Catalogue no. 12-001-X

Survey Methodology

December 2009

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 at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

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

To access and order this product

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

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

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

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

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

Standards of service to the public

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

Statistics Canada

Business Survey Methods Division

Survey Methodology

December 2009

Published by authority of the Minister responsible for Statistics Canada

December 2009

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

Catalogue no. 12-001-XPB
ISSN 0714-0045

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-X au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

SURVEY METHODOLOGY

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.

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. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $20 ($10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology

A Journal Published by Statistics Canada

Volume 35, Number 2, December 2009

Contents

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

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

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 121-122
Statistics Canada, Catalogue No. 12-001-X

In this issue

This issue of *Survey Methodology* opens with the ninth paper in the annual Waksberg Award invited paper series in honour of Joseph Waksberg's contributions to the theory and practice of survey methodology. The editorial board would like to thank the members of the selection committee – Bob Groves, chair, Leyla Mohadjer, Daniel Kasprzyk and Wayne Fuller – for having selected Graham Kalton as the author of this year's Waksberg Award paper.

In his paper entitled "Methods for oversampling rare subpopulations in social surveys" Kalton gives an overview of methods for sampling rare populations, what Kish called minor domains. After discussing general issues he describes several different methods including screening, stratification, two-phase sampling, multiple frames, multiplicity sampling, location sampling, and accumulating samples over time. He discusses the advantages and disadvantages of each method, and gives many examples of their use in surveys. In practice a combination of approaches is often used.

Randomized response strategies are often used in order to reduce nonsampling errors such as nonresponse and measurement errors. They can also be used in the context of statistical disclosure control for public use microdata files. In his paper, Quatember proposes a standardization of randomized response techniques. The statistical properties of the standardized estimator are derived. He applies the proposed method to a survey on academic cheating behaviour.

Xu and Lavallée consider the problem caused by link nonresponse when using the generalized weight share method in indirect sampling. Indirect sampling is used when selecting samples from a population that is not the target population of interest but is related to it. Biased estimates may occur when it is not known that a unit in the sampling population is related to a unit in the target population. The authors propose several weight adjustments to overcome the issue of link nonresponse.

In the context of unit nonresponse, the weights of the respondents are often adjusted by the inverse of the estimated response probability. Da Silva and Opsomer propose to estimate the response probabilities using local polynomial regression. Results of a simulation study are presented confirming the good performance of the proposed method.

In their paper, Van den Brakel and Krieg consider a multivariate structural time series model that accounts for the design of the Dutch Labour Force Survey. The model is used to estimate the unemployment rates. An empirical investigation demonstrates that the proposed model results in a significant increase in accuracy.

Zhang considers estimation of cross-classifications where one margin of the cross-classification corresponds to small areas and where non-response varies from area to area. He develops a double mixed model approach that combines the fixed effects and random area effects of the small area model with the random effects from the missing data mechanism. The associated conditional mean squared error of prediction is approximated in terms of a three-part decomposition, corresponding to a naive prediction variance, a positive correction that accounts for the hypothetical parameter estimation uncertainty based on the latent complete data, and another positive correction for the extra variation due to the missing data.

Souza, Moura and Migon propose a Bayesian small area estimation application using growth models that account for hierarchical and spatial relationships. They use this approach to obtain population predictions for the municipalities not sampled in the Brazilian Annual Household Survey and to increase the precision of the design-based estimates obtained for the sampled municipalities.

Shao and Thompson investigate the problem of variance estimation when a weight adjustment is applied to deal with nonresponse in stratified business surveys. They derive two consistent linearization variance estimators under weak assumptions. Naive jackknife variance estimators do not work well unless the sampling fraction is negligible, which is not the case when there are certainty strata. They propose a modified jackknife variance estimator that is consistent even when there are certainty strata but the non-certainty strata must not have a large sampling fraction. They evaluate their variance estimators empirically using real data and a simulation study.

In his paper, Preston investigates the bootstrap variance estimation for multistage designs when units are selected using simple random sampling without replacement at each stage. He proposes an extension to the commonly used rescaled bootstrap estimator that assumes with replacement sampling or negligible sampling fractions at the first stage. The proposed estimator is compared with the rescaled and Bernoulli bootstrap estimators.

Jang and Eltinge address the problem of estimating degrees of freedom values from stratified multistage designs when a small number of primary sampling units (PSUs) are selected per stratum. Due to the small number of PSUs selected, the traditional Satterthwaite-based degrees of freedom can be a severe underestimate. In their paper, they propose an alternative estimator of the degrees of freedom that uses the within PSU variances to provide auxiliary information on the relative magnitudes of the overall stratum-level variances. The proposed method is illustrated using data from the National Health and Nutrition Examination Survey (NHANES).

The article by Wang and Bellhouse explores an application of nonparametric regression techniques to study the relationship between the response variable and covariates, as well as prediction using auxiliary information in the context of complex surveys. The work is an extension of Bellhouse and Stafford (2001) that used a simple nonparametric regression function to the case of several independent variables, including indicator variables that often appear in regression analysis using survey data.

And finally, we are pleased to inform readers and authors that *Survey Methodology* will shortly be covered by SCOPUS in the Elsevier Bibliographic Databases starting with the June 2008 issue.

Harold Mantel, Deputy Editor

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 123-124
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

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

Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)
Alastair Scott (2006)
Carl-Erik Särndal (2007)
Mary Thompson (2008)
Graham Kalton (2009)
Ivan Fellegi (2010)

Nominations:

The author of the 2011 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, Daniel Kasprzyk, by email to DKasprzyk@Mathematica-Mpr.com. Nominations and suggestions for topics must be received by February 28, 2010.

2009 Waksberg Invited Paper

Author: Graham Kalton

Graham Kalton is Chairman of the Board of Directors and a Senior Vice President at Westat. He has a title of Research Professor in the Joint Program in Survey Methodology at the University of Maryland. Dr. Kalton has wide-ranging interests in survey methodology, and has published on several aspects of the subject, including sample design, nonresponse and imputation, panel surveys, question wording, and coding.He is a Fellow of the American Association for the Advancement of Science, a Fellow of the American Statistical Association, a National Associate of the National Academies, and an elected member of the International Statistical Institute. He delivered the annual Morris Hansen lecture in 2000.

Members of the Waskberg Paper Selection Committee (2009-2010)

Daniel Kasprzyk (Chair), *Mathematica Policy Research*
Wayne A. Fuller, *Iowa State University*
Elizabeth A. Martin
Mary Thompson, *University of Waterloo*

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)

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 125-141
Statistics Canada, Catalogue No. 12-001-X

Methods for oversampling rare subpopulations in social surveys

Graham Kalton [1]

Abstract

Surveys are frequently required to produce estimates for subpopulations, sometimes for a single subpopulation and sometimes for several subpopulations in addition to the total population. When membership of a rare subpopulation (or domain) can be determined from the sampling frame, selecting the required domain sample size is relatively straightforward. In this case the main issue is the extent of oversampling to employ when survey estimates are required for several domains and for the total population. Sampling and oversampling rare domains whose members cannot be identified in advance present a major challenge. A variety of methods has been used in this situation. In addition to large-scale screening, these methods include disproportionate stratified sampling, two-phase sampling, the use of multiple frames, multiplicity sampling, location sampling, panel surveys, and the use of multi-purpose surveys. This paper illustrates the application of these methods in a range of social surveys.

Key Words: Sample allocation; Screening; Disproportionate stratified sampling; Two-phase sampling; Multiple frames; Location sampling; Panel surveys; Multi-purpose surveys.

1. Introduction

I feel very privileged to have been invited to present this year's paper in the Waksberg Invited Paper Series, a series that honors Joe Waksberg for his numerous contributions to survey methodology. I was extremely fortunate to have had the opportunity to work with Joe at Westat for many years and, as did many others, I benefited greatly from that experience. When faced with an intractable sampling problem, Joe had a flair for turning the problem on its end and producing a workable solution. Since the problem often concerned the sampling of rare populations, I have chosen to review methods for sampling rare populations for this paper.

One of the major developments in survey research over the past several decades has been the continuously escalating demand for estimates for smaller and smaller subclasses (subpopulations) of the general population. This paper focuses on those subclasses – termed *domains* – that are planned for separate analysis at the sample design stage. Some examples of domains that have been taken into account in the sample designs of various surveys include a country's states or provinces, counties or districts; racial/ethnic minorities; households living in poverty; recent births; persons over 80 years of age; recent immigrants; gay men; drug users; and disabled persons. When the domains are small (also known as *rare populations*), the need to provide adequate sample sizes for domain analysis can create major challenges in sample design. This paper reviews the different probability sampling methods that are used to generate samples for estimating the characteristics of rare populations with required levels of precision. Sampling methods for estimating the size of a rare population are not explicitly addressed, although similar methods are often applicable. However, capture-recapture and related methods are not addressed in this paper.

An important issue for sample design is whether the aim of a survey is to produce estimates for a single domain or many domains. Although much of the literature on the sampling of rare populations discusses sample designs for a single rare domain (*e.g.*, drug users), in practice surveys are often designed to produce estimates for many domains (*e.g.*, each of the provinces in a country or several racial/ethnic groups). The U.S. National Health and Nutrition Examination Survey (NHANES) is an example of a survey designed to produce estimates for many domains, in this case defined by age, sex, race/ethnicity and low-income status (Mohadjer and Curtin 2008). In sample designs that include many domains, the domains may be mutually exclusive (*e.g.*, provinces or the cells of the cross-classification of age group and race/ethnicity) or they may be intersecting (*e.g.*, domains defined separately by age group and by race/ethnicity).

The size of a domain is a key consideration. Kish (1987) proposed a classification of *major domains* of perhaps 10 percent or more of the total population, for which a general sample will usually produce reliable estimates; *minor domains* of 1 to 10 percent, for which the sampling methods in this paper are needed; *mini-domains* of 0.1 to 1 percent, estimates for which mostly require the use of statistical models; and *rare types* comprising less than 0.01 percent of the population, which generally cannot be handled by survey sampling methods. Many surveys aim to produce estimates for some major domains, some minor domains and occasionally even some mini-domains.

1. Graham Kalton, Westat, 1600 Research Blvd., Rockville, MD 20850, U.S.A. E-mail: grahamkalton@westat.com.

Since the sample sizes for most surveys are sufficient to produce estimates of reasonable precision for major domains, there is generally no need to adopt the kinds of oversampling procedures reviewed in this paper. However, there are some important design features that should be considered. It is, for example, valuable to take major domains into account in creating the strata for the survey. This consideration is of particular importance with geographically defined domains and multistage sampling. If a geographic domain is not made into a design stratum, the number of primary sampling units (PSUs) selected in that domain is a random variable; the sampled PSUs in strata that cut across the domain boundaries may or may not be in the domain, creating problems for domain estimation. It is also valuable to have a sizable number of sampled PSUs in each geographical domain in order to be able to compute direct variance estimates of reasonable precision, implying the need to spread the sample across a large number of PSUs. At the estimation stage, it is preferable, where possible, to apply nonresponse and noncoverage post-stratification-type adjustments at the domain rather than the national level. Singh, Gambino and Mantel (1994) and Marker (2001) discuss design issues and Rao (2003, pages 9-25) discusses estimation issues for major domains. Major domains will receive little attention in this paper.

At the other end of the size continuum, even with the use of special probability sampling methods, the sample sizes possible for most surveys are not large enough to produce standard design-based, or direct, estimates of characteristics for multiple domains when many of the domains are mini-domains or rare types. An obvious exception is a national population census, but censuses too have their limitations. Since they are conducted infrequently (in many countries only once a decade), their estimates are dated – a particular concern for mini-domains, which can experience rapid changes. Also, the content of a census must be severely limited in terms of the range of topics and depth of detail. Very large continuous surveys such as the American Community Survey (U.S. Census Bureau 2009a; Citro and Kalton 2007), the French rolling census (Durr 2005) and the German Microcensus (German Federal Statistical Office 2009) have been developed to address the need for more up-to-date data for small domains, but a restriction on content remains (although the content of the German Microcensus does vary over time). Other exceptions occur at the border between mini-domains and minor domains. For example, since 2007 the Canadian Community Health Survey has provided estimates on the health status of the populations of each of Canada's 121 health regions based on an annual survey of around 65,000 persons aged 12 and over, with the production of annual and biennial data files (Statistics Canada 2008). By combining the samples across multiple years, researchers are able to produce estimates for rare populations of various types.

