we examine the performance of $\hat{Y}_L$, $\hat{Y}_{\mathrm{BC}}$ and their uncorrected counterpart $\hat{Y}_H$. We also examine bias-corrected estimators that can be used when the misclassification probabilities are not known. These can be constructed by replacing the known values of the probabilities in (4.3) and (4.7) with estimators of misclassification probabilities made from phase 2 subsamples. The subsample is selected from the original sample, using an accurate (and typically expensive) data collection method to ascertain true domain membership. Let $I_i^{A(2)}$ denote the phase 2 sample indicator for frame A, $\pi_i^{A(2)} = P\left\{I_i^{A(2)} = 1 \,\middle|\, I_i^{A} = 1\right\}$ its conditional selection probability, and $w_i^{A(2)} = 1/\pi_i^{A(2)}$ its phase 2 weight. Then ratio estimators of misclassification probabilities can be constructed; for example, $p_{a|a^*}$ can be estimated by

$$\hat{p}_{a|a^*}^{(2)} = \frac{\sum_{i=1}^{N_A} I_i^{A}\, I_i^{A(2)}\, w_i^{A}\, w_i^{A(2)}\, \delta_i\left(a^* \cap a\right)}{\sum_{i=1}^{N_A} I_i^{A}\, I_i^{A(2)}\, w_i^{A}\, w_i^{A(2)}\, \delta_i\left(a^*\right)} = \hat{N}_{a^*\cap a}^{(2)} \Big/ \hat{N}_{a^*}^{(2)}. \tag{4.8}$$

Lohr's estimator requires that the reverse conditional probabilities be estimated from the phase 2 sample; i.e., $\hat{p}_{a^*|a} = \hat{N}_{a^*\cap a}^{(2)} \Big/ \hat{N}_{a}^{(2)}$. Similar estimators can be constructed for the components of all four misclassification matrices. When these estimators are substituted for parameters in (4.3) and (4.7), the resulting bias-adjusted estimators are denoted by $\hat{\hat{Y}}_{\mathrm{BC}}$ and $\hat{\hat{Y}}_L$.

In this section, we present several comparisons among the five estimators. The first restricts the comparison of variances to only $\hat{Y}_H$, $\hat{Y}_L$, and $\hat{Y}_{\mathrm{BC}}$ for a simple scenario in which all domain means are equal, so that required assumptions about the means hold for both estimators. Its purpose is to illustrate the variance inflation that can occur in $\hat{Y}_L$. The second example partially replicates a simulation presented in Lohr (2011), but with the addition of $\hat{Y}_{\mathrm{BC}}$ and $\hat{\hat{Y}}_{\mathrm{BC}}$. It compares the mean squared error (MSE) of the estimators when both sets of equal mean assumptions are satisfied. We also examine the effect of a change of the phase 2 sample design from a simple random sample to one stratified by perceived domain. In example 3, we examine the robustness of both estimators to these assumptions by examining the simulated bias and MSE of the estimators when neither (4.2) nor (4.5) holds. In all cases, θ is fixed at 0.5.

4.3.1 Example 1: Variance inflation

We compared the variances of $\hat{Y}_{\mathrm{BC}}$ and $\hat{Y}_L$ with the MSE of $\hat{Y}_H$ when misclassification errors occur, but their rates are known. The populations are taken to be homogeneous with all domain means and variances constant: $\overline{Y}_a = \overline{Y}_{ab} = \overline{Y}_b = 2$ so that both equal mean assumptions hold and $\hat{Y}_{\mathrm{BC}}$ and $\hat{Y}_L$ are

unbiased. We set $S_a^2 = S_{ab}^2 = S_b^2 = 1$. For this example, the frames were taken to be small with substantial overlap: N_a = 3,000, N_{ab} = 2,000, N_b = 3,000. Simple random samples of sizes n_A = 100 and n_B = 50 were assumed. The MSE of the Hartley estimator was calculated from (3.4) and (3.5), while the variances of the bias-corrected estimators were computed using the expressions in Lin (2014, page 28 and 33). The following two cases were considered:

1. The misclassification probability for units in domain a varies from 0 to 1; no other misclassification errors exist.
2. The misclassification probability for units in domain ab (when sampled from frame A) varies from 0 to 1; no other misclassification errors exist.

The two panels of Figure 4.1 display the MSE's of the three estimators as functions of the two misclassification probabilities. For each condition, $\hat{Y}_{\text{BC}}$ has smaller MSE than that of $\hat{Y}_L$ over the range of misclassification probabilities, though the two estimators have very similar performance when the probabilities are small. The MSE of $\hat{Y}_{\text{BC}}$ varies little across the range, but the MSE of $\hat{Y}_L$ increases without bound as the probability approaches 1. This occurs because the components of the pseudoinverse (4.4) become large as the (single) misclassification approaches 1, which inflates the coefficients of subdomain estimates in (4.3).

In practice, if misclassification probabilities were known and exceeded $1/2$, the domain identification process would simply be reversed, so this scenario is not of practical importance. However, we will see in the next example that when misclassification probabilities are not known and must be estimated, the performance of $\hat{Y}_L$ is sensitive to the quality of their phase 2 probability estimators. This is because even if misclassification probabilities are small, their estimates can be large, especially when the phase 2 design is inefficient, and the same variance inflation can occur.

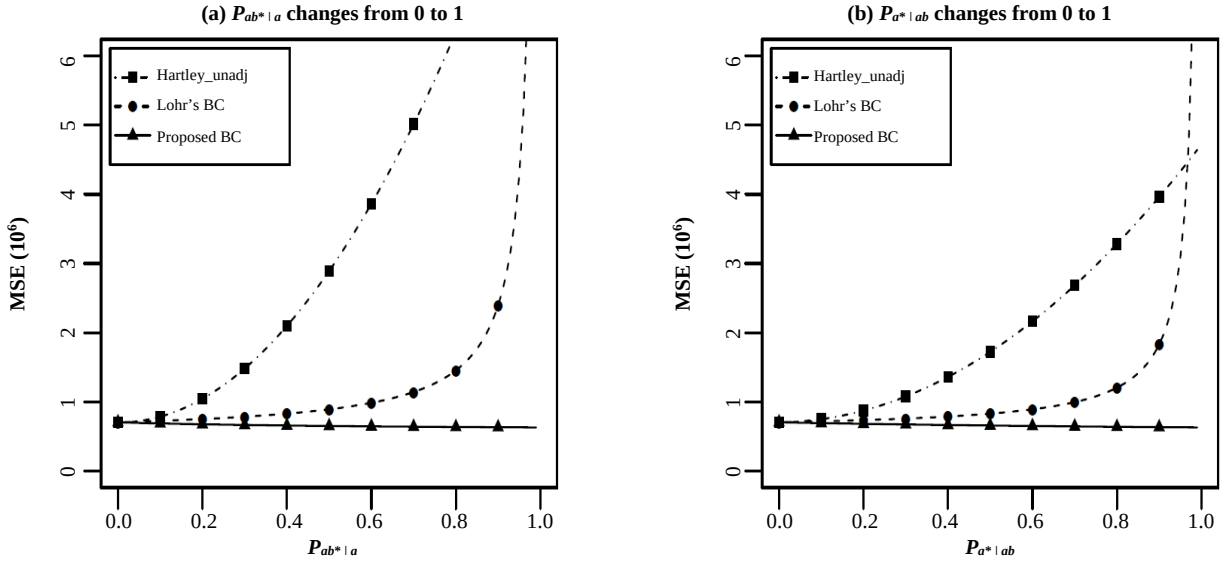


Figure 4.1 Comparison of MSE of $\hat{Y}_H$, $\hat{Y}_L$, and $\hat{Y}_{\text{BC}}$ over range of a single misclassification probability.

4.3.2 Example 2: Simulated relative efficiency of the five estimators

The second example uses simulation to compare the performance of $\hat{Y}_H$, $\hat{Y}_L$, $\hat{\hat{Y}}_L$, $\hat{Y}_{\text{BC}}$ and $\hat{\hat{Y}}_{\text{BC}}$ for a variety of misclassification patterns, first when both domain mean assumptions hold, and then when neither holds. The domain sizes chosen for this example were $N_a = 48{,}000$, $N_{ab} = 2{,}000$, and $N_b = 3{,}000$, which were meant to mimic the motivating angler survey, in which one frame (the address frame) was much larger than the other (the registration frame). The population was generated using the model $y_i \sim x_i + 1$, for $i = 1, \ldots, N$, with $x_i \sim$ Poisson(1), regardless of the domain. Thus the means and variances for the population were approximately the same as those in the first example. We simulated selection of simple random samples from each frame, with sample sizes of $n_A = 100$ and $n_B = 50$. For the phase 2 samples, we chose a simple random sample (SRS), with four sample sizes that varied from 20% to 80% of the phase 1 sample, resulting in frame A subsample sizes (m_A) between 20 and 80 and frame B subsample sizes (m_B) between 10 and 40. If any replicate had a phase 2 sample with fewer than 2 units from one of the subdomains, we assumed no misclassification for that domain; i.e., the estimators were constructed as if there was no misclassification error for that subdomain. 16 misclassification patterns were simulated, which were all pairs of the following four patterns for each frame:

1. Misclassification patterns for frame A (MPA)
   a) $p_{a^*|a} = 1,\ p_{ab^*(A)|ab} = 1$
   b) $p_{a^*|a} = 0.9,\ p_{ab^*(A)|a} = 0.1,\ p_{ab^*(A)|ab} = 1$
   c) $p_{a^*|a} = 0.9,\ p_{ab^*(A)|a} = 0.1,\ p_{ab^*(A)|ab} = 0.9,\ p_{a^*|ab} = 0.1$
   d) $p_{a^*|a} = 1,\ p_{ab^*(A)|ab} = 0.9,\ p_{a^*|ab} = 0.1$
2. Misclassification patterns for frame B (MPB)
   a) $p_{b^*|b} = 1,\ p_{ab^*(B)|ab} = 1$
   b) $p_{b^*|b} = 0.8,\ p_{ab^*(B)|b} = 0.2,\ p_{ab^*(B)|ab} = 1$
   c) $p_{b^*|b} = 0.8,\ p_{ab^*(B)|b} = 0.2,\ p_{ab^*(B)|ab} = 0.8,\ p_{b^*|ab} = 0.2$
   d) $p_{b^*|b} = 1,\ p_{ab^*(B)|ab} = 0.8,\ p_{b^*|ab} = 0.2$

Thus the simulation examined 64 settings (16 misclassification patterns × 4 phase 2 sample sizes),with 10,000 phase 1 samples generated under each setting and the appropriate estimates computed. The empirical MSE was calculated as the average squared deviation of each estimate from the true Y over the 10,000 replicates. The results are summarized in Table 4.1, Figure 4.2 and 4.3.

Table 4.1 shows the results for all 16 misclassification patterns for the 40% phase 2 sampling rate ($m_A = 40$ and $m_B = 20$). The conclusions we draw here were consistent for all sample sizes. We see that $\hat{Y}_{\text{BC}}$ and $\hat{\hat{Y}}_{\text{BC}}$ are less variable than $\hat{Y}_L$ and $\hat{\hat{Y}}_L$ for all misclassification patterns except when no error is present. Second, if misclassification occurs only in small domains (pattern a or d from frame A), there is little advantage to using bias correction, since $\hat{Y}_H$ performs about as well or better than the bias-corrected estimators. This suggests that bias correction may not be advantageous unless misclassification affects a

large portion of the population. Finally, we observe that $\hat{\hat{Y}}_L$ shows worse performance when misclassification pattern c holds in frame A. Now we examine this effect more closely.

Table 4.1
MSE $(\times 10^7)$ for dual frame estimators

		Known Prob.			Estimated Prob.	
MPA	MPB	$\hat{Y}_{\text{BC}}$	$\hat{Y}_L$	$\hat{Y}_H$	$\hat{\hat{Y}}_{\text{BC}}$	$\hat{\hat{Y}}_L$
a	a	2.53	2.53	2.53	2.53	2.53
b	a	2.45	2.86	4.77	2.66	3.00
c	a	2.44	2.96	4.60	2.66	3.80
d	a	2.52	2.54	2.53	2.55	2.72
a	b	2.53	2.55	2.57	2.54	2.55
b	b	2.45	2.88	5.37	2.66	3.02
c	b	2.44	2.98	5.18	2.67	3.82
d	b	2.52	2.56	2.55	2.56	2.74
a	c	2.53	2.57	2.54	2.55	2.63
b	c	2.45	2.90	4.99	2.67	3.10
c	c	2.44	3.00	4.81	2.67	3.90
d	c	2.52	2.59	2.54	2.56	2.81
a	d	2.53	2.54	2.56	2.54	2.54
b	d	2.45	2.87	4.44	2.66	3.01
c	d	2.44	2.97	4.28	2.66	3.81
d	d	2.52	2.55	2.57	2.55	2.73

Figure 4.2 shows the ratio of the MSE's of the probability-unknown to the probability-known bias-corrected estimators, where panel (a) shows $\text{MSE}(\hat{\hat{Y}}_{\text{BC}})\big/\text{MSE}(\hat{Y}_{\text{BC}})$ and (b) shows MSE $\text{MSE}(\hat{\hat{Y}}_L)\big/\text{MSE}(\hat{Y}_L)$. The four lines display the ratio for the four phase 2 sample sizes for the 16 misclassification patterns arrayed on the x-axis. The ratio's distance above 1 measures the variance penalty incurred from estimating the misclassification probabilities. The most significant feature of the figure is the large penalty suffered by $\hat{\hat{Y}}_L$ under frame A misclassification patterns c and d for small sample sizes, and the relative lack of this effect for $\hat{\hat{Y}}_{\text{BC}}$. Both patterns have non-zero misclassification error for the small overlap domain. Therefore, the phase 1 sample often has few units available from which to estimate $p_{a^*|ab}$, producing a noisy estimate of the misclassification probability. So even though the actual misclassification probability is not close to 1 under patterns c and d ($p_{a^*|ab} = 0.1$), the fact that it is not known and must be estimated from little data means that some samples produce extreme estimates, inflating its variance, as we noted in Figure 4.1.

We see from panel (a) of Figure 4.2 that this effect is also present to a lesser degree for $\hat{\hat{Y}}_{\text{BC}}$. The smallest phase 2 sample ($m_A = 20$, $m_B = 10$) produces disproportionately worse performance than the next larger sample size. The same cause is at play; a SRS from the phase 1 sample can provide too few units in domain $ab^*(A)$. However, improving the efficiency of the phase 2 design for estimation of $p_{a|ab^*(A)}$, is more

straightforward than improving it for $p_{a^*|ab}$. This is because the perceived domains are observable for all units in the phase 1 sample, while the actual domains are not, so the analyst can control the sample size for the former, but not the latter, by stratification.

To investigate how much advantage this would provide, we conducted an additional simulation to examine the performance of $\hat{\hat{Y}}_{\mathrm{BC}}$ when the phase 2 sample has a stratified design. We chose a design with equal sample sizes in each of the perceived domain strata. So for example, in the 20% sampling rate setting, we selected $m_A/2 = 20/2 = 10$ observations from domains a^* and $ab^*(A)$ from the phase 1 sample. If fewer than $m_A/2$ (or $m_B/2$) but more than 1 unit was available, then all of them were selected and the remainder chosen from the other domain. This simulation used the same 64 settings as the previous one, with the only difference being a stratified design at phase 2 rather than a SRS.

Results from this simulation are shown in Figure 4.3, which displays a ratio similar to a design effect (but for MSE rather than variance) of the phase 2 design. The graph shows that in some settings, meaningful improvement of the MSE of $\hat{\hat{Y}}_{\mathrm{BC}}$ is possible by stratifying the phase 2 sample. The gain is especially important for the small sample size.

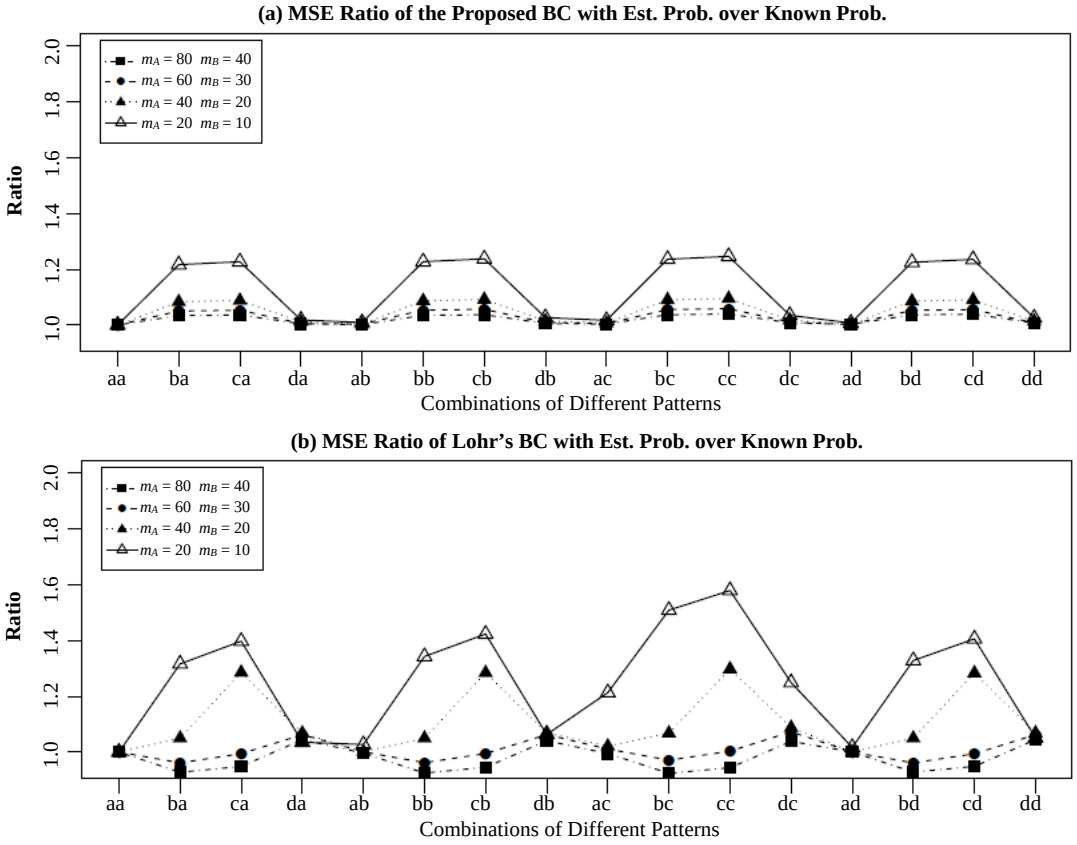


Figure 4.2 Ratio of simulated MSE's for estimators that assume misclassification probabilities are unknown: (a) MSE($\hat{\hat{Y}}_{\mathrm{BC}}$) / MSE($\hat{Y}_{\mathrm{BC}}$) and (b) MSE($\hat{\hat{Y}}_L$) / MSE($\hat{Y}_L$).

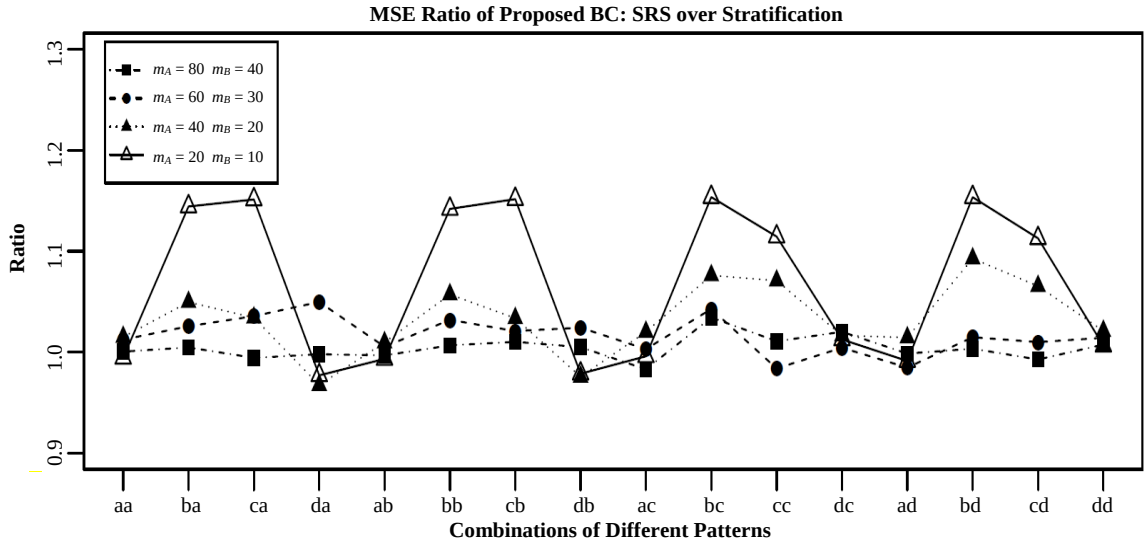


Figure 4.3 Ratio of simulated MSE's for estimators with two phase 2 designs: $MSE_{srs}(\hat{\hat{Y}}_{BC}) / MSE_{str}(\hat{\hat{Y}}_{BC})$.

4.3.3 Example 3: Robustness to violation of mean assumptions

Both bias-corrected estimators require equal means assumptions to guarantee unbiasedness in the known misclassification probabilities case, and those assumptions differ. Lin (2014, pages 43-47) derived expressions for the bias of $\hat{Y}_L$ and $\hat{Y}_{BC}$ when the assumptions do not hold. In this example, we use simulation to investigate the size of the bias for $\hat{\hat{Y}}_L$ and $\hat{\hat{Y}}_{BC}$ as well as $\hat{Y}_L$ and $\hat{Y}_{BC}$. The simulation settings in this example are similar to those of the previous one, except for the means. We also considered only 3 of the 16 misclassification patterns, which are those with domain misclassification in only the large (first-listed) frame: *ba*, *ca*, and *da*.