In general, however, the maximum sample size possible for a survey on a specific topic is not adequate to yield a large set of mini-domain estimates of acceptable precision. Yet policy makers are making increasing demands for local area data at the mini-domain level. This demand for estimates for mini-domains, mainly domains defined at least in part by geographical administrative units, is being addressed by the use of statistical modeling techniques, leading to model-dependent, indirect, small area estimates. Thus, for example, the U.S. Census Bureau's Small Area Income and Poverty Estimates program produces indirect estimates of income and poverty statistics for 3,141 counties and estimates of poor school-age children for around 15,000 school districts every year, based on data now collected in the American Community Survey and predictor variables obtained from other sources available at the local area level, such as tax data (U.S. Census Bureau 2009b). A comprehensive treatment of indirect estimation using small area estimation techniques, a methodology that falls outside the scope of this paper, can be found in Rao (2003).

Apart from location sampling, discussed in Section 3.6, this paper also does not address the various methods that have been developed for sampling other types of mini-domains of much interest to social researchers and epidemiologists, domains that are often "hidden populations" in that the activities defining them are clandestine, such as intravenous drug use (Watters and Biernacki 1989). A range of methods has been developed under the assumption that the members of the mini-domains know each other. The broad class of such designs is termed link-tracing designs (see the review by Thompson and Frank 2000). They are adaptive designs in that the units are selected sequentially, with those selected at later stages dependent on those selected earlier (Thompson and Seber 1996; Thompson 2002).

Snowball sampling was one of the early methods of an adaptive, chain-referral sample design. It starts with some initial sample of rare domain members (the seeds), and they in turn identify other members of the domain. While it bears a resemblance to network (multiplicity) sampling (described in Section 3.5), snowball sampling lacks the probability basis of the latter technique, *i.e.*, known, non-zero, selection probabilities for all members of the domain. A version of snowball sampling has been termed respondent-driven sampling (RDS) (Heckathorn 1997, 2007). Volz and Heckathorn (2008) develop a theory for RDS that is based on four assumptions: (1) that respondents know how many members of the network are linked to them (the degree); (2) that respondents recruit others from their personal network at random; (3) that network connections are reciprocal; and

(4) that recruitment follows a Markov process. The need for these modeling assumptions for statistical inference is the difference between chain-referral sample designs and the conventional probability sample designs used in surveys which do not need to invoke such assumptions. It is apparent that RDS is appropriate only for mini-domains for which clear networks exist. The method is used mainly in local area settings, but Katzoff, Sirken and Thompson (2002) and Katzoff (2004) have suggested that the seeds could come from a large-scale survey, such as the U.S. National Health Interview Survey.

This paper focuses on the use of probability sampling methods to produce standard design-based, or direct, estimates for characteristics of rare populations, building on previous reviews (*e.g.*, Kish 1965a; Kalton and Anderson 1986; Kalton 1993a, 2003; Sudman and Kalton 1986; Sudman, Sirken and Cowan 1988; and Flores Cervantes and Kalton 2008). Much of the literature deals with the sampling issues that arise when the rare population is the sole subject of study. However, as noted above, surveys are often required to produce estimates for many different domains as well as for the total population. Section 2 reviews the design issues involved when the survey has design objectives for multiple domains whose members can be identified from the sampling frame. The main part of the paper, Section 3, provides a review of a range of methods that have been used to sample rare populations whose members cannot be identified in advance. The paper ends with some concluding remarks in Section 4.

2. Multi-domain allocations

The issue of sample allocation arises when a survey is being designed to produce estimates for a number of different domains, for subclasses that cut across the domains, as well as for the total population. In most applications, domains vary considerably in size with at least some of them being rare domains.

Assume that there are H mutually exclusive and exhaustive domains that are identified on the sampling frame. Under the commonly made assumptions that the variance of an estimate for domain h can be expressed as V/n_h and that survey costs are the same across domains, the optimum allocation for estimating the overall population mean is $n_h \propto W_h$, where W_h is the proportion of the population in domain h. Assuming that the domain estimates are all to have the same precision, the optimum allocation is $n_h = n/H$ for all domains. These two allocations are in conflict when the W_h vary greatly, as often occurs when the domains are administrative areas of the country, such as states, provinces, counties or districts. In such cases, adopting the optimum allocation for one objective leads to a serious loss of precision for the other. However, a compromise allocation that falls between the two optimum allocations often works well for both objectives.

Several compromise solutions exist. One, proposed by Kish (1976, 1988), is to determine the domain sample sizes by the following formula:

$$n_h \propto \sqrt{IW_h^2 + (1-I)H^{-2}},$$

where I and $(1-I)$ represent the relative importance of the national estimate and the domain (*e.g.*, administrative district) estimates, respectively. If $I = 1$, the allocation is a proportionate allocation, as optimum for the national estimate, whereas if $I = 0$, the allocation is an equal allocation, as optimum for the domain estimates. The choice of I is highly subjective, but I have found that $I = 0.5$ is often a good starting point, after which a careful review of the allocation can lead to modifications. Bankier (1988) has proposed a similar compromise solution, termed a power allocation. Applied to the current example, the domain sample sizes are determined from $n_h \propto W_h^q$, where q is a power between 0 (equal allocation) and 1 (proportionate allocation). As an example, the 2007 Canadian Community Health Survey was designed to attach about equal importance to the estimates for provinces and health regions. The sample allocation to a province was based on its population size and its number of health regions. Within a province, the sample was allocated between health regions using the Bankier allocation with $q = 0.5$ (Statistics Canada 2008).

A limitation to the Kish and Bankier procedures is that they may not allocate sufficient sample to small domains to produce estimates at the required level of precision. This limitation can be addressed by revising the initial allocations to satisfy precision requirements. An alternative approach addresses this limitation directly: the allocation is determined by fixing a core sample that will satisfy one of the objectives and then supplementing that sample as needed to satisfy the other objective. Singh, Gambino and Mantel (1994) describe such a design for the Canadian Labour Force Survey, with a core sample to provide national and provincial estimates and, where needed, supplemental samples to provide subprovincial estimates of acceptable precision.

The Kish and Bankier schemes assume that the same precision level is required for all small domains. Longford (2006) describes a more general approach in which 'inferential priorities' P_d are assigned to each domain d. As an example, he proposes setting the priorities as $P_d = N_d^a$, where N_d is the population size of domain d and a is a value chosen between 0 and 2. The value $a = 0$ corresponds to the Kish and Bankier equal domain sample size assumption and $a = 2$ corresponds to an overall proportionate

allocation. An intermediate value of a attaches greater priority to larger domains. Longford also extends the approach to incorporate an inferential priority for the overall estimate.

A more general approach to sample allocation is via mathematical programming, as has been proposed by a number of researchers (see, for example, Rodriguez Vera 1982). This approach can accommodate unequal variances across domains, intersecting domains, and multiple estimates for each domain. The U.S. Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) provides an example with intersecting domains, with the sample selected from birth certificate records that contained the requisite domain information. There were 10 domains of interest for the ECLS-B: births classified by race (5 domains), birth weight (3 domains) and twins or non-twins (2 domains). The approach adopted first determined a minimum effective sample size (*i.e.*, the actual sample size divided by the design effect) for each domain. With the 30 cells of the cross-classification of birth weight, race/ethnicity and twin/non-twin treated as strata, an allocation of the sample across the strata was then determined to minimize the overall sample size while satisfying the effective sample size requirements for all the domains (Green 2000).

When there are multiple domains of interest and multi-stage sampling is to be used, a variant of the usual measure of size for probability proportional to size (PPS) sampling can be useful for controlling the sample sizes in the sampled clusters (PSUs, second-stage units, *etc.*), provided that reasonable estimates of the domain population sizes are available by cluster. The requirements that all sampled clusters have approximately the same overall subsample size and that sampled units in each domain have equal probabilities of selection can both be met by sampling the clusters with standard PPS methods, but with a composite measure of size that takes account of the differing sampling rates for different domains (Folsom, Potter and Williams 1987). As an example, in a survey of men in English prisons, the desired sampling fractions were 1 in 2 for civil prisoners (C), 1 in 21 for "star" prisoners who are normally serving their first term of imprisonment (S) and 1 in 45 for recidivists (R). Prisons were selected at the first stage of sampling, with prison i being selected with probability proportional to its composite measure of size $R_i + 2.2S_i + 20.3C_i$, where the multipliers are the sampling rates relative to the rate for recidivists (Morris 1965, pages 303-306).

3. Methods for oversampling rare domains

The main focus of this paper is on the use of probability sampling methods to produce standard design-based, or direct, estimates for characteristics of rare populations, often minor domains in Kish's terminology. As preparation for the subsequent discussion, it will be useful to note some features of different types of rare populations that, together with the survey's mode of data collection, are influential in the choice of sampling methods that can be applied to generate required sample sizes for all domains. Some important features for consideration are summarized below:

- Is a separate frame(s) available for sampling a rare population? Can those sampled be located for data collection? How up-to-date and complete is the frame? If an existing up-to-date frame contains only the rare population (with possibly a few other listings) and provides almost complete coverage, then sampling can follow standard methods. If no single frame gives adequate coverage but there are multiple frames that between them give good coverage, issues of multiple routes of selection arise (Section 3.4).
- Is the rare population concentrated in certain, identifiable parts of the sampling frame, or is it fairly evenly spread throughout the frame? If it is concentrated, disproportionate stratification can be effective (Section 3.2).
- If a sample is selected from a more general population, can a sampled person's membership in the rare population be determined inexpensively, such as from responses to a few simple questions? If so, standard screening methods may be used (Section 3.1). If accurate determination requires expensive procedures, such as medical examinations, a two-phase design may be useful (Section 3.3). A related issue is whether some members of a rare population consider their membership to be sensitive; the likelihood that members may be tempted to deny their membership may influence the choice of survey administration mode and other aspects of screening.
- Are members of the rare population readily identified by others? If so, some form of network, or multiplicity, sampling may be useful (Section 3.5).
- Are members of the rare population to be found at specific locations or events? If so, location sampling may be useful (Section 3.6).
- Is the rare population defined by a constant characteristic (*e.g.*, race/ethnicity) or by a recent event (*e.g.*, a hospital stay)? The distinction between these two types of characteristics is important in considering the utility of panel surveys for sampling rare populations (Section 3.7).

The following sections review a range of methods for sampling rare populations. Although the methods are discussed individually, some are interrelated and, in practice, a combination of methods is often used.

3.1 Screening

Some form of screening is generally needed when the sampling frame does not contain domain identifiers. This section considers a straightforward application of a screening design in which a large first-phase sample is selected to identify samples of the members of the domains of interest, without recourse to the techniques described in later sections. The first-phase sample size is the minimum sample size that will produce the required (or larger) sample sizes for all of the domains. The minimum first-phase sample size is determined by identifying the required sample size for one of the domains, with all of the sample members of that domain then being included in the second-phase sample. Subsamples of other domains are selected for the second-phase sample at rates that generate the required domain sample sizes. If the survey is designed to collect data for only a subset of the domains (often only one domain), then none of the members of the other domains is selected for the second-phase sample.

Since a very large screening sample size is needed to generate an adequate domain sample size when one (or more) of the domains of interest is a rare population, the cost of screening becomes a major concern. In addition to the sampling methods discussed in later sections, there are several strategies that can be employed to keep costs low:

- Use an inexpensive mode of data collection, such as telephone interviewing or a mail questionnaire, for the screening. The second-phase data collection may be by the same mode or a different mode.
- When possible and useful, permit the collection of screening data from persons other than those sampled. For example, other household members may be able to accurately report the rare population status of the sampled member. See the discussion below and also Section 3.5 on multiplicity sampling.
- When screening is carried out by face-to-face interviewing in a multistage design, it is efficient to select a large sample size in each cluster. Compact clusters can also be used. Costs are reduced, and the precision of domain estimates is not seriously harmed because the average domain sample sizes in the clusters will be relatively small.