We simulated the populations so that one of the four subdomain means was about 3 ($y \sim x_i + 2$, with $x_i \sim$ Poisson(1)) while the others remained at about 2 ($y \sim x_i + 1$). Therefore, in each case, the equal means assumptions (4.2) and (4.5) were violated for both bias-corrected estimators. The empirical bias, variance, and MSE for each of the five estimators was computed from 10,000 replicated samples for each of 48 settings (3 misclassification patterns $\times$ 4 equal-mean violation patterns $\times$ 4 phase 2 sample sizes). We present results for a representative subset of these settings in Figure 4.4.

Figure 4.4 displays boxplots of the simulated estimates from $\hat{Y}_H$, $\hat{\hat{Y}}_{BC}$, and $\hat{\hat{Y}}_L$ for one of the phase 2 sample sizes (m_A = 40 and m_B = 20). The columns of the plot show results for the four equal mean assumption violations, and the rows show the results for three misclassfication patterns. The horizontal line in each plot shows the true population total. The figure shows that both $\hat{\hat{Y}}_{BC}$ and $\hat{\hat{Y}}_L$ have smaller bias than $\hat{Y}_H$ for all settings, so that making the wrong assumption about the means is better for bias than assuming that no misclassification error exists. For the settings considered, it appears that $\hat{\hat{Y}}_{BC}$ in particular is not too sensitive to small violations of the equal mean assumptions.

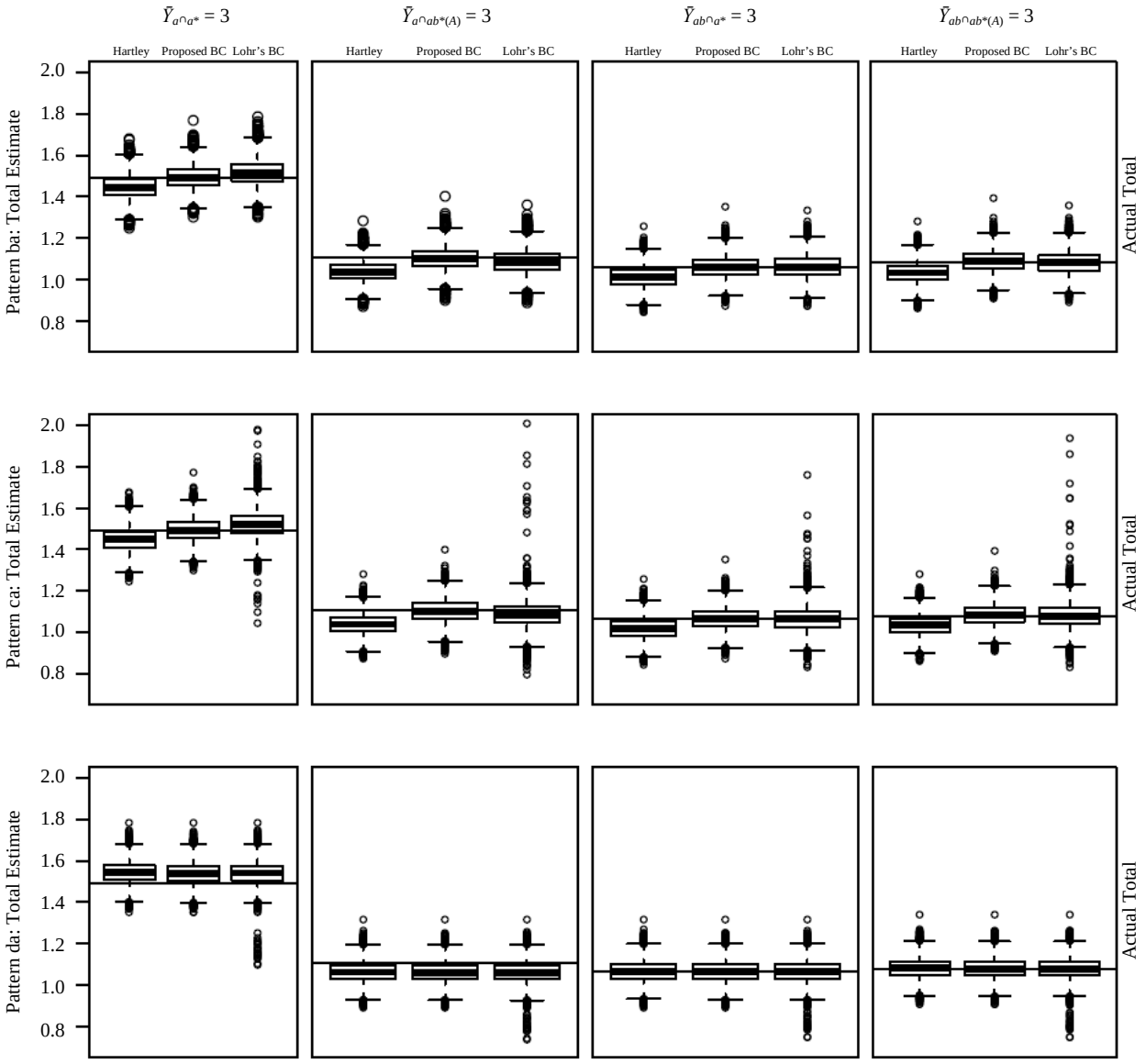


Figure 4.4 Estimation of total $(\times 10^5)$ under different violation of equal mean assumption.

5 Inference for $\hat{\hat{Y}}_{\mathrm{BC}}$

Lohr (2011, Section 4) noted that inference for dual frame estimators with non-random constructed weights is straightforward using standard survey software. This is true for both $\hat{Y}_L$ and $\hat{Y}_{\mathrm{BC}}$. For $\hat{Y}_{\mathrm{BC}}$, the constructed weights for units in the four domains are $\tilde{w}_i^A = w_i^A \left(p_{a|a^*} + \theta\left(1 - p_{a|a^*}\right)\right)$ and $\tilde{w}_i^A = w_i^A \left(p_{a|ab^*(A)} + \theta\left(1 - p_{a|ab^*(A)}\right)\right)$ for units in domains a^* and ab^* sampled from frame A, and $\tilde{w}_i^B = w_i^B \left((1-\theta)\left(1 - p_{b|b^*}\right) + p_{b|b^*}\right)$ and $\tilde{w}_i^B = w_i^B \left((1-\theta)\left(1 - p_{b|ab^*(B)}\right) + p_{b|ab^*(B)}\right)$ for units in domains b^* and ab^* sampled from frame B. Then $\hat{Y}_{\mathrm{BC}}$ and its standard error can be calculated by providing to the software files containing data and weights from both frames. When the misclassification probabilities must

be estimated, as they are for $\hat{\hat{Y}}_{\mathrm{BC}}$, however, the variances are inflated, as illustrated in Figure 4.2, panel (a). In this case, linearization, jackknife, or bootstrap methods could be used to accommodate this increased variance.

Lin (2014, Section 5.2.2) produced an approximate variance expression for $\hat{\hat{Y}}_{\mathrm{BC}}$ for the case of a SRS at both phase 1 and phase 2, based on the linearization method. However, implementing the method requires special purpose coding, and would have to be adapted for complex designs at the two phases. Thus, we chose to investigate the accuracy of an approximate method that ignores the additional variability introduced by estimation of the misclassification probabilities and produces confidence intervals using standard survey software. The only pre-processing required is that the estimated misclassification probabilities must be computed from the phase 2 sample and used to replace the known values in the weight expressions above.

To test this method, we simulated a population with the characteristics of that of example 2 in the previous section. We examined a subset of the misclassification patterns considered there: *db*, *cb* and *ca*. The sample sizes for the phase 1 samples in this case were set to $n_A = 400$ and $n_B = 200$, and three phase 2 sampling rates were chosen, ranging from 5% to 20% (i.e., from $m_A = 20$ and $m_B = 10$ to $m_A = 80$ and $m_B = 40$), chosen as SRS and stratified designs. 10,000 replicates were generated under each setting. From each sample, misclassification probabilities were estimated and substituted in the weight expressions above. For comparison purposes, we also produced weights using the known misclassification probabilities to test the performance of confidence intervals based on $\hat{Y}_{\mathrm{BC}}$. All computation was done using R's *survey* package. The software-produced estimates of standard error and a 95% confidence interval for the population total (based on *survey*'s standard jackknife procedure) were obtained from each replicate.

The results are summarized in Table 5.1. The columns labeled *Sim.Var.* displays the variance of each estimator as computed from the 10,000 replicates, which is our best assesment of the true variances. The columns labeled *Est.Var.* show the average over the 10,000 replicates of the software-produced variance estimates of $\hat{Y}_{\mathrm{BC}}$ and $\hat{\hat{Y}}_{\mathrm{BC}}$. The columns labeled *Suc.Rate* shows the proportion of the replicates for which the confidence interval includes the true total. Panels (a) and (b) display results for the phase 2 SRS and stratified sample designs, respectively.

Table 5.1
Variance estimation and confidence interval coverage

			Known Prob.			Estimated Prob.								
						$m_A = 80$ $m_B = 40$			$m_A = 40$ $m_B = 20$			$m_A = 20$ $m_B = 10$		
		pattern	Sim. Var.	Est. Var.	Suc. Rate	Sim. Var.	Est. Var.	Suc. Rate	Sim. Var.	Est. Var.	Suc. Rate	Sim. Var.	Est. Var.	Suc. Rate
(a) SRS Method $(\times 10^6)$	SRS	db	6.42	6.45	0.95	6.60	6.45	0.94	7.00	6.45	0.94	7.14	6.46	0.94
		cb	6.36	6.26	0.94	7.42	6.28	0.93	8.46	6.31	0.91	11.6	6.34	0.85
		ca	6.26	6.27	0.95	7.20	6.29	0.93	8.41	6.33	0.91	11.8	6.37	0.85
(b) Stratification Method $(\times 10^6)$	STR	db	6.35	6.46	0.95	6.59	6.46	0.95	6.91	6.45	0.94	7.54	6.45	0.93
		cb	6.18	6.27	0.95	6.65	6.27	0.94	7.16	6.27	0.93	8.38	6.27	0.91
		ca	6.50	6.28	0.94	6.94	6.28	0.94	7.00	6.28	0.93	7.86	6.29	0.92

By comparing the *Sim.Var*. and *Est.Var*. columns, we show that the variance of $\hat{\hat{Y}}_{\mathrm{BC}}$ is underestimated, on average for all settings. As would be expected, the underestimation is worse for the smallest sample size and the inefficient (SRS) design. As a result, the confidence interval coverage is less than its nominal value for most settings. However the undercoverage is small (less than 5%) for all cases except the smallest sample size for a phase 2 SRS. Since the coverage of the confidence intervals based on $\hat{Y}_{\mathrm{BC}}$ held their nominal values, we can conclude that the undercoverage was due completely to the estimation of misclassification probabilities. We suggest that the additional variation added from estimation of misclassification probabilities can be safely ignored in inference if an efficient phase 2 design is used, unless the sample sizes are very small. Based on this simulation, if misclassification probabilities are estimated from at least 10 units in the each perceived domain, the coverage probabilities were no more than a few percent off. This can be accomplished with a smaller total sample size when the phase 2 sample is stratified, than when a SRS is used, so a stratified phase 2 design is recommended.

6 Example: Angler survey

We illustrate the use of the proposed bias correction procedure with its application to a dual frame mail survey of anglers in North Carolina (NC) in 2009. It was a pilot survey testing several changes to an ongoing program collecting recreational marine angler effort by NOAA, where effort is defined as the number of fishing trips during a specified time period. The two frames were a NC address frame and a license frame, which included the names and addresses of anglers who had any of several types of licenses. The target population of the pilot survey was recreational anglers who fished in NC saltwater, regardless of where they lived. The target time period of fishing was Wave 6 of 2009 (November - December). These two frames together had some undercoverage because unlicensed anglers whose home address was outside NC were not included in the union of the two frames.

6.1 Sample design

The address frame was obtained from the US Postal Service and covered all households in NC. The license frame included all persons listed on the NC database of licensed anglers as of the date of the license pull, which was several days before the mailing of the surveys. Independent samples of addresses were drawn from the two frames. Estimates were made of the fishing effort in NC during Wave 6, 2009 using the Hartley estimator with $\theta = 1/2$. The sample from the address frame was a complex sample, and itself involved two phases. The sample from the license frame required only one phase. In this application, units were at risk of misclassification only if they were chosen from the larger of the two frames, the address frame, since it was not known whether the persons in those households owned fishing licenses. The analysts did know that all persons selected from the license frame had a valid license during the wave and also knew whether or not they had an address in North Carolina, since the their address was available from the frame.

The sample design for the address frame was conducted in two phases. A random sample of 1,800 addresses was selected first, stratified by geography. The strata were defined as addresses in coastal and non-coastal counties of NC, with samples of 900 each. A screening questionnaire asked whether any household member fished in saltwater in the last 12 months. The second phase sample consisted of one randomly chosen angler from every household that reported fishing by any household member in the first phase. One additional angler was selected from households reporting more than one active adult angler. The reason for this two-phase construction was to avoid sending a lenthy questionnaire to non-angling households, in order to decrease cost and increase response rate.

The license frame was obtained from the NC license database. All individuals who were listed on the database on the day the frame was pulled and were licensed to fish during the target period (Wave 6, 2009) were included. The license frame was preprocessed to make it suitable for sampling. Multiple records with the same core data (name, date of birth and address) were deleted, as were anglers identified as being under 18. The license frame was divided into three strata: coastal, non-coastal, and out-of-state. The file was sorted by address, and a systematic sample of 450 anglers was selected from each stratum. Sampling in the license frame was conducted in a single phase, and used a questionnaire identical to the second phase questionnaire for the address frame sample. As in the address frame, a supplemental sample was selected from addresses with more than one licensed angler present on the frame.

The common questionnaire used for both frames included an item that asked whether the respondent had a NC marine recreational fishing license. This question was included to determine domain membership for those chosen from the address frame. However, analysts observed that some respondents from the license frame reported they did not have a license, which alerted them to the possible presence of domain misclassification error. As a result, an operation was undertaken to determine true domain membership for respondents from the address frame. We attempted to match 100% of the sampled addresses to the license frame. The last part of this process involved a human matcher trying to identify if a particular angler within a matched address appeared to be the licensed angler, based on available data from the license frame and survey responses. This was a time-consuming operation, which motivated the search for alternatives. The goal was to develop methods for the operational survey that allowed for determinination of true domain status for only a subset of the sample. However, since we did have access to the true domain status for the entire sample, we were able to examine misclassification probabilities and subdomain means, as well as to compare $\hat{\hat{Y}}_{BC}$'s results with an estimate made from "true" data.

Even though we observed that some on the license frame made errors concerning their license status, this did not cause a domain misclassification error for the license frame because the true license status was known. For the license frame, domain misclassfication could occur only if the in or out-of-state status of the household could not be determined accurately. It is possible such errors could occur. For example, if a household with an out-of-state address on the license frame were sampled, but it had a second in-state address that appeared on the address frame, then the domain assignment would be incorrect. However, the

incidence of such cases was believed to be small enough that it could be ignored, so we treated the misclassification probabilities as if they were known to be 0 for the units on the license frame in our analysis.

6.2 Sample analysis

The domain misclassification rates for the sample from the address frame are shown separately by stratum in Table 6.1. In this case domain ab^* contains those respondents from the address frame who report that they are licensed, while domain a^* contains those reporting they are not licensed. Anglers who reported that they are unlicensed have about a 5% error rate in both strata. Those who reported they are licensed have extremely high error rates, with those in non-coastal counties more likely to be wrong than right! We point out that the address frame respondents from which these estimates were reported are those who were in the second phase of the address frame sample. This means that they had screened in because their household had at least one person who had fished in the last 12 months. As a result, a very high fraction of these respondents were anglers compared to the general population.

Table 6.1
Misclassification rates calculated from full sample (Address frame, Wave 6, 2009)

	Proportion of those who report not being licensed who are $\left(\hat{p}_{ab\mid a^*}\right)$	Proportion of those who report being licensed who are not $\left(\hat{p}_{a\mid ab^*}\right)$
Coastal Stratum	0.04	0.46
Non-coastal Stratum	0.06	0.63

We also examined the equal means assumptions using data from the address frame sample. The estimated mean effort in each of the four categories of domain and perceived domain membership are shown in Table 6.2. The columns classify respondents into perceived domains, while the rows classify according to their true domain. The table shows that respondents' fishing behavior is consistent with what they report their license status to be rather than what their true status is. Thus, we believe that the equal means assumption of our proposed method is more reasonable for the angler survey data than Lohr's equal mean assumption.

Table 6.2
Estimated mean #of fishing trips (SE) by subdomain for Wave 6 2009 NC Address frame

$\bar{y}$ for subdomains	reported no license (a^*)	reported license (ab^*)
true no license (a)	0.34 (0.14)	0.88 (0.41)
true license (ab)	0.35 (0.46)	0.98 (0.24)

The sample data contained weights provided by the survey designers that accounted for the complex design and nonresponse adjustment. Because the domain misclassification probabilities differed by stratum, we adjusted the weights as described in Section 5 separately by stratum, using individual estimates of misclassification for each address frame domain. We assumed no domain misclassification for the license frame. Six estimates of effort were computed and are shown in Table 6.3:

1) Uncorrected Hartley estimator (labeled *Unadj.* in table): The perceived domain membership was used to estimate the total, using the Hartley estimator as in (3.3);
2) 20%, 40%, 100%-subsampled estimator: Units from each stratum of the phase 1 address frame sample were subsampled, and their true domains were used to estimate the misclassification probabilities. The weight adjustments were calculated based on the estimated misclassification probabilities;
3) Corrected Hartley estimator (*True*): The true domain membership ascertained from the matching operation was used to estimate the total number using the Hartley estimator with the original weights, as in (3.1). This is considered the best available estimate since it requires no assumptions for unbiasedness.

The first row contains the five estimates, the second row contains an estimate of bias for each, and the third row shows the square root of the sum of estimated variance and squared bias. The bias displayed in row 2 is the difference between each estimate and the corrected Hartley estimate (*True* column). We acknowledge that the address matching algorithm is undoubtedly not perfect, which means that the "True" estimator may still contain bias in addition to its sampling variability. Still, taking this as our best assessment of bias, we see that after applying the bias correction method, the estimated bias is reduced by using the bias-corrected estimator from $211K$ to $40K - 80K$. The difference between $\hat{Y}_H$ and $\hat{\hat{Y}}_{\mathrm{BC}}$ with 100% subsampling may reflect failure of the required equal mean assumptions. The estimated RMSE is reduced by using a bias-adjusted method instead of the unadjusted Hartley estimator by about $70K$.

Table 6.3
Estimated total fishing trips (Address frame, Wave 6, 2009)

	Unadj.	20% sub. $m_A = 36$	40% sub. $m_A = 71$	100% sub.	True
Estimate	731,430	889,860	863,488	905,947	942,360
Bias	210,930	52,500	78,872	36,413	0
RMSE	244,531	181,809	180,311	176,954	213,966

7 Discussion

An important observation from this application was that respondents may have difficulties providing accurate information about domain membership, even when it is defined by a straightforward concept, like having a fishing license. This poses a particular problem for dual frame estimation. When domain

membership is available from other sources with minimal cost, there is a simple solution. But when the cost is high, it may be beneficial to use a bias-correction method. Another observation that was important to our application was that the only previously available method that we were aware of required an unstated assumption that units are homogeneous within true domains. This would seem to be a natural and benign assumption for many applications, but appeared not to hold for the key item in our questionnaire.

In this study, the reason for the domain misclassification error was that the respondent was unable or unwilling to provide the correct information. In other cases, it could be that the frame itself contains errors. In either case, identifying the mechanism causing the error could help to determine which assumption about means is most plausible. In our application, it seems obvious in retrospect that a person who admits to frequent fishing might be reluctant to admit to a government agency conducting the survey that they are not licensed.

Many applications of dual frame estimation are for the purpose of improving efficiency rather than coverage, as in our application, where the license frame was included to reduce the cost of contacting anglers. In that case, membership in one of the frames is likely to be predictive of key response variables in the survey and so means of the responses for the subgroups of the population may vary widely. However, this is not always the case. If neither frame is directly related to the topic of the survey itself, it may be more likely that the true domain determines the mean response, or even that all four subgroups of the population have the same mean. (In the latter case, neither bias-correction method would be incorrect.) For example, suppose the two frames for this survey were land-line and cell-phone frames, and a respondent sampled from the cell frame, for example, is asked if they have a land-line phone in order to determine if they are in the overlap domain. Responses to this question are likely to have some measurement error. However, it seems unlikely that whether a respondent *says* he or she has a land-line is more predictive of angling avidity than whether he or she actually *has* a land-line. (In fact the latter will probably be related to fishing avidity because both are correlated with age.) Predicting which mean assumption will hold will benefit from the advice of experts on the topic of the survey. However, in the end, examining the data from the survey itself before a decision is made about bias-correction will be necessary.

The cost of bias-correction is increased variance. This penalty is significant, especially if little information is available about the misclassification rates. Therefore, if it can be determined that there are few errors, either because the domains subject to errors are a small fraction of the population or the error rate is very small, then bias-correction may not be worthwhile. Calculating the bias-corrected estimators is straightforward with survey software, once misclassification estimates are available. Their variance estimates, along with the difference in bias-corrected and not bias-corrected estimates themselves can help guide the choice.