One possible means of reducing screening costs is to share the costs across more than one survey. For instance, the child component of the ongoing U.S. National Immunization Survey (NIS) is a quarterly telephone survey that screens households with landline telephone numbers to locate children aged 19 to 35 months, in order to ascertain vaccination coverage levels (Smith, Battaglia, Huggins, Hoaglin, Roden, Khare, Ezzati-Rice and Wright 2001; U.S. National Center for Health Statistics 2009b). The NIS large-scale screening is also used to identify members of domains of interest for the State and Local Area Integrated Telephone Survey (SLAITS) program, which addresses a variety of other topics over time (U.S. National Center for Health Statistics 2009c). When sharing screening costs across a number of surveys, it is advantageous if the domains for the surveys are fairly disjoint sets in order to minimize the problems associated with screening some respondents into more than one survey.

When no one is at home to complete a face-to-face screening for a household, it may be possible to obtain information from knowledgeable neighbors as to whether the household contains a member of the rare population (*e.g.*, a child under 3 years of age). This approach (which is used in NHANES) can appreciably reduce data collection costs when a large proportion of the households do not contain members of the rare population. However, there is a danger that the approach may result in undercoverage; some protection is provided by requiring that, if the first neighbor interviewed indicates that the household does not include a member of the rare population(s), the other neighbor is also interviewed. Ethical issues also must be considered, particularly for the identification of rare populations that are sensitive in nature.

An extension of the approach of collecting screening information from neighbors is known as focused enumeration. This technique, which is a form of multiplicity sampling (see Section 3.5), involves asking the respondent at each sampled, or "core", address about the presence of members of the rare population in the n neighboring addresses on either side. In essence, the sample consists of $2n + 1$ addresses for each core address. If the respondent is unable to provide the screening information for one or more of the linked addresses, then the interviewer must make contact at another address. Focused enumeration has been used with $n = 2$ in the British Crime Survey (Bolling, Grant and Sinclair 2008) and the Health Survey of England (Erens, Prior, Korovessis, Calderwood, Brookes and Primatesta 2001) to oversample ethnic minorities. A limitation of the technique is that it will likely produce some (possibly substantial) undercoverage. Evidence of the extent of undercoverage can be obtained by comparing the prevalence of the rare population in the core sample with that in the linked addresses.

In surveys that sample persons by first sampling households, survey designers often prefer to select one person per

household – perhaps allowing two persons to be sampled in large households – to avoid contamination effects and prevent a within-household clustering homogeneity effect on design effects. This design is not always the best (Clark and Steel 2007), and this particularly applies when rare populations are sampled. When rare population members are concentrated in certain households (*e.g.*, minority populations), the size of the screening sample can be appreciably reduced if more than one person – even all eligible persons – can be taken in some households (see Hedges 1973). Elliott, Finch, Klein, Ma, Do, Beckett, Orr and Lurie (2008) suggest that, for oversampling American Indian/Alaskan Native and Chinese minorities in the United States, taking all eligible persons in a household has potential for U.S. health surveys. The NHANES maximizes the number of sampled persons per household. Since each respondent is remunerated for participation, households with more respondents receive more remuneration, a factor thought to increase response rates (Mohadjer and Curtin 2008). Note that within-household homogeneity will have little effect on design effects when the data are analyzed by subgroup characteristics (*e.g.*, age and sex) that cut across households.

The use of large-scale screening to identify rare populations raises three issues, each of which could lead to a failure to achieve planned sample sizes unless precautions are taken. The first results from the fact that, with screening, the sample size for a rare population is a random variable. As a result, the achieved sample size may be larger or smaller than expected. When a minimum sample size is specified for a rare population, it may be wise to determine the sampling fraction to be used to ensure that there is, say, a 90 percent probability that the achieved sample size will be at least as large as the specified minimum. This procedure was used in determining the sampling fractions for the many age, sex and income subdomains for the Continuing Survey of Food Intakes by Individuals 1994-96 (Goldman, Borrud and Berlin 1997).

The second issue raised by large-scale screening is that the overall nonresponse rate must be considered. A sampled member of a rare population will be a nonrespondent if the screener information is not obtained, or if a member of the rare population is identified (perhaps by a proxy informant) but does not respond to the survey items. The overall nonresponse rate may well be much higher than would occur without the screening component. Furthermore, the survey designers must consider the nature of the rare domain and the ways in which members of that domain will react to the survey content. A survey in which new immigrants are asked about their immigration experiences might have a very different response rate than a survey in which war veterans are asked about the medical and other support services they are receiving.

The third issue is that noncoverage can be a significant problem when large-scale screening is used to identify rare populations. One source of noncoverage relates to the sampling frame used for the screener sample. Even though a frame has good overall coverage, its coverage of a rare domain may be inadequate. For example, the noncoverage of a frame of landline telephone numbers is much higher for households of younger people than for the total population. The designers of landline telephone surveys of such rare domains as young children and college students therefore must carefully consider the potential for noncoverage biases. To address the problem of the substantial noncoverage of poor people in telephone surveys, the National Survey of America's Families, which was designed to track the well-being of children and adults in response to welfare reforms, included an area sample of households without telephones in conjunction with the main random digit dialing (RDD) telephone sample (Waksberg, Brick, Shapiro, Flores Cervantes and Bell 1997).

Another source of noncoverage is a failure to identify some members of the rare population at the screening stage. In particular, when a survey aims to collect data only for members of a rare domain, some screening phase respondents may falsely report, and some interviewers may falsely record, that the sampled persons are not members of that domain. These misclassifications may be inadvertent or they may be deliberately aimed at avoiding the second-phase data collection. Misclassification error can give rise to serious levels of noncoverage, particularly when the rare population classification is based on responses to several questions, misreports to any one of which leads to a misclassification (Sudman 1972, 1976). When the survey oversamples one or more rare domains as part of a survey of the general population, misclassifications are uncovered at the second phase, thus avoiding noncoverage. However, misclassifications still result in a smaller sample sizes for rare domains; in addition, the variation in sampling weights between respondents selected as members of the rare domain and those sampled as members of another domain can lead to a serious loss of precision. Noncoverage is more likely to arise when screener data are collected from proxy informants. It is a particular problem with focused enumeration.

In a number of surveys of rare populations, the proportion of rare population members identified has been much lower than prevalence benchmarks. For example, the 1994 NIS had an appreciable shortfall in the identified proportion of children aged 19 to 35 months (4.1 percent compared to the predicted rate of 5 percent) (Camburn and Wright 1996). In the National Longitudinal Survey of Youth of 1997, only 75 percent of youth aged 12 to 23 years were located (Horrigan, Moore, Pedlow and Wolter 1999). These findings could be the result of higher nonresponse rates for

members of the rare population, frame noncoverage of various types, or misclassifications of domain membership. To produce the required sample size, an allowance for under-representation must be made at the design stage. The noncoverage of an age domain appears to be greatest at the domain boundaries, perhaps because respondents do not know exact ages (with those falsely screened out being lost and those falsely screened in being detected and dropped later) or because of deliberate misreporting to avoid the follow-up interview. To counteract this effect, it can be useful to start with an initial screening for all household members or for a broader age range and then narrow down to the required age range later on.

Weighting adjustments can be used in an attempt to mitigate biases caused by nonresponse and noncoverage, but they are necessarily imperfect. Adjustments for a domain specific level of nonresponse require knowledge of the domain membership of nonrespondents, but that is often not available. Adjustments for noncoverage of a rare domain require accurate external data for the domain, data that are often not available. Indeed, one of the purposes for some rare domain surveys is to estimate the domain size. Noncoverage is a major potential source of error in the estimation of domain size.

3.2 Disproportionate stratification

A natural extension of the screening approach is to try to identify strata where the screening will be more productive. In the ideal circumstance, one or more strata that cover all of the rare population and none from outside that population are identified. That case requires no screening process. Otherwise, it is necessary to select samples from all the strata (apart from those known to contain no rare population members) to have complete coverage of the rare population. The use of disproportionate stratification, with higher sampling fractions in the strata where the prevalence of the rare population is higher, can reduce the amount of screening needed.

3.2.1 Theoretical background

Consider initially a survey designed to provide estimates for a single rare population. Waksberg (1973) carried out an early theoretical assessment of the value of disproportionate stratification for this case. Subsequent papers on this topic include those by Kalton and Anderson (1986) and Kalton (1993a, 2003). The theoretical results show that three main factors must be considered in determining the effectiveness of disproportionate stratification for sampling a single rare population: the prevalence rate in each stratum, the proportion of the rare population in each stratum, and the ratio of the full cost of data collection for members of the rare population to the screening cost involved in identifying members of that population. If it is assumed that (1) the element variances for the rare population are the same across strata and (2) the costs of data collection for members of the rare and non-rare populations are the same across strata, then, with simple random sampling within strata, the optimum sampling fraction in stratum h for minimizing the variance of an estimated mean for the rare population, subject to a fixed total budget, is given by

$$f_h \propto \sqrt{\frac{P_h}{P_h(c-1)+1}},$$

where P_h is the proportion of the units in stratum h that are members of the rare population and c is the ratio of the data collection cost for a sampled member of the rare population to the cost for a member of the non-rare population (Kalton 1993a). The following formula provides the ratio of the variance of the sample mean with the optimum disproportionate stratified sampling fractions to that with a proportionate stratified sample of the same total cost:

$$R = \frac{\left[\Sigma A_h \sqrt{P(c-1)+P/P_h}\right]^2}{P(c-1)+1},$$

where A_h is the proportion of the rare population in stratum h and P is the prevalence of the rare population in the full population.

In general, the variability in the optimum sampling fractions across the strata, and the gains in precision for the sample mean, decline as c increases. Thus, if the main survey data collection cost is high – as, for instance, when the survey involves an expensive medical examination – or if the screening cost is very low, then disproportionate stratification may yield only minor gains in precision.

When the main data collection cost adds nothing to the screening cost, the ratio of main data collection cost to screening cost will be $c = 1$. In this limiting situation, the formulas given above simplify to $f_h \propto \sqrt{P_h}$ and $R = (\Sigma\sqrt{A_h W_h})^2$, where W_h is the proportion of the total population in stratum h. These simple formulas provide a useful indication of the maximum variation in optimum sampling fractions and the maximum gains in precision that can be achieved. The square root function in the optimum sampling fraction formula makes clear that the prevalences in the strata must vary a good deal if the sampling fractions are to differ appreciably from a proportionate allocation. For example, even if the prevalence in stratum A is four times as large as that in stratum B, the optimum sampling fraction in stratum A is only twice as large as that in stratum B. The gains in precision $(1 - R)$ are large when A_h is large when W_h is small and vice versa. With only two strata, a stratum with a prevalence five times as large as the overall prevalence (*i.e.*, $P_h / P = 5$) will yield gains in precision of 25 percent or more ($(1 - R) \geq 0.25$) only if that stratum

includes at least 60 percent of the rare population (Kalton 2003, Table 1).

In summary, while generally useful, disproportionate stratification will yield substantial gains in efficiency only if three conditions hold: (1) the rare population must be much more prevalent in the oversampled strata; (2) the oversampled strata must contain a high proportion of the rare population; and (3) the cost of the main data collection per sampled unit must not be high. In many cases, not all of these three conditions can be met, in which case the gains will be modest.

Furthermore, the results presented above are based on the assumption that the true prevalence of the rare population in each stratum is known, whereas in practice it will be out of date (for example, based on the last census) or will perhaps simply have been guesstimated. Errors in the prevalence estimates will reduce the precision gains achieved with disproportionate stratification and could even result in a loss of precision. A major overestimation of the prevalence of the rare population, and hence of the optimum sampling fraction, in the high-density stratum can result in a serious loss of precision for the survey estimates. It is therefore often preferable to adopt a conservative strategy, that is, to adopt a somewhat less disproportionate allocation, one that moves in the direction of a proportionate allocation.