Our research was done in the context of Hartley's dual frame estimator since that was the estimator being used in our application. Many different dual-frame estimators are available, and all require knowledge of domain membership. Some, such as the Fuller-Burmeister estimator, could be adjusted using methods similar to those outlined here. Others, such as the pseudo-maximum likelihood estimators, would require a different approach.

References

Andrews, R., Brick, J.M., Mathiowetz, N. and Stokes, L. (2010). Pilot test of a dual frame two-phase mail survey of anglers in North Carolina. Final report for National Oceanic and Atmospheric Administration.

Bankier, M. (1986). Estimators based on several stratied samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Fuller, W., and Burmeister, L. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

Hartley, H. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section,* American Statistical Association, 203-206.

Kalton, G., and Anderson, D. (1986). Pilot test of a dual frame two-phase mail survey of anglers in North Carolina. *Journal of the Royal Statistical Society, Series A*, 149, 65-82.

Lin, D. (2014). Dissertation: *Measurement Error in Dual Frame Estimation*.

Lohr, S. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology*, 37, 2, 197-213. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11608-eng.pdf.

National Research Council (NRC) (2006). Review of recreational sheries survey methods. Technical report.

Skinner, C., and Rao, J. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Survey Methodology, December 2019
Vol. 45, No. 3, pp. 567-576
Statistics Canada, Catalogue No. 12-001-X

On a new estimator for the variance of the ratio estimator with small sample corrections

Paul Knottnerus and Sander Scholtus[1]

Abstract

The widely used formulas for the variance of the ratio estimator may lead to serious underestimates when the sample size is small; see Sukhatme (1954), Koop (1968), Rao (1969), and Cochran (1977, pages 163-164). In order to solve this classical problem, we propose in this paper new estimators for the variance and the mean square error of the ratio estimator that do not suffer from such a large negative bias. Similar estimation formulas can be derived for alternative ratio estimators as discussed in Tin (1965). We compare three mean square error estimators for the ratio estimator in a simulation study.

Key Words: Bias; Product moments; Sample variance; Taylor series expansion.

1 Introduction

Consider a population of N distinct units with values (x_i, y_i) $(i = 1, \ldots, N)$ of the variables x and y $(x_i > 0)$. Denote the corresponding population means by $\bar{X}$ and $\bar{Y}$, that is $\bar{X} = \sum_{i=1}^{N} x_i \big/ N$ and $\bar{Y} = \sum_{i=1}^{N} y_i \big/ N$. Define R by $R = \bar{Y} \big/ \bar{X}$. Suppose that a simple random sample of size n is selected from the population. When $\bar{X}$ is known, $\bar{Y}$ can be estimated by the ratio estimator

$$\hat{\bar{Y}}_R = \hat{R}\bar{X}, \tag{1.1}$$

where $\hat{R} = \bar{y}_s \big/ \bar{x}_s$ with $\bar{y}_s = \sum_{i=1}^{n} y_i \big/ n$ and $\bar{x}_s = \sum_{i=1}^{n} x_i \big/ n$; see Cochran (1977, page 151). For large n, the well-known approximation for the variance of $\hat{\bar{Y}}_R$ is

$$\text{var}\left(\hat{\bar{Y}}_R\right) \approx \frac{1-f}{n} S_e^2, \tag{1.2}$$

where $f = n/N$, $S_e^2 = \sum_{i=1}^{N} e_i^2 \big/ (N-1)$ and $e_i = y_i - Rx_i$ $(i = 1, \ldots, N)$; note that $\bar{E} = \sum_{i=1}^{N} e_i \big/ N = 0$. When n is small, the approximation error of (1.2) can be considerable; see Koop (1968). Moreover, this error may increase when, in practice, S_e^2 in (1.2) is replaced by its standard estimator $s_{\hat{e}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \hat{e}_i^2$ where $\hat{e}_i = y_i - \hat{R}x_i$ $(i = 1, \ldots, n)$; see Cochran (1977, page 163). As stated by Koop (1968), the cause of the discrepancy relative to the true variance lies in neglecting terms in $1/n^2$ and $1/n^3$, and perhaps also those of higher orders.

The three main aims of this paper are: (*i*) to improve approximation (1.2) for small values of n by using a second-order Taylor series expansion of $1/\bar{x}_s$; (*ii*) to derive a new estimator for S_e^2 that is less biased than $s_{\hat{e}}^2$; and (*iii*) to derive a new variance estimator for the ratio estimator. Although a normal distributional approximation might be imprecise at the sample sizes considered in this paper, such a more accurate

1. Paul Knottnerus, Statistics Netherlands, P.O. Box 24500, 2490 HA, The Hague, The Netherlands. E-mail: pkts@cbs.nl; Sander Scholtus, Statistics Netherlands, P.O. Box 24500, 2490 HA, The Hague, The Netherlands. E-mail: sshs@cbs.nl.

variance estimator is useful in order to get some more insight into the precision of the ratio estimator in comparison with that of other estimators. For instance, in case of small samples from small strata the combined ratio estimate for $\bar{Y}$ is to be recommended rather than the separate ratio estimate certainly when the ratios (say, R_h) are constant from stratum to stratum; see Cochran (1977, page 167).

The outline of the paper is as follows. Using some results of Nath (1968), we derive in Section 2 an alternative approximation formula for the variance of $\hat{R}$ with an error of order $1/n^3$. In addition, we derive a new approximation formula for the bias of the residual sampling variance $s_{\hat{e}}^2$ of order $1/n$. Furthermore, we propose two new estimators for the mean square error (MSE) of $\hat{\bar{Y}}_R$. In Section 3 we carry out a simulation study in order to compare the standard variance estimator with the new estimators proposed in Section 2. Section 4 summarizes the main conclusions.

2 A new variance estimator

Noting that $\hat{R} - R = \bar{e}_s / \bar{x}_s$, where $\bar{e}_s$ is the sample mean of e_i, and using the second-order Taylor series expansion

$$\frac{1}{\bar{x}_s} = \frac{1}{\bar{X}} - \frac{1}{\bar{X}^2}(\bar{x}_s - \bar{X}) + \frac{1}{\bar{X}^3}(\bar{x}_s - \bar{X})^2 + O_p\left(\frac{1}{n^{1.5}}\right),$$

it is seen that the third-order Taylor series expansion of $\hat{R} - R$ is

$$\hat{R} - R = \frac{\bar{e}_s}{\bar{X}} - \frac{1}{\bar{X}^2}(\bar{x}_s - \bar{X})\bar{e}_s + \frac{1}{\bar{X}^3}(\bar{x}_s - \bar{X})^2 \bar{e}_s + O_p\left(\frac{1}{n^2}\right). \tag{2.1}$$

Hence, using $\hat{\bar{Y}}_R = \bar{X}\hat{R}$, we obtain

$$\begin{aligned}\operatorname{var}\left(\hat{\bar{Y}}_R\right) &= \operatorname{var}(\bar{e}_s) + \frac{1}{\bar{X}^2}\operatorname{var}\{(\bar{x}_s - \bar{X})\bar{e}_s\} - \frac{2}{\bar{X}}\operatorname{cov}\{\bar{e}_s, (\bar{x}_s - \bar{X})\bar{e}_s\} \\ &\quad + \frac{2}{\bar{X}^2}\operatorname{cov}\{\bar{e}_s, (\bar{x}_s - \bar{X})^2 \bar{e}_s\} + O\left(\frac{1}{n^3}\right).\end{aligned} \tag{2.2}$$

In (2.2) we omitted one variance and one covariance because the underlying fifth and sixth moments are of order $1/n^3$; see David and Sukhatme (1974). All (co)variances in (2.2) can be evaluated by using the following results on product moments of four arbitrary sample means, say $\bar{x}_{sa}$, $\bar{x}_{sb}$, $\bar{x}_{sc}$ and $\bar{x}_{sd}$,

$$E(\bar{x}_{sa}\bar{x}_{sb}\bar{x}_{sc}) = (1 - f)(1 - 2f)S_{abc}/n^2 + O(n^{-3}) \tag{2.3}$$

$$E(\bar{x}_{sa}\bar{x}_{sb}\bar{x}_{sc}\bar{x}_{sd}) = \gamma(S_{ab}S_{cd} + S_{ac}S_{bd} + S_{ad}S_{bc}) + O(n^{-3}) \tag{2.4}$$

$$\operatorname{cov}(\bar{x}_{sa}\bar{x}_{sb}, \bar{x}_{sc}\bar{x}_{sd}) = \gamma(S_{ac}S_{bd} + S_{ad}S_{bc}) + O(n^{-3}) \tag{2.5}$$

$$E(\bar{x}_{sa}^2\bar{x}_{sb}^2) = \gamma(S_{aa}S_{bb} + 2S_{ab}^2) + O(n^{-3}), \tag{2.6}$$

where $\gamma = (1-f)^2/n^2$, $S_{ab} = \sum_{i=1}^{N} x_{ia}x_{ib}/(N-1)$ and $S_{abc} = \sum_{i=1}^{N} x_{ia}x_{ib}x_{ic}/(N-1)$. Without loss of generality, it is assumed for expediency that the population means are zero, that is, $\bar{X}_a = \bar{X}_b = \bar{X}_c = \bar{X}_d = 0$. Formulas (2.3) and (2.4) follow from Theorems 1 and 2 of Nath (1968) while (2.5) and (2.6) follow from (2.4). From (2.2)-(2.6) it follows that

$$\begin{aligned}\operatorname{var}\left(\hat{\bar{Y}}_R\right) &= \frac{1-f}{n}S_e^2 + \left(\frac{1-f}{n\bar{X}}\right)^2 (S_x^2 S_e^2 + S_{xe}^2) - \frac{2}{n^2\bar{X}}(1-f)(1-2f)S_{xee} \\ &\quad + 2\left(\frac{1-f}{n\bar{X}}\right)^2 (S_x^2 S_e^2 + 2S_{xe}^2) + O(n^{-3}) \\ &= \frac{1-f}{n}S_e^2\left\{1 + 3\left(\frac{1-f}{n\bar{X}^2}\right)S_x^2\right\} + 5\left(\frac{1-f}{n\bar{X}}\right)^2 S_{xe}^2 - 2\frac{(1-f)(1-2f)}{n^2\bar{X}}S_{xee} + O(n^{-3}), \quad (2.7)\end{aligned}$$

where

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2, \quad S_{xe} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})e_i$$

$$S_{xee} = S_{ge} = \frac{1}{N-1}\sum_{i=1}^{N} e_i^2(x_i - \bar{X}), \quad g_i = (x_i - \bar{X})e_i.$$

Similar formulas in terms of cumulants are derived by Tin (1965) using some results from Kendall and Stuart (1958). Unfortunately, the numerous cumulants in Tin's formulas give little insight into the structure of $\operatorname{var}(\hat{\bar{Y}}_R)$ and, consequently, small sample corrections for the variance estimator require somewhat tedious calculations. In contrast, from (2.7) it is seen that for sufficiently large n approximation (1.2) leads to an underestimate unless $S_{xee}(= S_{ge})$ is very positive. In addition, Tin also discusses three alternative estimators for a ratio but small sample corrections when estimating the various variances are ignored by him.

It follows from (2.1) and (2.3) that

$$\operatorname{bias}\left(\hat{\bar{Y}}_R\right) = -\frac{1-f}{n\bar{X}}S_{xe} + O\left(\frac{1}{n^2}\right); \quad (2.8)$$

also see Cochran (1977, page 161). Subsequently, using $S_{xee} = S_{ge}$, it follows from (2.7) and (2.8) that the mean square error of $\hat{\bar{Y}}_R$ is

$$\operatorname{MSE}\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n}S_e^2\left\{1 + 3\left(\frac{1-f}{n\bar{X}^2}\right)S_x^2\right\} + 6\left(\frac{1-f}{n\bar{X}}\right)^2 S_{xe}^2 - 2\frac{(1-f)(1-2f)}{n^2\bar{X}}S_{ge} + O(n^{-3}). \quad (2.9)$$

When the variation coefficient $C_x\ (\equiv S_x/\bar{X})$ is known, it is useful to write (2.9) as

$$\operatorname{MSE}\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n}S_e^2\left\{1 + 3\left(\frac{1-f}{n}\right)C_x^2(1+2\rho_{xe}^2) - 2\frac{(1-2f)}{n}C_x\rho_{ge}\,S_g/(S_xS_e)\right\}, \quad (2.10)$$

where $\rho_{xe} = S_{xe}/S_x S_e$, $\rho_{ge} = S_{ge}/S_g S_e$ and $S_g^2 = \frac{1}{N-1}\sum_{i=1}^{N}(g_i - \bar{G})^2$. In practice, $\mathrm{MSE}(\hat{\bar{Y}}_R)$ in (2.10) can be estimated by

$$\widehat{\mathrm{MSE}}_1\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n} s_{\hat{e}}^2 \left\{1 + 3\left(\frac{1-f}{n}\right) C_x^2 (1 + 2\hat{\rho}_{x\hat{e}}^2) - 2\frac{(1-2f)}{n} C_x \hat{\rho}_{\hat{g}\hat{e}}\, s_{\hat{g}}/(s_x s_{\hat{e}})\right\}, \quad (2.11)$$

where $\hat{\rho}_{x\hat{e}} = s_{x\hat{e}}/s_x s_{\hat{e}}$, $\hat{\rho}_{\hat{g}\hat{e}} = s_{\hat{g}\hat{e}}/s_{\hat{g}} s_{\hat{e}}$ and

$$s_{\hat{g}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{g}_i - \bar{\hat{g}}_s\right)^2 \quad [\text{note: } \hat{g}_i = (x_i - \bar{x}_s)\hat{e}_i]$$

$$s_{x\hat{e}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)\hat{e}_i, \quad s_{\hat{g}\hat{e}} = s_{x\hat{e}\hat{e}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)\hat{e}_i^2.$$

However, the estimator in (2.11) does not take into account the bias of $s_{\hat{e}}^2$ defined above.

In order to examine the bias of $s_{\hat{e}}^2 = \sum_{i=1}^{n}\hat{e}_i^2/(n-1)$, we use some additional symbols

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)^2, \quad s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e}_s)^2$$

$$s_{xe} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_s)e_i, \quad q_i = (x_i - \bar{X})^2 \ (i = 1, \ldots, N).$$

Now we can write $E(s_{\hat{e}}^2)$ as

$$\begin{aligned} E(s_{\hat{e}}^2) &= E\frac{1}{n-1}\sum_{i=1}^{n}\left\{y_i - \bar{y}_s - R(x_i - \bar{x}_s) - \left(\hat{R} - R\right)(x_i - \bar{x}_s)\right\}^2 \\ &= E(s_e^2) + E\left\{\left(\hat{R} - R\right)^2 s_x^2\right\} - 2E\left\{\left(\hat{R} - R\right) s_{xe}\right\} \\ &= S_e^2 + E\left\{\left(\hat{R} - R\right)^2 \bar{q}_s\right\} - 2E\left\{\left(\hat{R} - R\right)\bar{g}_s\right\} + O(n^{-2}), \end{aligned} \quad (2.12)$$

where $\bar{q}_s$ and $\bar{g}_s$ are sample means of q_i and g_i, respectively. In (2.12) we used

$$s_x^2 = \left\{\bar{q}_s - (\bar{x}_s - \bar{X})^2\right\}\left(1 + \frac{1}{n-1}\right), \quad s_{xe} = \left\{\bar{g}_s - (\bar{x}_s - \bar{X})\bar{e}_s\right\}\left(1 + \frac{1}{n-1}\right)$$

and hence, using (2.1), (2.3) and (2.4), we get

$$\begin{aligned} E\left\{\left(\hat{R} - R\right)^2(\bar{q}_s - s_x^2)\right\} &= E\left\{\bar{e}_s^2(\bar{x}_s - \bar{X})^2/\bar{X}^2\right\}\{1 + o(1)\} = O(n^{-2}) \\ E\left\{\left(\hat{R} - R\right)(\bar{g}_s - s_{xe})\right\} &= E\left\{\bar{e}_s(\bar{x}_s - \bar{X})\bar{e}_s/\bar{X}\right\}\{1 + o(1)\} = O(n^{-2}). \end{aligned}$$

From (2.1) and (2.12) it is seen that

$$\begin{aligned} \mathrm{bias}(s_{\hat{e}}^2) &= E\left\{\left(\hat{R} - R\right)^2(\bar{q}_s - \bar{Q} + \bar{Q})\right\} - 2E\left\{\left(\hat{R} - R\right)(\bar{g}_s - \bar{G} + \bar{G})\right\} + O(n^{-2}) \\ &= \frac{1-f}{n\bar{X}^2} S_e^2 \bar{Q} - 2E\left[\left\{\frac{\bar{e}_s}{\bar{X}} - \frac{1}{\bar{X}^2}(\bar{x}_s - \bar{X})\bar{e}_s\right\}(\bar{g}_s - \bar{G} + \bar{G})\right] + O(n^{-2}) \\ &= \frac{1-f}{n\bar{X}^2} S_e^2 S_x^2 - 2\frac{1-f}{n}\left(\frac{S_{ge}}{\bar{X}} - \frac{S_{xe}^2}{\bar{X}^2}\right) + O(n^{-2}), \end{aligned} \quad (2.13)$$

where we used $\bar{Q} = S_x^2(1 - N^{-1})$ and $\bar{G} = S_{xe}(1 - N^{-1})$. Note that it follows from (2.13) that for sufficiently large n, the quantity $s_{\hat{e}}^2$ leads to an overestimate of S_e^2 unless S_{ge} is very positive. To our best knowledge, formula (2.13) is not mentioned elsewhere in the literature.

Based on (2.13), an alternative estimator of $\text{MSE}(\hat{\bar{Y}}_R)$ that takes the bias of $s_{\hat{e}}^2$ into account is

$$\widehat{\text{MSE}}_2\left(\hat{\bar{Y}}_R\right) = \frac{1-f}{n}\hat{S}_e^2\left\{1 + 3\left(\frac{1-f}{n}\right)C_x^2(1 + 2\hat{\rho}_{x\hat{e}}^2) - 2\frac{(1-2f)}{n}C_x\hat{\rho}_{\hat{g}\hat{e}}s_{\hat{g}}/(s_x s_{\hat{e}})\right\}, \quad (2.14)$$

where $\hat{S}_e^2$ is adjusted for the relative bias of $s_{\hat{e}}^2$ that follows from (2.13). That is,

$$\hat{S}_e^2 = s_{\hat{e}}^2\left[1 - \frac{1-f}{n}C_x^2(1 + 2\hat{\rho}_{x\hat{e}}^2) + 2\frac{1-f}{n}C_x\frac{\hat{\rho}_{\hat{g}\hat{e}}s_{\hat{g}}}{s_x s_{\hat{e}}}\right].$$

Note that we used here $S_{xe}^2/\bar{X}^2S_e^2 = \rho_{xe}^2C_x^2$ and $S_{ge}/\bar{X}S_e^2 = \rho_{ge}S_gC_x/S_eS_x$. Finally, it should be noted that the other estimators $\hat{\rho}_{x\hat{e}}^2$ and $\hat{\rho}_{\hat{g}\hat{e}}s_{\hat{g}}/(s_x s_{\hat{e}})$ in (2.11) are also biased. However, it is less straightforward to derive that kind of bias. It is hoped that by taking all (co)variances from the sample, including s_x^2, their bias is modest. In addition, in the simulations of Section 3, we found that replacing C_x by $s_x/\bar{x}_s$ did not improve the results.

3 A simulation study

3.1 Set-up and main results

In this section we apply the above results to eleven populations. Populations 1-5 are taken from Cochran (1977, pages 152, 182, 203, 325), populations 6 and 7 from Sukhatme (1954, pages 183-184), population 8 from Kish (1995, page 42) and populations 9-11 are taken from Koop (1968). The population sizes vary between 10 and 49. The correlation coefficients between y and x vary between 0.32 and 0.98, while the coefficients of variation of x vary between 0.14 and 1.19. For further details, see Table 3.1.

We considered simple random samples without replacement of sizes $n = 4, 6, \ldots, 14$ from these populations (excluding cases where $n \geq N$). For each population, we simulated all $\binom{N}{n}$ possible samples of size n provided that this number is not larger than one million. When $\binom{N}{n} > 10^6$, we restricted ourselves to drawing one million random samples of size n from the population. From these simulated samples, we computed (an accurate estimate of) the true mean square error of $\hat{\bar{Y}}_R$ for a given population and a given sample size, to be used as a benchmark.

For each sample, we calculated the standard variance estimator for $\hat{\bar{Y}}_R$, say $\widehat{\text{var}}(\hat{\bar{Y}}_R)$, based on (1.2) with S_e^2 replaced by $s_{\hat{e}}^2$. This estimator is also the standard estimator of the mean square error of $\hat{\bar{Y}}_R$, say $\widehat{\text{MSE}}_0(\hat{\bar{Y}}_R)$, with an error of order $1/n^2$. Furthermore, we calculated the new estimators $\widehat{\text{MSE}}_1(\hat{\bar{Y}}_R)$ and $\widehat{\text{MSE}}_2(\hat{\bar{Y}}_R)$ for the mean square error of $\hat{\bar{Y}}_R$ from (2.11) and (2.14). It is expected that these estimators are more accurate than the standard estimator, as they have an error of order $1/n^3$.