3.2.2 Applications

When area sampling is used, data available from the last census and other sources can be used to allocate the area clusters to strata based on their prevalence estimates for the rare population. See Waksberg, Judkins and Massey (1997) for a detailed investigation of this approach for oversampling various racial/ethnic populations and the low-income population using U.S. census blocks and block groups as clusters. Based on data from the 1990 Census, Waksberg and his colleagues found that the approach generally worked well for Blacks and Hispanics but not for the low-income population. While the low-income population did exhibit high concentrations in some blocks and block groups, those areas did not cover a high proportion of that population.

When the survey designers have access to a list frame with names, the names can be used to construct strata of likely members of some racial/ethnic groups. This situation arises, for instance, with lists of names and telephone numbers and when names are merged onto U.S. Postal Service (USPS) Delivery Sequence File addresses (no name merge is made in some cases). The allocation to strata can be based on surnames only or on a combination of surname and first name (and even other names also). Since women often adopt their husbands' surnames, the allocation is generally more effective for men than women. Names can be reasonably effective for identifying Hispanics, Filipinos, Vietnamese, Japanese and Chinese, but not Blacks. A number of lists of names associated with different racial/ethnic groups have been compiled, such as the list of Spanish names compiled by the U.S. Census Bureau for the 1990s (Word and Perkins 1996). Several commercial vendors have developed complex algorithms to perform racial/ethnic classifications based on names (see Fiscella and Fremont 2006 for further details). The use of names in identifying race and ethnicity has been of considerable interest to epidemiologists and demographers, who have conducted a number of evaluations of this method (*e.g.*, Lauderdale and Kestenbaum 2000; Elliott, Morrison, Fremont, McCaffrey, Pantoja and Lurie 2009). They often assess the effectiveness of the method in terms of positive predictive value and sensitivity, which are the equivalents of prevalence and the proportion of members of the domain who are identified as such by the instrument used for the classification. In the sampling context, besides limitations in the instrument, researchers also need to take into account that sometimes names are not available and that some available names may be incorrect (for example, with address-based sampling, the names may be out-of-date, because the original family has moved out and a new family has moved into an address). These additional considerations serve to reduce the effectiveness of the name stratification, and depending on the particular circumstances, the reduction in effectiveness may be sizable.

As with stratification in general, the stratification factors used for sampling rare populations do not have to be restricted to objective measures. They can equally be subjective classifications. The only consideration is how well they serve the needs of the stratification (see Kish 1965b, pages 412-415, for an example of the effectiveness of the use of listers' rapid classifications of dwellings into low, medium or high socio-economic status for disproportionate stratification). Elliott, McCaffrey, Perlman, Marshall and Hambarsoomians (2009) describe an effective application of subjective stratification for sampling Cambodian immigrants in Long Beach, California. A local community expert rated all individual residences in sampled blocks as likely or unlikely to contain Cambodian households, based on externally observable cultural characteristics such as footwear outside the door and Buddhist altars. The residences allocated to the "likely" stratum (approximately 20 percent) were then sampled at four times the rate than the rest.

Sometimes, when the survey is concerned with producing estimates only for a very rare population, disproportionate stratification may still require an excessive amount of screening. In that circumstance, it may be necessary to sample from the strata where the prevalence is

highest, dropping the other strata and accepting some degree of noncoverage (or redefining the survey population to comprise only members of the rare population in the strata that were sampled). The Hispanic Health and Nutrition Examination Survey of 1982-84 (HHANES) provides an illustration. For its samples of Mexican Americans in the Southwest and Puerto Ricans in the New York City area, the HHANES sampled only from counties with large numbers and/or percentages of Hispanics, based on 1980 Census counts (Gonzalez, Ezzati, White, Massey, Lago and Waksberg 1985).

As another example of this approach, Hedges (1979) describes a procedure for sampling a minority population that is more concentrated in some geographical districts, such as census enumeration districts. In this procedure, the districts are listed in order of their prevalence of members of the rare population (obtained, say, from the last census), and then the survey designers produce Lorenz curves of the cumulative distribution of rare population prevalence and the cumulative distribution of the proportions of rare population members covered. With the cumulative prevalence declining as the cumulative coverage increases, the survey designers can use these distributions to select the combination of prevalence and proportion covered that best fulfills their requirements. The issue then to be faced is whether to make inferences to the covered population, or whether to make inferences to the full population by applying population weighting adjustments in an attempt to address the noncoverage bias.

When a domain is very rare but a portion of it is heavily concentrated in a stratum, researchers sometimes sample that stratum at a rate much higher than the optimum in order to generate a sizable number of cases. Although this approach may produce a large sample of the rare population, the effective sample size (*i.e.*, the sample size divided by the design effect) will be smaller than if the optimum sampling fractions had been used. Thus, from the perspective of the standard survey design-based mode of inference, this approach is not appropriate. However, the researchers using this approach often argue for a model-based mode of inference in which the sampling weights are ignored. In my view, ignoring the sampling weights is problematic. However, discussion of this issue is outside the scope of this paper.

3.3 Two-phase sampling

The screening approach treated in Sections 3.1 and 3.2 assumes that identification of rare population members is relatively easy. When accurate identification is expensive, a two-phase design can be useful, starting with an imperfect screening classification at the first phase, to be followed up with accurate identification for a disproportionate stratified subsample at the second phase. Whether the two-phase approach is cost-effective depends in part on the relative costs of the imperfect classification and accurate identification: since the imperfect classifications use up some of the study's resources, they must be much less expensive than the accurate identification. Deming (1977) suggests that the ratio of the per-unit costs of the second- to the first-phase data collections should be at least 6:1. Also, the imperfect classification must be reasonably effective in order to gain major benefits from a second-phase disproportionate stratification.

Two- or even three-phase sampling can often be useful in medical surveys of persons with specific health conditions. The first phase of the survey often consists of a screening questionnaire administered by survey interviewers, and the second phase is generally conducted by clinicians, often in a medical center. As one example, in a survey of epilepsy in Copiah County, Mississippi, Haerer, Anderson and Schoenberg (1986) first had survey interviewers administer to all households in the county a questionnaire that had been pretested to ensure that it had a high level of sensitivity for detecting persons with epilepsy. To avoid false negatives at this first phase, a broad screening net was used in identifying persons who would continue to the second phase. All those so identified were the subjects for the second phase of the survey, which consisted of brief neurological examinations conducted by a team of four senior neurologists in a public health clinic.

A second example illustrates the use of another survey to serve as the first-phase data collection for studying a rare domain. In this case, the Health and Retirement Study (HRS) was used as the first phase for a study of dementia and other cognitive impairment in adults aged 70 or older. The HRS collects a wide range of measures on sample respondents, including a battery of cognitive measures. Using these measures, the HRS respondents were allocated to five cognitive strata, with a disproportionate stratified sample being selected for the second phase. The expensive second-phase data collection consisted of a 3- to 4-hour structured in-home assessment by a nurse and neuropsychology technician. The results of the assessment were then evaluated by a geropsychiatrist, a neurologist and a cognitive neuroscientist to assign a preliminary diagnosis for cognitive status, which was then reassessed in the light of data in the person's medical records (Langa, Plassman, Wallace, *et al.* 2005).

A third example is a three-phase design that was used in a pilot study to identify persons who would qualify for disability benefits from the U.S. Social Security Administration if they were to apply for them (Maffeo, Frey and Kalton 2000). At the first phase, a knowledgeable household respondent was asked to provide information about the

disability beneficiary status and impairment status of all adults aged 18 to 69 years in the household. At the second phase, all those classified into a stratum of severely disabled nonbeneficiaries and samples of the other strata were interviewed in person and were then reclassified as necessary into likely disability strata for the third phase. At the third phase, a disproportionate stratified sample of persons was selected to undergo medical examinations in mobile examination centers.

A fairly common practice with two-phase designs is to take no second- (or third-) phase sample from the stratum of those classified as nonmembers of the rare domain based on their responses at the previous phase. The proportion of the population in that stratum is usually very high, and the prevalence of the rare domain in it is very low (indeed, as in the Haerer, Anderson and Schoenberg (1986) study, the stratum is often conservatively defined with the aim of avoiding the inclusion of those who might possibly be members of the rare domain). As a result, a moderate-sized sample from this stratum will yield almost no members of the rare domain. However, the cut-off strategy of taking no sample from this stratum is risky. If the prevalence of the rare domain in this large stratum is more than minimal, a substantial proportion of the domain may go unrepresented in the final sample.

3.4 Multiple frames

Sometimes sampling frames exist that are more targeted on a rare population than a general frame, but they cover only part of the rare population. In this situation, it can be efficient to select the sample from more than one frame. For example, in the common case of oversampling ethnic minorities, there is sometimes a list frame available. The persons on the list can be classified based on their names as being likely to belong to a given ethnic group (*e.g.*, Chinese, Korean, Pacific Islanders, Vietnamese) to create a second, incomplete sampling frame from which to sample, in addition to a more complete frame that has a lower prevalence of the rare population (see, *e.g.*, Elliott *et al.* 2008; Flores Cervantes and Kalton 2008). As with disproportionate stratification (Section 3.2), major benefits derive from this approach only when the second frame has a high prevalence and covers a sizable fraction of the rare population. See Lohr (2009) for a review of the issues involved in sampling from multiple frames.

With multiple frames, some members of the rare population may be included on several frames, in which case they may have multiple routes of being selected into the sample. There are three broad approaches for addressing these multiplicities (Anderson and Kalton 1990; Kalton and Anderson 1986). When all the frames are list frames, as sometimes occurs in health studies, it may be possible to combine the frames into a single unduplicated list; however, this can often involve difficult record linkage problems. An alternative approach is to make the frames non-overlapping by using a unique identification rule that associates each member of the rare population with only one of the frames, treating the listings on the other frames as blanks (Kish 1965b, pages 388-390). Samples are selected from each of the frames without regard to the duplication, but only the non-blank sampled listings are accepted for the final sample. This approach works best when searches can be made for each sampled unit on the other frames; if the frames are put in a priority order and the unit is found on a prior frame to the one from which the selection was made, the sampled listing would be treated as a blank. In this case, the frames are strata; the sampled units are treated as subclasses within the strata, allowing for the blank listings (Kish 1965b, pages 132-139), and the analysis follows standard methods.

The use of the unique identification approach can, however, be inefficient when the persons sampled from one frame have to be contacted to establish whether their listings are to be treated as real or blank for that frame. In this case, it is generally more economical to collect the survey data for all sampled persons (*i.e.*, to accept the multiple routes of selection). There are, however, exceptions, as in the case of the National Survey of America's Families. That survey used a combination of an area frame and an RDD telephone frame, with the area frame being used to cover only households without telephones (Waksberg, Brick *et al.* 1997). It proved to be efficient to conduct a quick screening exercise with households on the area frame to eliminate households with telephones, retaining only the non-telephone households for the survey.

There are two general approaches for taking multiple routes of selection into account in computing selection probabilities (Bankier 1986; Kalton and Anderson 1986). One method calculates each sampled unit's overall selection probability across all the frames and uses the inverse of that probability as the base weight for the analysis (leading to the Horvitz-Thompson estimator). For example, the overall selection probability for sampled unit i on two frames is $p_i = (p_{1i} + p_{2i} - p_{1i}p_{2i}) = [1 - (1 - p_{1i})(1 - p_{2i})]$, where p_{fi} is the probability of the unit's selection from frame $f = 1, 2$. A variant is to replace the overall selection probability with the expected number of selections (leading to the Hansen-Hurwitz estimator), which is easier to compute when multiple frames are involved. With only two frames, the expected number of selections is $(p_{1i} + p_{2i})$. When selection probabilities are small, there is little difference between these two estimators.

Adjustments to compensate for nonresponse and to calibrate sample totals to known population totals can either be made to the overall selection probabilities p_i or they can

be made to the p_{fi} individually. A problem that can occur is that the survey designers do not know whether a nonresponding unit sampled from one frame is on another frame since that information is only collected in the interview. In this situation the p_i for nonresponding units cannot be directly computed and must be estimated in some fashion. When adjustments are made to the p_{fi} individually, it is not possible to form nonresponse weighting classes that take membership on other frames into account. Instead, the designers must assume that, within weighting classes, the response rates are the same no matter how many frames a unit is on.

In general, the application of the approach described above requires knowledge of each sampled unit's selection probabilities for all of the frames, information that is not always available. When selection probabilities are not known for frames other than the frame(s) from which the unit is sampled (but presence/absence on the frames is known), an alternative approach, termed a weight share method by Lavallée (1995, 2007), can be used. Unbiased estimates of population totals are obtained if the weight for unit i is given by $w_i = \Sigma_j \alpha_{ij} w'_{ij}$ where α_{ij} are any set of constants such that $\Sigma_j \alpha_{ij} = 1$ when summed across the j frames, $w'_{ij} = 1/p_{ij}$ if unit i is selected from frame j with probability p_{ij} and $w'_{ij} = 0$ otherwise (Kalton and Brick 1995; Lavallée 2007). For many applications, it is reasonable to set $\alpha_{ij} = \alpha_j$ and then a good choice of α_j is $\alpha_j = \tilde{n}_j / \Sigma \tilde{n}_j$, where $\tilde{n}_j$ is the effective sample size based on some average design effect (Chu, Brick and Kalton 1999).

The second general approach for dealing with multiple routes of selection uses the multiple-frame methodology introduced by Hartley (1974), and the subject of much recent research (see, *e.g.*, Lohr and Rao 2000 and 2006 and the references cited in those papers). In the case of two frames (A and B), the population can be divided into three mutually exclusive subsets labeled $a = A \cap \overline{B}$, $b = \overline{A} \cap B$ and $ab = A \cap B$. The sample can be divided into samples from a, b and ab, where the ab sample can be separated into respondents sampled from frame A and those sampled from frame B. The samples in subsets a and b have only one route of selection, and hence are readily handled in estimation. Totals for ab could be estimated from the sample from frame A or the sample from frame B, say, $\hat{Y}^A_{ab}$ or $\hat{Y}^B_{ab}$. The Hartley methodology takes a weighted average of these two estimators, $\hat{Y}_{ab} = \theta \hat{Y}^A_{ab} + (1 - \theta)\hat{Y}^B_{ab}$, where θ is chosen to minimize the variance of $\hat{Y}_{ab}$, taking into account that sample sizes and design effects differ between the two samples. Note that the dual-frame methodology is estimator specific, with different values of θ for different estimators. Skinner (1991), Skinner and Rao (1996) and Lohr and Rao (2006) have proposed an alternative, pseudo-maximum likelihood estimation approach that has the attraction of avoiding the problems associated with different values of θ for different variables. Wu and Rao (2009) propose a multiplicity-based pseudo empirical likelihood approach for multiple frame surveys, including what they term a single-frame multiplicity-based approach that incorporates Lavallée's weight share method as described above.

When a dual- or multiple-frame design is used, it is often the case that one frame has complete coverage but a low prevalence of the rare population (*e.g.*, an area frame) and the other frame(s) has a much higher prevalence of the rare population but incomplete coverage. Metcalf and Scott (2009), for example, combined an area sample with an electoral roll sample for the Auckland Diabetes, Heart and Health Survey, in which Pacific Islanders, Maoris and older people were domains of special interest. The electoral roll frame had the advantage of containing information about electors' ages, as well as a special roll on which those who considered themselves to be of Maori descent could enroll. Furthermore, many people of Pacific descent could likely be identified by their names, since Pacific languages use fewer letters than English. A disproportionate stratified sample was selected from the electoral roll frame to oversample the domains of interest, and the sample from the area frame brought in people not on the electoral rolls.

The National Incidence Study of Child Abuse and Neglect provides an example of a more complex situation (Winglee, Park, Rust, Liu and Shapiro 2007). That survey used many frames to increase its overall coverage of abused and neglected children. Child Protective Services (CPS) agencies in the sampled PSUs were the basis of the main sampling frame, while police, hospitals, schools, shelters, daycare centers and other agencies were the sources of other frames. The samples from CPS agencies were selected from list frames, but the samples from other agencies were drawn by sampling agencies, constructing rosters of relevant professional staff, and sampling staff who acted as informants about maltreated children. With these procedures, duplication across agencies cannot be ascertained, except in the case of CPS agencies and any of the other agencies. The design was therefore treated as a dual-frame design, with CPS as one frame and the combination of the other frames as the second frame (*i.e.*, assuming no overlap between the other frames).

3.5 Network sampling

Network (or multiplicity) sampling expands on the standard screening approach by asking sampled persons (or addresses) to also serve as proxy informants to provide the screening information for persons who are linked to them in a clearly specified way (Sudman *et al.* 1988; Sirken 2004, 2005). Relatives such as parents, siblings and children are

often used as the basis of linkages. A key requirement is that every member of the linkage must know and be willing to report the rare population membership statuses of all those linked to them. In a pilot study of male Vietnam veterans, Rothbart, Fine and Sudman (1982) included aunts and uncles as informants as well as parents and siblings, but found that aunts and uncles identified far fewer Vietnam veterans than expected. This apparent failure of aunts and uncles to report some veterans gives rise to a potential sampling bias, thus making their inclusion in the linkage rules problematic.

The multiple routes of selection with network sampling need to be taken into account in determining selection probabilities in a similar manner to that described for multiple frames in the previous section. Conceptually, one can consider each member of the rare population divided into, say, l parts corresponding to the l informants for that member; it is then these parts that are sampled for the survey. See Lavallée (2007) for some theory behind the technique.

When network sampling is used in surveys that collect data on the characteristics of rare population members, direct contact must be made with the members of the rare population identified by the initial informant. In this case, the informant has to be able to provide contact information for the rare population members. The linkage definition may be structured to facilitate the follow-up data collection. For example, with face-to-face interviewing, the linkage may be restricted to relatives living in a defined area close to the informant.

Sudman and Freeman (1988) describe the application of network sampling in a telephone survey about access to health care, in which an oversample of persons with a chronic or serious illness was required. During an initial contact with the head of the household, linkages to the respondent's or spouse's parents, stepparents, siblings, grandparents and grandchildren under age 18 were identified and data were collected on their health status. The use of this network sampling design increased the number of chronically or seriously ill adults identified by about one-third. However, about one in eight of the initial network informants with relatives were unable or unwilling to provide illness information for their network members, and 70 percent did not provide complete location information, including 28 percent who provided neither name nor location information (thus making tracing impossible). The use of network sampling led to some false positives (persons reported as being chronically or seriously ill by the initial respondent but reporting themselves as well). A more serious concern is that the survey was not able to provide information on false negatives (this would have required following up a sample of network members reported to be well by the initial informant).

Some forms of linkage have the added benefit that they can incorporate some rare population members who are not on the original sampling frame and would therefore otherwise be a component of noncoverage. For example, Brick (1990) describes a field test for the telephone-based National Household Education Survey (NHES) that used multiplicity sampling to increase the sample of 14- to 21-year-olds, with a focus on school drop-outs. In a subsample of households, all women aged 28 to 65 were asked to provide information for all their 14- to 21-year-old children currently living elsewhere. Some of these children lived in telephone households and hence had two routes of selection. Others lived in non-telephone households and hence would not have been covered by the survey; their inclusion via the multiplicity design increased the coverage rate in 1989 by about 5 percent. However, the response rate for out-of-household youth was much lower than that for in-household youth because of failure to reach the youth, particularly the youth living in non-telephone households.

Tortora, Groves and Peytcheva (2008) provide another example, in this case using multiplicity sampling in an attempt to cover persons with only mobile telephones via an RDD sample of landline telephone numbers. Respondents to the RDD survey (itself a panel survey) were asked to provide information about parents, siblings and adult children living in mobile-only households. The results demonstrate some of the general issues with multiplicity sampling: knowledge about the mobile-only status of the network members depended on the cohesion of the network; there was widespread unwillingness to provide mobile telephone numbers; and many of those identified as mobile-only households in fact also had a landline telephone.

Network sampling has not been widely used in practice for surveys of rare population members. Some of the limitations of the method are illustrated by the studies described above. There is the risk that the sampled informant may not accurately report the rare population status of other members of the linkage, either deliberately or through lack of knowledge. Nonresponse for the main survey data collection is another concern. In addition, ethical issues can arise when sampled persons are asked about the rare population membership of those in their linkage when that membership is a sensitive matter. The benefits of network sampling are partially offset by the increased sampling errors arising from the variable weights that the method entails, and by the costs of locating the linked rare population members.

3.6 Location sampling

Location sampling is widely used to sample populations that have no fixed abode for both censuses and surveys: nomads may, for example, be sampled at waterpoints when

they take their animals for water, and homeless persons may be sampled at soup kitchens when they go for food (*e.g.*, Kalton 1993a; Ardilly and Le Blanc 2001). A central feature of such uses of location sampling is that there is a time period involved, resulting in issues of multiplicity (Kalsbeek 2003). A serious concern with the use of the technique is that it fails to cover those who do not visit any of the specified locations in the particular time period.

Location sampling is used to sample rare mobile populations such as passengers at airports and visitors to a museum or national park. In such cases, the question arises as to whether the unit of analysis should be the visit or the visitor. When the visit is the appropriate unit, no issues of multiplicity arise (see, for example, the report on the U.S. National Hospital Discharge Survey by DeFrances, Lucas, Buie and Golosinskiy 2008). However, when the visitor is the unit of analysis, the fact that visitors may make multiple visits during the given time period must be taken into account (Kalton 1991; Sudman and Kalton 1986). One approach is to treat visits as eligible only if they are the first visits made during the time period for the survey. Another approach is to make multiplicity adjustments to the weights in the analysis; however, determining the number of visits made is problematic because some visits will occur after the sampled visit.

Location sampling has also been used for sampling a variety of rare – often very rare – populations that tend to congregate in certain places. For example, Kanouse, Berry and Duan (1999) employed the technique to sample street prostitutes in Los Angeles County by sampling locations where street prostitution was known to occur, and by sampling time periods (days and shifts within days). Location (center) sampling has also been used to sample legal and illegal immigrants in Italy (Meccati 2004). For a 2002 survey of the immigrant population of Milan, 13 types of centers were identified, ranging from centers that provide partial lists from administrative sources (*e.g.*, legal and work centers, language courses), centers that have counts of those attending (*e.g.*, welfare service centers, cultural associations), to centers with no frame information (*e.g.*, malls, ethnic shops).

Location sampling has often been used to sample men who have sex with men, with the locations being venues that such men frequent, such as gay bars, bathhouses and bookstores (Kalton 1993b, MacKellar, Valleroy, Karon, Lemp and Janssen 1996). Based on a cross-sectional telephone survey, Xia, Tholandi, Osmond, Pollack, Zhou, Ruiz and Catania (2006) found that men who visited gay venues more frequently had higher rates of high-risk sexual behaviors and also that the rates of high-risk behaviors varied by venue. These findings draw attention to the difficulty of generating a representative sample by location sampling.

McKenzie and Mistiaen (2009) carried out an experiment to compare location (intercept) sampling with both area sampling and snowball techniques, for sampling Brazilians of Japanese descent (Nikkei) in Sao Paulo and Parana. The locations included places where the Nikkei often went (*e.g.*, a sports club, a metro station, grocery stores and a Japanese cultural club) and events (*e.g.*, a Japanese film and a Japanese food festival). Based on this experiment, they conclude that location sampling (and snowball sampling) oversampled persons more closely connected with the Nikkei community and thus did not produce representative samples. This not-unexpected finding highlights the concern about the use of location sampling for sampling rare populations in general, although not for sampling visits to specified sites.