Table 3.1
Key features of the eleven populations used in the simulation study

	Source	N	$\bar{Y}$	$\bar{X}$	R	S_e^2	C_x	ρ_{xy}	ρ_{xe}	ρ_{ge}
1	Cochran, page 152	49	128	103	1.24	621	1.01	0.98	-0.34	0.02
2	Cochran, page 182	34	2.91	8.37	0.35	5.72	1.03	0.72	-0.24	0.56
3	Cochran, page 182	34	2.59	4.92	0.53	4.81	1.02	0.73	-0.14	0.38
4	Cochran, page 203	10	54.3	56.9	0.95	6.71	0.17	0.97	0.38	-0.01
5	Cochran, page 325	10	101	58.8	1.72	150	0.14	0.65	-0.29	-0.29
6	Sukhatme, pages 183-184	34	201	218	0.92	3,304	0.77	0.93	-0.23	0.93
7	Sukhatme, pages 183-184	34	218	765	0.29	8,735	0.62	0.83	0.05	0.44
8	Kish, page 42	20	12.8	21.8	0.59	17.8	1.19	0.97	0.23	0.75
9	Koop, population 1	20	4.40	6.30	0.70	0.41	0.67	0.98	-0.06	0.50
10	Koop, population 2	20	4.50	51.2	0.09	4.87	0.44	0.42	-0.50	-0.85
11	Koop, population 3	20	15.6	30.0	0.52	36.3	0.40	0.32	-0.88	0.11

To compare the accuracy of these three estimators, we evaluated their relative bias with respect to the benchmark value for the true mean square error of $\hat{\bar{Y}}_R$:

$$\mathrm{RB}_k = \frac{E\left\{\widehat{\mathrm{MSE}}_k\left(\hat{\bar{Y}}_R\right)\right\} - \mathrm{MSE}\left(\hat{\bar{Y}}_R\right)}{\mathrm{MSE}\left(\hat{\bar{Y}}_R\right)} \times 100\%, \quad k \in \{0, 1, 2\}.$$

The mean square error $\mathrm{MSE}(\hat{\bar{Y}}_R)$ consists of $\mathrm{bias}^2(\hat{\bar{Y}}_R)$ and $\mathrm{var}(\hat{\bar{Y}}_R)$. For all populations in this study we found that, in spite of the small sample sizes, the bias of $\hat{\bar{Y}}_R$ as an estimator for $\bar{Y}$ was more or less negligible. In fact, the largest relative bias of $\hat{\bar{Y}}_R$ always occurred for $n = 4$ and varied between -4% and +4%. In other words, in this study the true and estimated mean square errors were dominated by their variance components.

Table 3.2 gives the results. Firstly, it is seen that the standard estimator $\widehat{\mathrm{MSE}}_0(\hat{\bar{Y}}_R)$ usually underestimates the true mean square error. The negative bias of this estimator can be very large (up to more than -60% for population 8). Secondly, it is striking that for the three populations in Koop's paper (populations 9-11), $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$ always estimates the true MSE of $\hat{\bar{Y}}_R$ with a relative bias of less than 5%. For the other populations, the relative bias is always less than 7% except for populations 1, 6 and 8 with $n = 4$ and $n = 6$. For $n \geq 10$, $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$ is always more accurate than $\widehat{\mathrm{MSE}}_0(\hat{\bar{Y}}_R)$, and in fact this is also true for most cases with $n < 10$. For $n \geq 8$, $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$ nearly always performs better than $\widehat{\mathrm{MSE}}_1(\hat{\bar{Y}}_R)$, which shows that correcting for the bias in $s_{\hat{e}}^2$ is useful. Furthermore, it can be seen from Table 3.2 that, in general, $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$ suffers much less from a negative bias than $\widehat{\mathrm{MSE}}_0(\hat{\bar{Y}}_R)$ while $\widehat{\mathrm{MSE}}_1(\hat{\bar{Y}}_R)$ suffers from a positive bias.

Table 3.2
Relative bias RB_k for the three estimators of $MSE(\hat{\bar{Y}}_R)$

population	estimator	$n = 4$	$n = 6$	$n = 8$	$n = 10$	$n = 12$	$n = 14$
1	$\widehat{MSE}_0$	-48.2%	-35.6%	-27.1%	-21.6%	-17.2%	-14.2%
	$\widehat{MSE}_1$	27.4%	15.8%	10.9%	7.7%	6.3%	5.1%
	$\widehat{MSE}_2$	-30.9%	-11.7%	-5.6%	-3.5%	-2.1%	-1.4%
2	$\widehat{MSE}_0$	-34.9%	-27.7%	-22.3%	-18.7%	-16.1%	-13.6%
	$\widehat{MSE}_1$	32.6%	10.1%	3.3%	0.5%	-0.9%	-0.9%
	$\widehat{MSE}_2$	2.8%	3.4%	1.7%	0.4%	-0.5%	-0.5%
3	$\widehat{MSE}_0$	-37.2%	-28.4%	-22.4%	-17.9%	-14.4%	-11.6%
	$\widehat{MSE}_1$	26.1%	7.7%	2.6%	1.0%	0.6%	0.7%
	$\widehat{MSE}_2$	-2.8%	-0.6%	-1.3%	-1.3%	-1.1%	-0.6%
4	$\widehat{MSE}_0$	-1.0%	-0.4%	-0.1%			
	$\widehat{MSE}_1$	1.4%	0.5%	0.2%			
	$\widehat{MSE}_2$	0.7%	0.3%	0.1%			
5	$\widehat{MSE}_0$	0.4%	0.7%	0.8%			
	$\widehat{MSE}_1$	2.0%	1.0%	0.5%			
	$\widehat{MSE}_2$	0.8%	0.4%	0.2%			
6	$\widehat{MSE}_0$	-19.2%	-17.3%	-15.8%	-14.7%	-14.1%	-13.5%
	$\widehat{MSE}_1$	21.1%	0.8%	-5.4%	-7.4%	-7.9%	-7.8%
	$\widehat{MSE}_2$	20.6%	10.2%	4.9%	2.3%	0.7%	-0.3%
7	$\widehat{MSE}_0$	-17.8%	-12.0%	-8.7%	-6.7%	-5.3%	-4.3%
	$\widehat{MSE}_1$	4.9%	0.3%	-0.1%	0.0%	0.0%	0.0%
	$\widehat{MSE}_2$	0.0%	-0.6%	-0.5%	-0.3%	-0.3%	-0.2%
8	$\widehat{MSE}_0$	-62.3%	-45.8%	-34.9%	-28.0%	-23.4%	-20.3%
	$\widehat{MSE}_1$	-11.1%	-8.2%	-6.5%	-5.7%	-5.3%	-4.8%
	$\widehat{MSE}_2$	-34.4%	-13.3%	-6.4%	-4.0%	-3.3%	-3.2%
9	$\widehat{MSE}_0$	-20.1%	-13.2%	-9.7%	-7.6%	-6.2%	-5.2%
	$\widehat{MSE}_1$	7.4%	1.0%	-0.5%	-0.8%	-0.8%	-0.7%
	$\widehat{MSE}_2$	0.4%	0.1%	-0.2%	-0.3%	-0.4%	-0.4%
10	$\widehat{MSE}_0$	-8.9%	-2.0%	0.9%	2.5%	3.5%	4.2%
	$\widehat{MSE}_1$	21.1%	15.4%	10.9%	7.7%	5.4%	3.7%
	$\widehat{MSE}_2$	0.9%	2.1%	2.0%	1.7%	1.4%	1.1%
11	$\widehat{MSE}_0$	-17.5%	-10.1%	-6.5%	-4.4%	-3.0%	-2.1%
	$\widehat{MSE}_1$	3.4%	3.0%	2.3%	1.7%	1.2%	0.8%
	$\widehat{MSE}_2$	-4.3%	-1.2%	-0.3%	0.0%	0.0%	0.1%
mean	$\widehat{MSE}_0$	-24.2%	-17.4%	-13.3%	-13.0%	-10.7%	-8.9%
	$\widehat{MSE}_1$	12.4%	4.3%	1.7%	0.5%	-0.1%	-0.4%
	$\widehat{MSE}_2$	-4.2%	-1.0%	-0.5%	-0.6%	-0.6%	-0.6%

3.2 Discussion of two specific results

Referring back to Table 3.1, it may be noted that both populations 1 and 8, where the largest relative negative errors occur for $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$, involve a strong correlation ρ_{xy} in combination with a relatively large value of C_x in comparison to the other populations in our study ($\rho_{xy} \geq 0.97$ and $C_x \geq 1.01$). It is therefore interesting to examine the effect of these quantities on the accuracy of the estimated mean square error more closely.

Firstly, suppose that the following transformation is applied to the values of x, e and y in a given population:

$$x' := x, \quad e' := ae, \quad y' := Rx + e',$$

with $a \neq 0$. Under this transformation, the ratio of the two variables does not change ($R' = \bar{Y}' / \bar{X}' = R$) but their correlation coefficient does ($\rho_{x'y'} \neq \rho_{xy}$ unless $a = 1$). It is obvious that $C_{x'} = C_x$ and $S_{e'}^2 = a^2 S_e^2$. Now using expressions (1.2), (2.8), (2.11) and (2.14), it is not difficult to see that $E\{\widehat{\mathrm{MSE}}_k(\hat{\bar{Y}}_R')\} = a^2 E\{\widehat{\mathrm{MSE}}_k(\hat{\bar{Y}}_R)\}$ for all $k \in \{0, 1, 2\}$. Moreover, it can be seen from (2.1) that the error in $\hat{R}$ is linear in $\bar{e}_s$ and hence it follows that the identity $\mathrm{MSE}(\hat{\bar{Y}}_R') = a^2 \mathrm{MSE}(\hat{\bar{Y}}_R)$ holds exactly. Thus, it is seen that this transformation has no effect on the relative bias RB_k of any of the mean square error estimators in this study. This suggests that this bias is not affected by a change in the correlation ρ_{xy} when other features of the population remain constant. In particular, this suggests that the large values of ρ_{xy} in populations 1 and 8 alone do not explain the lack of accuracy of $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$ in these populations.

Secondly, consider the following alternative transformation:

$$x'' := \bar{X} + b(x - \bar{X}), \quad e'' := be, \quad y'' := \bar{Y} + b(y - \bar{Y}),$$

with $0 < b \leq 1$. In this case, it can be shown that $R'' = R$, $\rho_{x''y''} = \rho_{xy}$ and $C_{x''} = bC_x \leq C_x$. Thus, this transformation can be used to reduce the coefficient of variation of x in a given population, while holding the ratio and correlation of y and x fixed.

We have applied this transformation to populations 1 and 8 for $n = 4$, with $b = 1.0, 0.9, \ldots, 0.2$. Table 3.3 shows the resulting relative bias of $\widehat{\mathrm{MSE}}_k(\hat{\bar{Y}}_R'')$ for the transformed populations, obtained by simulating all $\binom{49}{4} = 211{,}876$ and $\binom{20}{4} = 4{,}845$ possible samples, respectively. It is seen that all three estimators for the mean square error tend to become less biased as the coefficient of variation of x is reduced. In particular, $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R'')$ becomes reasonably accurate (considering that $n = 4$) once the coefficient of variation of x drops below 0.8 for population 1 and below 1 for population 8.