3.7 Accumulating or retaining samples over time

When survey data collection is repeated over time, survey designers can take advantage of that feature in sampling rare populations (Kish 1999). An important distinction to be made is that between repeated and panel surveys. Samples of rare population members can readily be accumulated over time in repeated surveys. For example, the U.S. National Health Interview Survey is conducted on a weekly basis with nationally representative samples; samples of rare populations can be accumulated over one or more years until a sufficient sample size is achieved (U.S. National Center for Health Statistics 2009a). With accumulation over time, the estimates produced are period, rather than point-in-time, estimates that can be difficult to interpret when the characteristics of analytic interest vary markedly over time (Citro and Kalton 2007). For example, how is a 3-year period poverty rate for a rare minority population to be interpreted when the poverty rate has varied a great deal over the period?

In considering the sampling of rare populations in panel surveys, it is important to distinguish between rare populations that are defined by static versus non-static characteristics. No accumulation over time can be achieved in panel surveys for rare populations defined by static characteristics such a race/ethnicity. However, if a sample of a static rare population is taken at one point in time, it can be useful to follow that sample in a panel to study that population's characteristics at later time points, possibly with supplementary samples added to represent those who entered that population after the original sample was selected. Fecso, Baskin, Chu, Gray, Kalton and Phelps (2007) describe how this approach has been applied in sampling U.S. scientists and engineers over a decade. For the decade of the 1990s, the National Survey of College Graduates (NSCG) was conducted in 1993 with a stratified sample of college graduates selected from the 1990 Census of Population long-form sample records. Those found to be

scientists or engineers were then resurveyed in the NSCG in 1995, 1997 and 1999. To represent new entrants to the target population, another survey – the Survey of Recent College Graduates – was conducted in the same years as the NSCG. A subsample of the recent college graduates was added in to the next round of the NSCG panel on each occasion.

Panel surveys can be used to accumulate samples of non-static rare populations, especially persons experiencing an event such as a birth or a divorce. The U.S. National Children's Study, for instance, plans to follow a large sample of eligible women of child-bearing age over a period of about four years, enrolling those who become pregnant in the main study, a longitudinal study that will follow the children through to age 21 (National Children's Study 2007, Michael and O'Muircheartaigh 2008).

Finally, a large sample can be recruited into a panel and provide data that will identify members of a variety of rare populations that may be of future interest. They are then followed in the panel and, based on their rare population memberships, included in the samples for the surveys for which they qualify. Körner and Nimmergut (2004) describe a German "access panel" that could be used in this way, and there are now several probability-based Web panels that can serve this purpose (Callegaro and DiSogra 2008). However, a serious concern with such panels is the low response rates that are generally achieved.

4. Concluding remarks

This paper has presented a brief overview of the range of methods used in sample surveys for sampling and oversampling rare populations, primarily those classified by Kish as minor domains (the references cited provide more details). Although the methods have been discussed separately, in practice they are often combined, particularly when there are several rare domains of interest. As an example, the California Health Interview Survey, conducted by telephone, has used a combination of disproportionate stratification (oversampling telephone exchanges where the prevalence of the Korean and Vietnamese populations of interest is higher) and a dual-frame design (RDD methods supplemented with a frame of likely Korean and Vietnamese names). In many cases, the art of constructing an effective probability sample design for a rare population is to apply some combination of methods in a creative fashion.

As another example, the Pew Research Center telephone survey of Muslim Americans employed three sampling methods to sample this very rare population (Pew Research Center 2007). One component of the design was a geographically stratified RDD sample, with disproportionate stratified sampling from strata defined in terms of the prevalence of Muslim Americans. The stratum with the lowest prevalence was treated as a cut-off stratum and excluded. The second component was a recontact sample of Muslim Americans drawn from Pew's interview database of recent surveys. The third component was an RDD sample selected from a list of likely Muslim Americans provided by a commercial vendor. To avoid duplicate routes of selection between the geographical strata and the commercial vendor list, telephone numbers selected from the geographical strata were matched against the commercial vendor list and dropped from the geographical strata sample if a match was found.

Not only are the various sampling techniques often used in combination in sample designs for rare populations, but several of the techniques are interrelated. For example, multiple frames can be treated by unique identification (see Section 3.4), which in effect is simply disproportionate stratification. Whereas the whole population is classified into strata for disproportionate stratification, the same approach is adopted with two-phase sampling, but the classification into strata is applied only to members of the first-phase sample. The theory of network sampling is similar to that of multiple-frame sampling, when the latter technique uses inverse overall selection probabilities as weights in the analysis. These interrelationships help to explain the similarities in the theoretical underpinnings of the techniques.

Acknowledgements

I would like to thank Daniel Levine and Leyla Mohadjer for helpful reviews of a draft of this paper, Daifeng Han and Amy Lin for constructive comments on an earlier, shorter version of the paper, and to Mike Brick, Marc Elliott, and Jon Rao for advice on some specific points.

References

Anderson, D.W., and Kalton, G. (1990). Case-finding strategies for studying rare chronic diseases. *Statistica Applicata*, 2, 309-321.

Ardilly, P., and Le Blanc, D. (2001). Sampling and weighting a survey of homeless persons: A French example. *Survey Methodology*, 27, 109-118.

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *American Statistician*, 42, 174-177.

Bolling, K., Grant, C. and Sinclair, P. (2008). *2006-07 British Crime Survey* (*England and Wales*). Technical Report. Volume I. Available at http://www.homeoffice.gov.uk/rds/pdfs07/bcs0607tech1.pdf.

Brick, J.M. (1990). Multiplicity sampling in an RDD telephone survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 296-301.

Callegaro, M., and DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, 72, 1008-1032.

Camburn, D.P., and Wright, R.A. (1996). *Predicting eligibility rates for rare populations in RDD screening surveys*. Available at http://www.cdc.gov/nis/pdfs/sample_design/camburn1996.pdf.

Chu, A., Brick, J.M. and Kalton, G. (1999). Weights for combining surveys across time or space. *Bulletin of the International Statistical Institute*, *Contributed Papers*, 2, 103-104.

Citro, C.F., and Kalton, G. (Eds.) (2007). *Using the American Community Survey*: *Benefits and Challenges*. Washington, DC: National Academies Press.

Clark, R.G., and Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society*, 170, Series A, 63-82.

DeFrances, C.J., Lucas, C.A., Buie, V.C. and Golosinskiy, A. (2008). *2006 National Hospital Discharge Survey*. National Health Statistics Reports Number 5. U.S. National Center for Health Statistics, Hyattsville, MD.

Deming, W.E. (1977). An essay on screening, or on two-phase sampling, applied to surveys of a community. *International Statistical Review*, 45, 29-37.

Durr, J.-M. (2005). The French new rolling census. *Statistical Journal of the United Nations Economic Commission for Europe*, 22, 3-12.

Elliott, M.N., Finch, B.K., Klein, D., Ma, S., Do, D.P., Beckett, M.K., Orr, N. and Lurie, N. (2008). Sample designs for measuring the health of small racial/ethnic subgroups. *Statistics in Medicine*, 27, 4016-4029.

Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., Pantoja, P. and Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services Outcomes Research Methods*, 9, 69-83.

Elliott, M.N., McCaffrey, D., Perlman, J., Marshall, G.N. and Hambarsoomians, K. (2009). Use of expert ratings as sampling strata for a more cost-effective probability sample of a rare population. *Public Opinion Quarterly*, 73, 56-73.

Erens, B., Prior, G., Korovessis, C., Calderwood, L., Brookes, M. and Primatesta, P. (2001). Survey methodology and response. In *Health Survey for England – The Health of Minority Ethnic Groups '99. Volume 2: Methodology and Documentation*. (Eds., B. Erens, P. Primatesta and G. Prior). The Stationery Office, London.

Fecso, R.S., Baskin, R., Chu, A., Gray, C., Kalton, G. and Phelps, R. (2007). *Design Options for SESTAT for the Current Decade*. Working Paper SRS 07-021. Division of Science Resource Statistics, U.S. National Science Foundation.

Fiscella, K., and Fremont, A.M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41, 1482-1500.

Flores Cervantes, I., and Kalton, G. (2008). Methods for sampling rare populations in telephone surveys. In *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P. Lavrakas, M.W. Link and R.L. Sangster). Hoboken, NJ: Wiley, 113-132.

Folsom, R.E., Potter, F.J. and Williams, S.K. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 792-796.

German Federal Statistical Office (2009). *Microcensus*. Available at http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/EN/press/abisz/Mikrozensus__e,templateId=renderPrint.psml.

Goldman, J.D., Borrud, L.G. and Berlin, M. (1997). An overview of the USDA's 1994-96 Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 796-801.

Gonzalez, J.F., Ezzati, T.M., White, A.A., Massey, J.T., Lago, J. and Waksberg, J. (1985). Sample design and estimation procedures. In *Plan and Operation of the Hispanic Health and Nutrition Examination Survey*, *1982-84*. (Ed., K.R. Maurer). Vital and Health Statistics, Series 1, No. 19. U.S. Government Printing Office, Washington, DC, 23-32.

Green, J. (2000). Mathematical programming for sample design and allocation problems. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 688-692.

Haerer, A.F., Anderson, D.W. and Schoenberg, B.S. (1986). Prevalence and clinical features of epilepsy in a biracial United States population. *Epilepsia*, 27, 66-75.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyã*, 36, 99-118.

Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.

Heckathorn, D.D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37, 151-208.

Hedges, B.M. (1973). *Sampling Minority Groups*. Thomson Medal Awards. Thomson Organization, London.

Hedges, B.M., (1979). Sampling minority populations. In *Social and Educational Research in Action* (Ed., M.J. Wilson) London: Longman, 245-261.

Horrigan, M., Moore, W., Pedlow, S. and Wolter, K. (1999). Undercoverage in a large national screening survey for youths. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 570-575.

Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.

Kalton, G. (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17, 183-194.

Kalton, G. (1993a). *Sampling Rare and Elusive Populations*. Department for Economic and Social Information and Policy Analysis, United Nations, New York.

Kalton, G. (1993b). Sampling considerations in research on HIV risk and illness. In *Methodological Issues in AIDS Behavioral Research*. (Eds., D.G. Ostrow, R.C. Kessler). New York: Plenum Press.

Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition*, 6, 491-501.

Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, *A*, 149, 65-82.

Kalton, G., and Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.

Kanouse, D.E., Berry, S.H. and Duan, N. (1999). Drawing a probability sample of female street prostitutes in Los Angeles County. *Journal of Sex Research*, 36, 45-51.

Katzoff, M.J. (2004). Applications of adaptive sampling procedures to problems in public health. In *Proceedings of Statistics Canada Symposium 2004, Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at http://www.statcan.gc.ca/pub/11-522-x/2004001/8751-eng.pdf

Katzoff, M.J., Sirken, M.G. and Thompson, S.K. (2002). Proposals for adaptive and link-tracing sampling designs in health surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1772-1775.

Kish, L. (1965a). Selection techniques for rare traits. In *Genetics and the Epidemiology of Chronic Diseases*. Public Health Service Publication No. 1163.

Kish, L. (1965b). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1976). Optima and proxima in linear sample design. *Journal of the Royal Statistical Society*, *A*, 139, 80-95.

Kish, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.

Kish, L. (1988). Multipurpose sample design. *Survey Methodology*, 14, 19-32.

Kish, L. (1999). Cumulating/combining population surveys. *Survey Methodology*, 25, 129-138.

Körner, T., and Nimmergut, A. (2004). A permanent sample as a sampling frame for difficult-to-reach populations? In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at http://www.statcan.gc.ca/pub/11-522-x/2004001/8752-eng.pdf.

Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fultz, N.H., Hurd, M.D., Potter, G.G., Rodgers, W.L., Steffens, D.C., Weir, D.R. and Willis, R.J. (2005). The Aging, Demographics, and Memory Study: Study design and methods. *Neuroepidemiology*, 25, 181-191.

Lauderdale, D.S., and Kestenbaum, B. (2000). Asian American ethnic identification by surname. *Population Research and Policy Review*, 19, 283-300.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lohr, S.L. (2009). Multiple-frame surveys. In *Handbook of Statistics. Volume 29A: Sample Surveys*: *Design, Methods, and Applications*. (Eds., D. Pfeffermann and C.R. Rao). Burlington, MA: Elsevier B.V.

Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.

MacKellar, D., Valleroy, L., Karon, J., Lemp, G. and Janssen, R. (1996). The Young Men's Survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111, Supplement 1, 138-144.

Maffeo, C., Frey, W. and Kalton, G. (2000). Survey design and data collection in the Disability Evaluation Study. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 79-88.

Marker, D.A. (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183-188.

McKenzie, D.J., and Mistiaen, J. (2009). Surveying migrant households: A comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society*, 172, 339-360.

Meccati, F. (2004). Center sampling: A strategy for sampling difficult-to-sample populations. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to Reach Populations*. Available at http://www.statcan.gc.ca/pub/11-522-x/2004001/8740-eng.pdf.

Metcalf, P., and Scott, A. (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28, 1512-1523.

Michael, R.T., and O'Muircheartaigh, C.A. (2008). Design priorities and disciplinary perspectives: The case of the US National Children's Study. *Journal of the Royal Statistical Society*, *A*, 171, 465-480.

Mohadjer, L., and Curtin, L.R. (2008). Balancing sample design goals for the National Health and Nutrition Examination Survey. *Survey Methodology*, 34, 119-126.

Morris, P. (1965). *Prisoners and Their Families*. Allen and Unwin, London, 303-306.

National Children's Study (2007). Study design. In *The National Children's Study Research Plan*. Version 1.3. Available at http://www.nationalchildrensstudy.gov/research/studydesign/researchplan/Pages/Chapter_6_032008.pdf

Pew Research Center (2007). *Muslim Americans, Middle Class and Mostly Mainstream*. Pew Research Center, Washington, DC.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rodriguez Vera, A. (1982). Multipurpose optimal sample allocation using mathematical programming. Biostatistics Doctoral Dissertation. Ann Arbor: University of Michigan.

Rothbart, G.S., Fine, M. and Sudman, S. (1982). On finding and interviewing the needles in the haystack: The use of multiplicity sampling. *Public Opinion Quarterly*, 46, 408-421.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology,* 20, 3-22.

Sirken, M.G. (2004). Network sample surveys of rare and elusive populations: A historical review. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at http://www.statcan.gc.ca/pub/11-522-x/2004001/8614-eng.pdf.

Sirken, M.G. (2005). Network sampling developments in survey research during the past 40+ years. *Survey Research*, 36, 1, 1-5. Available at http://www.srl.uic.edu/Publist/Newsletter/pastissues.htm.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Smith, P.J., Battaglia, M.P., Huggins, V.J., Hoaglin, D.C., Roden, A., Khare, M., Ezzati-Rice, T.M. and Wright, R.A. (2001). Overview of the sampling design and statistical methods used in the National Immunization Survey. *American Journal of Preventive Medicine*, 20(4S), 17-24.

Statistics Canada (2008). *Canadian Community Health Survey* (*CCHS*). Available at http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3226&lang=en&db=imdb&adm=8&dis=2#b3.

Sudman, S. (1972). On sampling of very rare human populations. *Journal of the American Statistical Association*, 67, 335-339.

Sudman, S. (1976). *Applied Sampling*. New York: Academic Press.

Sudman, S., and Freeman, H.E. (1988). The use of network sampling for locating the seriously ill. *Medical Care*, 26, 992-999.

Sudman, S., and Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.

Sudman, S., Sirken, M.G. and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.

Thompson, S.K. (2002). *Sampling*. 2nd Edition. New York: John Wiley & Sons, Inc.

Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.

Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

Tortora, R., Groves, R.M. and Peytcheva, E. (2008). Multiplicity-based sampling for the mobile telephone population: Coverage, nonresponse, and measurement issues. In *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link and R.L. Sangster). Hoboken, NJ: Wiley, 133-148.

U.S. Census Bureau (2009a). *Design and Methodology*, *American Community Survey*. U.S. Government Printing Office, Washington, DC.

U.S. Census Bureau (2009b). *Small Area Income and Poverty Estimates*. Available at http://www.census.gov/did/www/saipe/methods/statecounty/index.html.

U.S. National Center for Health Statistics (2009a). *National Health Interview Survey* (*NHIS*). Available at http://www.cdc.gov/nchs/nhis/methods.htm.

U.S. National Center for Health Statistics (2009b). *The National Immunization Survey (NIS)*. Available at http://www.cdc.gov/nis/about_eng.htm.

U.S. National Center for Health Statistics (2009c). *State and Local Area Integrated Telephone Survey* (*SLAITS*). Available at http://www.cdc.gov/nchs/about/major/slaits/nsch.htm.

Volz, E., and Heckathorn, D.D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24, 79-97.

Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section*, American Statistical Association, 429-434.

Waksberg, J., Brick, J.M., Shapiro, G., Flores Cervantes, I. and Bell, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 713-718.

Waksberg, J., Judkins, D. and Massey, J.T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 61-71.

Watters, J.K., and Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.

Winglee, M., Park, I., Rust, K., Liu, B. and Shapiro, G. (2007). A case study in dual-frame estimation methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3195-3202.

Word, D.L., and Perkins, R.C. (1996). *Building a Spanish Surname List for the 1990's – A New Approach to an Old Problem.* Population Division Technical Working Paper No. 13. U.S. Census Bureau, Washington, DC.

Wu, C., and Rao, J.N.K. (2009). Empirical likelihood methods for inference from multiple frame surveys. *Proceedings of the International Statistical Institute*, Durban, South Africa.

Xia, Q., Tholandi, M., Osmond, D.H., Pollack, L.M., Zhou, W., Ruiz, J.D. and Catania, J.A. (2006). The effect of venue sampling on estimates of HIV prevalence and sexual risk behaviors in men who have sex with men. *Sexually Transmitted Diseases*, 33, 545-550.

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 143-152
Statistics Canada, Catalogue No. 12-001-X

A standardization of randomized response strategies

Andreas Quatember [1]

Abstract

Randomized response strategies, which have originally been developed as statistical methods to reduce nonresponse as well as untruthful answering, can also be applied in the field of statistical disclosure control for public use microdata files. In this paper a standardization of randomized response techniques for the estimation of proportions of identifying or sensitive attributes is presented. The statistical properties of the standardized estimator are derived for general probability sampling. In order to analyse the effect of different choices of the method's implicit "design parameters" on the performance of the estimator we have to include measures of privacy protection in our considerations. These yield variance-optimum design parameters given a certain level of privacy protection. To this end the variables have to be classified into different categories of sensitivity. A real-data example applies the technique in a survey on academic cheating behaviour.

Key Words: Privacy protection; Statistical disclosure control; Nonresponse; Untruthful answering.

1. Introduction

The occurence of nonresponse and the unwillingness to provide the true answers are natural in survey sampling. They may result in an estimator of population parameters, which has a bias of unknown magnitude and a high variance. A responsible user therefore cannot ignore the presence of nonresponse and untruthful answering.

Let U be the universe of N population units and U_A be a subset of N_A elements, that belong to a class A of a categorial variable under study. Moreover let U_A^c be the group of N_A^c elements, that do not belong to this class $(U = U_A \cup U_A^c, U_A \cap U_A^c = \emptyset, N = N_A + N_A^c)$. Let

$$x_i = \begin{cases} 1 & \text{if unit } i \in U_A, \\ 0 & \text{otherwise} \end{cases}$$

$(i = 1, 2, ..., N)$ and the parameter of interest be the relative size π_A of subpopulation U_A:

$$\pi_A = \frac{\sum_U x_i}{N} = \frac{N_A}{N} \qquad (1)$$

$(\sum_U x_i$ is abbreviated notation for $\sum_{i \in U} x_i)$. In a probability sample s (see for instance: Särndal, Swensson and Wretman 1992, page 8f) an estimator of π_A can be calculated from the Horvitz-Thompson estimator of N_A by

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \sum_s \frac{x_i}{\pi_i} \qquad (2)$$

$(\pi_i > 0$ is the probability that unit i will be included in the sample), if the question "Are you a member of group U_A?" (or an equivalent question) is asked directly (dir). This estimator is unbiased, if all x_i's $(i = 1, 2, ..., n)$ are observed truthfully. In the presence of unit or item nonresponse with respect to a variable under study the sample s is divided into a "response set" $r \subset s$ of size n_r and a "missing set" $m \subset s$ of size n_m $(s = r \cup m, r \cap m = \emptyset, n = n_r + n_m)$. For variables of a highly personal, embarrassing matter (like drug addiction, diseases, sexual behaviour, tax evasion, alcoholism, domestic violence or involvement in crimes) r is furthermore divided into a set t of n_t sample units, who answer truthfully, and a set u of size n_u, who answer untruthfully $(r = t \cup u, t \cap u = \emptyset, n_r = n_t + n_u)$. Estimator (2) must then be rewritten as:

$$\hat{\pi}_A^{\text{dir}} = \frac{1}{N} \cdot \left(\sum_t \frac{x_i}{\pi_i} + \sum_u \frac{x_i}{\pi_i} + \sum_m \frac{x_i}{\pi_i} \right). \qquad (3)$$

Evidently the elements of set u cannot be identified and the x_i's of m are not observable. This imposes errors of measurement and nonreponse on the estimation. Therefore everything should be done to keep the untruthful answering rate as well as the nonresponse rate as low as possible.

Survey design features, which clearly affect both the quantity and the quality of the information asked from the respondents (see for instance: Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau 2004, Section 6.7), are strongly related to the sample units' concerns about "data confidentiality" and "perceived protection of privacy". The first term refers to the respondents' desire to keep replies out of hands of uninvolved persons, whereas the second refers to the wish to withhold information from absolutely anybody. Singer, Mathiowetz and Couper (1993) and Singer, van Hoewyk and Neugebauer (2003) report on two successive U.S. population surveys, that the higher these concerns are the lower is the probability of the respondent's participation in the survey (page 470ff and page 375ff).

1. Andreas Quatember is Assistant Professor at the IFAS-Department of Applied Statistics, Johannes Kepler University Linz, Altenberger Str. 69, A-4040 Linz, Austria, Europe. Web address: www.ifas.jku.at. E-mail: andreas.quatember@jku.at.

What can statisticians contribute to this important field of research? For awkward questions the use of *randomized response strategies* at the survey's design stage may reduce the rates of nonresponse and of untruthful answering due to a perceived increase of privacy protection. A common characteristic of these methods is that instead of the direct questioning on the sensitive subject a questioning design is used, which does not enable the data collector to identify the (randomly selected) question on which the respondent has given the answer, although it does still allow to estimate the parameter under study. The idea is to reduce in this way the individuals' fear of an embarrassing "outing" to make sure that the responding person is willing to cooperate. To achieve this goal the respondent clearly has to understand how the questioning design does protect his or her privacy (*cf.* Landsheer, van der Heijden and van Gils 1999, page 6ff).

Pioneering work in this field was published by Warner (1965). In his questioning design each respondent has to answer randomly either with probability p_1 the question "Are you a member of group U_A?" or with probability $p_2 = 1 - p_1$ the alternative "Are you a member of group U_A^c?" $(0 < p_1 < 1)$. Since then various randomized response techniques with differing randomization devices have been proposed (for a review see: Chaudhuri and Mukerjee 1987, Nathan 1988 or Tracy and Mangat 1996). All of these strategies make use of randomly selected questions or answers, though some of them use different random devices depending on the respondent's possession or nonpossession of a certain attribute (see for example: Kuk 1990; Mangat 1994; Kim and Warde 2005).