This suggests that the value of C_x – which is known in practice – is an important factor for the (negative) bias of our proposed estimator $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$. Assuming that the set of natural populations in this simulation study contains sufficient variation to represent most populations that will be encountered in practice, we may tentatively conclude that even for $n = 4$, $\widehat{\mathrm{MSE}}_2(\hat{\bar{Y}}_R)$ is an accurate estimator of the mean square error of the ratio estimator without a large negative bias when $C_x < 0.8$. For $C_x \geq 0.8$, this need not be the case.

Table 3.3
Relative bias RB_k for transformed versions of populations 1 and 8, with $n = 4$

	Population 1				Population 8			
		relative bias				relative bias		
b	C_{x^b}	$\widehat{\text{MSE}}_0$	$\widehat{\text{MSE}}_1$	$\widehat{\text{MSE}}_2$	C_{x^b}	$\widehat{\text{MSE}}_0$	$\widehat{\text{MSE}}_1$	$\widehat{\text{MSE}}_2$
1.0	1.01	-48.2%	27.4%	-30.9%	1.19	-62.3%	-11.1%	-34.4%
0.9	0.91	-39.1%	32.0%	-16.5%	1.07	-48.4%	7.6%	-12.9%
0.8	0.81	-31.0%	31.8%	-6.2%	0.95	-38.3%	14.2%	-0.7%
0.7	0.71	-24.0%	28.5%	0.3%	0.83	-30.0%	15.0%	5.9%
0.6	0.61	-17.8%	23.4%	3.6%	0.72	-23.1%	12.5%	8.4%
0.5	0.51	-12.5%	17.6%	4.6%	0.60	-17.2%	8.6%	8.0%
0.4	0.40	-8.2%	11.9%	4.1%	0.48	-12.3%	4.2%	6.0%
0.3	0.30	-4.7%	6.8%	2.8%	0.36	-8.1%	0.4%	3.5%
0.2	0.20	-2.1%	3.0%	1.4%	0.24	-4.7%	-1.9%	1.2%

4 Conclusions

In this paper we have derived a new approximation formula for $\text{MSE}(\hat{\bar{Y}}_R)$ of order $1/n^2$ and a new formula for the bias of $s_{\hat{e}}^2$ of order $1/n$. The new estimator $\widehat{\text{MSE}}_2(\hat{\bar{Y}}_R)$ which takes into account the bias of $s_{\hat{e}}^2$ appears to be less biased than $\widehat{\text{MSE}}_0(\hat{\bar{Y}}_R) = \widehat{\text{Var}}(\hat{\bar{Y}}_R)$ and $\widehat{\text{MSE}}_1(\hat{\bar{Y}}_R)$. For $n \geq 8$, the bias of $\widehat{\text{MSE}}_2(\hat{\bar{Y}}_R)$ was in all cases of the simulation study less than 7% which is much better than the standard variance estimator; in most cases, this result even holds for $n \geq 4$. For very small n, $\widehat{\text{MSE}}_2(\hat{\bar{Y}}_R)$ may have a large negative bias if the population has a large coefficient of variation C_x. From our simulation study this issue appears to be unlikely to occur as long as $C_x < 0.8$.

Finally, recall that for the populations in this simulation study, the bias of the ratio estimator itself was consistently small, even for $n = 4$. In general, for other populations this bias may not be negligible. Cochran (1977, pages 174-175) discusses several alternative ratio estimators that are unbiased.

References

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

David, I.P., and Sukhatme, B.V. (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association*, 69, 464-466.

Kendall, M.G., and Stuart, A. (1958). *The Advanced Theory of Statistics, Volume I*. London: Charles Griffin and Company.

Kish, L. (1995). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Koop, J.C. (1968). An exercise in ratio estimation. *The American Statistician*, 22, 29-30.

Nath, S.N. (1968). On product moments from a finite universe. *Journal of the American Statistical Association*, 63, 535-541.

Rao, J.N.K. (1969). Ratio and regression estimators. In *New Developments in Survey Sampling*, (Eds., N.L. Johnson and H. Smith), New York: John Wiley & Sons, Inc., 213-234.

Sukhatme, P.V. (1954). *Sampling Theory of Surveys with Applications*, Iowa State College Press, Ames, IA.

Tin, M. (1965). Comparison of some ratio estimators. *Journal of the American Statistical Association*, 60, 294-307.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2019.

- M. Ballin, *ISTAT*
- J.-F. Beaumont, *Statistics Canada*
- R. Benedetti, *Università degli Studi*
- M. Brick, *Westat Inc.*
- P. Brodie, *Office of National Statistics*
- S. Cai, *Carleton University*
- P.J. Cantwell, *U.S. Census Bureau*
- A.-S. Charest, *Université Laval*
- G. Chauvet, *ENSAI*
- J. Chipperfield, *Australian Bureau of Statistics*
- R. Clark, *Australian National University*
- H.F.C. Coelho, *Federal University of Paraiba*
- W. Davis, *University of Wollongong*
- J. Dever, *RTI International*
- D. Di Cecco, *ISTAT*
- D. Draper, *University of California*
- J.L. Eltinge, *U.S. Bureau of Labor Statistics*
- W.A. Fuller, *Iowa State University*
- S. Gabler, *GESIS – Leibniz Institute for the Social Sciences*
- J. Gambino, *Statistics Canada*
- M. Ganninger, *F. Hoffmann - La Roche, Diagnostics Information Solutions*
- W.F. Gross, *University of Wollongong*
- D. Haziza, *Université de Montréal*
- M.A. Hidiroglou, *Statistics Canada*
- J. Hong, *U.S. Census Bureau*
- B. Hulliger, *University of Applied and Arts Sciences Northwestern Switzerland*
- D. Judkins, *ABT Associates Inc Bethesda*
- P. Kelly, *Statistics Canada*
- F. Keusch, *University of Mannheim*
- J. Kim, *Iowa State University*
- P. Kott, *RTI International*
- P. Lahiri, *JPSM, University of Maryland*
- E. Lesage, *INSEE*
- B. Levine, *RTI international*
- P. Lynn, *University of Essex*
- D. Malec, *National Center for Health Statistics*
- K. McConville, *Reed College*
- K. McGeenney, *Penn Schoen Berland*
- S. Merad, *Office of National Statistics*
- Molina, *Universidad Carlos III de Madrid*
- J.F. Muñoz Rosas, *University of Granada*
- A. Norberg, *Statistics Sweden*
- M. Oguz-Alper, *Statistics Norway*
- K. Olson, *University of Nebraska-Lincoln*
- J. Opsomer, *Westat Inc*
- D. Pfeffermann, *University of Southampton*
- D. Piccone, *U.S. Bureau of Labor Statistics*
- J.N.K. Rao, *Carleton University*
- L.-P. Rivest, *Université Laval*
- F. Scheuren, *National Opinion Research Center*
- S. Scholtus, *Statistics Netherlands*
- P.L.N.D. Silva, *Escola Nacional de Ciências Estatísticas*
- C. Skinner, *The London School of Economics and Political Science*
- P. Smith, *University of Southampton*
- D. Steel, *University of Wollongong*
- Sukasih, *RTI*
- M. Sverchkov, *BLS*
- S.M. Tam, *Australian Bureau of Statistics*
- J.-L. Tambay, *Statistics Canada*
- Y. Tillé, *Université de Neuchâtel*
- M. Torabi, *University of Manitoba*
- R. Tourangeau, *Westat Inc*
- D. Toth, *U.S. Bureau of Labor Statistics*
- A.-A. Vallée, *Université Laval*
- J. van den Brakel, *Statistics Netherlands*
- M. van der Loo, *Statistics Netherlands*
- Z. Wang, *Xieman University*
- West, *University of Michigan*
- Wu, *University of Waterloo*
- Zaslavsky, *Harvard University*
- P. Zhao, *Yunnan University*
- L.-C. Zhang, *University of Southampton*

Acknowledgements are also due to those who assisted during the production of the 2019 issues: Céline Ethier, Cynthia Bocci and Jean-Sébastien Provençal of International Cooperation and Methodology Innovation Centre; Joana Bérubé of Economic Statistics Methods Division; Frédéric Picard of Statistical Integration Methods Division; Gilbert Côté of Official Release, Language Services and Respondent Communications Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Kathy Charbonneau, Christina Jaworski, Joseph Prince, Travis Robinson and Jenna Waite as well as our partners in the Communications Division.

ANNOUNCEMENTS

Nominations Sought for the 2021 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium and give the 2021 Waksberg Invited Address at the Statistics Canada Symposium. The paper will be published in an upcoming issue of *Survey Methodology*.

The author of the 2021 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2020 to the chair of the committee, Bob Fay (bobfay@westat.com).

Previous Waksberg Award honorees and their invited papers are:

2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.

2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.

2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.

2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.

2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.

2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.

2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.

2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.

2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.

2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.

2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.

2012 Lars **Lyberg**, "Survey Quality". *Survey Methodology*, vol. 38, 2, 107-130.

2013 Ken **Brewer**, "Three controversies in the history of survey sampling". *Survey Methodology*, vol. 39, 2, 249-262.

2014 Constance F. **Citro**, "From Multiple Modes for Surveys to Multiple Data Sources for Estimates". *Survey Methodology*, vol. 40, 2, 137-161.

2015 Robert M. **Groves**, "Towards a Quality Framework for Blends of Designed and Organic Data". Proceedings: *Symposium 2016, Growth in Statistical Information: Challenges and Benefits*.

2016 Don **Dillman**, "The promise and challenge of pushing respondents to the Web in mixed-mode surveys". *Survey Methodology*, vol. 43, 1, 3-30.
2017 Donald B. **Rubin**, "Conditional calibration and the sage statistician". *Survey Methodology*, vol. 45, 2, 187-198.
2018 Jean-Claude **Deville**, "De la pratique à la théorie : l'exemple du calage à poids bornés". 10ème Colloque Francophone sur les sondages, Université Lumière Lyon 2.
2019 Chris **Skinner**, Manuscript topic under consideration.
2020 Roger **Tourangeau**, Manuscript topic under consideration.

Members of the Waksberg Paper Selection Committee (2019-2020)

Bob Fay, *Westat* (Chair)
Jean Opsomer, *Westat*
Jack Gambino, *Statistics Canada Alumni*
Elizabeth Stuart, *John Hopkins Bloomberg School of Public Health*

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)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)
Cynthia Clark (2013 - 2014)
Louis-Paul Rivest (2014 - 2015)
Tommy Wright (2015 - 2016)
Kirk Wolter (2016 - 2017)
Danny Pfeffermann (2017 - 2018)
Mike Hidiroglou (2018-2019)

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 35, No. 2, June 2019

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 35, No. 3, September 2019

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 47, No. 1, March/mars 2019

CONTENTS TABLE DES MATIÈRES

Volume 47, No. 2, June/juin 2019

Original Articles

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The documents should be divided into numbered sections with suitable verbal titles.

1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.

4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.