Warner (1971) was the first to note that these techniques are also applicable as methods of masking confidential micro-data sets to allow their release for public use (*cf.* ibd., page 887). Such microdata sets might contain variables, which allow the direct identification of survey units like the name or an identification number, but also variables, which contain sensitive information on an individual. To protect the survey units against disclosure it might not suffice to delete the variables, which are directly linked to entities, because some of the units might still be identifiable by the rest of their records. Statistical disclosure control is nothing else but a balancing act between the protection of the anonymity of the survey units and the preservation of information contained in the data (*cf.* Skinner, Marsh, Openshaw and Wymer 1994). Methods of data masking can be classified into three categories (*cf.* Domingo-Ferrer and Mateo-Sanz 2002 or Winkler 2004): (1) The *global recoding* of variables into less detailed categories or larger intervals (see for instance: Willenborg and de Waal 1996, page 5f) or the *local recoding* using different grouping schemes at unit level (*cf.* Hua and Pei 2008, page 215f). (2) The *local suppression* of certain variables for survey units with a high risk of re-identification by simply setting their values at "missing" (*cf.* Willenborg and de Waal 1996, page 77). (3) The *substitution* of true values of a variable by other values.

One of the strategies of the third category is the *micro-aggregation* of variables (*cf.* Defays and Anwar 1998). Therein the true variable values are for example sorted by size and then divided into (small) groups. Within each group data aggregates are released instead of the original observations. Another such method is *data-swapping*, where data from units with a high risk of re-identification are interchanged with data from another subset of survey units (*cf.* Dalenius and Reiss 1982). Another technique of substituting identifying or sensitive information is the *addition of noise* to the observed values, meaning that the outcome of a random experiment is added to each datum (*cf.* Dalenius 1977 or Fuller 1993). Finally also the randomized response techniques can be used to mask identifying or sensitive variables. In this case either the survey units already perform the data masking at the survey's design stage or the statistical agency applies the probability mechanism of the technique before the release of the microdata file (*cf.* Rosenberg 1980, Kim 1987, Gouweleeuw, Kooiman, Willenborg and de Wolf 1998, or van den Hout and van der Heijden 2002).

All methods of statistical disclosure control protect the survey units' privacy by a loss of information, which can be seen as the price that has to be paid for it. To be able to appropriately adjust the estimation process the user of the microdata file has to be informed about the details of the masking procedure.

A new standardization of the techniques of randomized response follows in Section 2 of this paper. Furthermore the statistical properties of the standardized estimator are derived for general probability sampling. In Section 3 the essential perspective of privacy protection is described. The question, which of the special cases included in the standardization is most efficient, is answered in the subsequent Section 4. Section 5 contains a real-data example, which demonstrates the application of the recommendations of Section 4 in a survey on academic cheating behaviour.

2. Standardizing randomized response strategies

Let us formulate the following standardization of the randomized response strategies: Each respondent has either to answer randomly with probability

- p_1 the question "Are you a member of group U_A?",
- p_2 the question "Are you a member of group U_A^c?"

or

– p_3 the question "Are you a member of group U_B? "
or is instructed just to say
– "yes" with probability p_4 or
– "no" with probability p_5

($\sum_{i=1}^{5} p_i = 1,\ 0 \le p_i \le 1$ for $i = 1, 2, ..., 5$). The N_B elements of group U_B are characterized by the possession of a completely innocuous attribute B (for instance a season B of birth), that should not be related to the possession or nonpossession of attribute A. This nonsensitive question on membership of group U_B was introduced as an alternative to the question on membership of U_A by Horvitz, Shah and Simmons (1967) to further reduce the respondent's perception of the sensitivity of the procedure. $\pi_B = N_B/N$ (with $0 < \pi_B < 1$) is the relative size of group U_B . π_B and the probabilities $p_1, p_2, ..., p_5$ are the *design parameters* of our standardized randomized response technique.

Let

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ answers "yes",} \\ 0 & \text{otherwise} \end{cases}$$

($i = 1, 2, ..., n$). For an element i the probability of a "yes"-answer with respect to the randomized response questioning design R is for given x:

$$P_R(y_i = 1) = p_1 \cdot x_i + p_2 \cdot (1 - x_i) + p_3 \cdot \pi_B + p_4 = a \cdot x_i + b \qquad (4)$$

with $a \equiv p_1 - p_2$ and $b \equiv p_2 + p_3 \cdot \pi_B + p_4$. Then the term

$$\hat{x}_i = \frac{y_i - b}{a}$$

is unbiased for the true value $x_i (a \neq 0)$. Using these "substitutes" for x_i (and assuming full cooperation of the respondents) the following theorems apply:

Theorem 1: For a probability sampling design with inclusion probabilities π_i the following unbiased estimator of parameter π_A is given:

$$\hat{\pi}_A = \frac{1}{N} \cdot \sum_s \frac{\hat{x}_i}{\pi_i}. \qquad (5)$$

Theorem 2: For a probability sampling design P the variance of the standardized estimator $\hat{\pi}_A$ (5) is given by

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left(V_P\left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b \cdot (1-b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i} \right). \qquad (6)$$

For the proofs of both theorems see the Appendix. The first summand within the outer brackets of (6) refers to the variance of the Horvitz-Thompson estimator for the total $\sum_U x_i$ for a probability sampling design P when the question on membership of U_A is asked directly. The second one can be seen as the price we have to pay in terms of accuracy for the privacy protection offered by the randomized response questioning design. Apparently this variance can be estimated unbiasedly by inserting an unbiased estimator $\hat{V}_P(\sum_s x_i / \pi_i)$ for $V_P(\sum_s x_i / \pi_i)$ and $\sum_s \hat{x}_i / \pi_i^2$ for $\sum_U x_i / \pi_i$.

For simple random sampling without replacement for instance estimator (5) is given by

$$\hat{\pi}_A = \frac{\hat{\pi}_y - b}{a} \qquad (7)$$

with $\hat{\pi}_y = \sum_s y_i / n$, the proportion of "yes"-answers in the sample. In this case the variance (6) of the standardized estimator $\hat{\pi}_A$ is given by

$$V(\hat{\pi}_A) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N - n}{N - 1} + \frac{1}{n} \cdot \left(\frac{b \cdot (1-b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \pi_A \right). \qquad (8)$$

This theoretical variance is unbiasedly estimated by

$$\hat{V}(\hat{\pi}_A) = \frac{\hat{\pi}_A \cdot (1 - \hat{\pi}_A)}{n - 1} \cdot \frac{N - n}{N} + \frac{1}{n} \cdot \left(\frac{b \cdot (1-b)}{a^2} + \frac{1 - 2 \cdot b - a}{a} \cdot \hat{\pi}_A \right). \qquad (9)$$

To be able to calculate $\hat{\pi}_A$ at all, the question on membership of U_A (or U_A^c, but we will ignore this possibility subsequently without loss of generality) must be included in the questioning design with $p_1 > 0$. There is a total of 16 combinations of this question with the four other questions or answers (see: Table 1). These combinations can be described as special cases of our standardized response strategy. For example choosing $p_1 = 1$ leads to the direct questioning on the subject. If we let $0 < p_1 < 1$ and $p_2 = 1 - p_1$ the standardized questioning design turns into Warner's procedure. For $0 < p_1 < 1$ and $p_3 = 1 - p_1$ one gets Horvitz *et al.*'s technique with known π_B (see: Greenberg, Abul-Ela, Simmons and Horvitz 1969). (For other special cases already published as to the best of our knowledge, the reader is referred to the "References"-column of Table 1).

The question, that arises directly from these considerations, is how to choose the design parameters of the standardized response technique to find out the strategies that perform best. We will answer this question in Section 4. But for this purpose we have to include the level of privacy protection, which results from choosing these parameters differently, in our considerations.

Table 1
All special cases of the standardized randomized response strategy

Design	Questions/Answers					References
	U_A	U_{A^c}	U_B	yes	no	
*ST*1	•					Direct questioning
*ST*2	•	•				Warner (1965)[1]
*ST*3	•		•			Greenberg *et al.* (1969)[2]
*ST*4	•			•		
*ST*5	•				•	
*ST*6	•	•	•			
*ST*7	•	•		•		
*ST*8	•	•			•	Quatember (2007)[3]
*ST*9	•		•	•		
*ST*10	•		•		•	Singh, Horn, Singh and Mangat (2003)[4]
*ST*11	•			•	•	Fidler and Kleinknecht (1977)[5]
*ST*12	•	•	•	•		
*ST*13	•	•	•		•	
*ST*14	•	•		•	•	
*ST*15	•		•	•	•	
*ST*16	•	•	•	•	•	

1. A two-stage version was presented by Mangat and Singh (1990)
2. A two-stage version was presented by Mangat (1992)
3. This is a one-stage version of Mangat, Singh and Singh (1993)
4. This is a one-stage version of Singh, Singh, Mangat and Tracy (1994)
5. A two-stage version was presented by Singh, Singh, Mangat and Tracy (1995)

3. Privacy protection

To be able to compare the efficiency of questioning designs with different design parameters it is apparently inevitable to measure the loss of the respondents' privacy induced by these parameters. The following ratios λ_1 and λ_0 of conditional probabilities may be used for this purpose (*cf.* for example the similar "measures of jeopardy" in Leysieffer and Warner 1976, page 650):

$$\lambda_j = \frac{\max[P(y_i = j \mid i \in U_A),\ P(y_i = j \mid i \in U_A^c)]}{\min[P(y_i = j \mid i \in U_A),\ P(y_i = j \mid i \in U_A^c)]} \quad (10)$$

$(1 \le \lambda_j \le \infty;\ j = 1,\ 0)$.

For $j = 1$ (10) refers to the privacy protection with respect to a "yes"-, for $j = 0$ with respect to a "no"-answer. For the standardized questioning design these "λ-measures" of loss of privacy are given by

$$\lambda_1 = \frac{\max[a + b;\ b]}{\min[a + b;\ b]} \quad (11)$$

and

$$\lambda_0 = \frac{\max[1 - (a + b);\ 1 - b]}{\min[1 - (a + b);\ 1 - b]}. \quad (12)$$

$\lambda_1 = \lambda_0 = 1$ indicates a totally protected privacy. This means that the answer of the responding unit contains absolutely no information on the subject under study. This applies for $a = 0$. The more the λ-measures differ from unity, the more information about the characteristic under study is contained in the answer on the record. At the same time the efficiency of the estimation increases (see below), but the individual's protection against the data collector decreases. For the direct questioning design with $p_1 = 1$, where no masking of the variable is done at all, these measures are given by $\lambda_1 = \lambda_0 = \infty$.

Let the values $\lambda_{1,\,\mathrm{opt}}$ and $\lambda_{0,\,\mathrm{opt}}$ be the maximum λ-values of (11) and (12), that the agency considers to allow enough disclosure protection for the records. In the case of the strategy's usage as to avoid nonresponse and untruthful answering in surveys we may also model the respondents' willingness to cooperate as a function of perceived privacy protection. If the privacy of the respondents is sufficiently protected by the randomization device their full cooperation is assumed. Exceeding the limits $\lambda_{1,\,\mathrm{opt}}$ and/or $\lambda_{0,\,\mathrm{opt}}$ would then automatically introduce untruthful answering and nonresponse into the survey and therefore set us back to the starting point of the problem. Fidler and Kleinknecht (1977) showed in their study for design $ST11$ (Table 1) containing nine variables of very different levels of sensitivity, that their choice of the design parameters ($p_1 = 10/16$, $p_4 = p_5 = 3/16$) yielded nearly full and truthful response for each variable including sexual behaviour (ibd., page 1048). Inserting these values in (11) and (12) gives $\lambda_1 = \lambda_0 = 13/3$. This finding corresponds in the main with results that can be derived from the experiment by Soeken and Macready (1982) and with recommendations given by Greenberg *et al.* (1969). Therefore choosing $\lambda_{1,\,\mathrm{opt}}$ and/or $\lambda_{0,\,\mathrm{opt}}$ close to a value of 4 could be a good choice for most variables, when the standardized randomized response method is used to avoid refusals and untruthful answering of respondents in a survey.

Without loss of generality let us assume subsequently, that we will choose the two categories of the variable under study in such way, that the membership of U_A is at least as sensitive as the membership of U_A^c $(1 \le \lambda_{1,\,\mathrm{opt}} \le \lambda_{0,\,\mathrm{opt}} \le \infty)$. From (11) and (12) the terms a and b can be expressed by the λ-values λ_1 and λ_0. Their sum is given by:

$$a + b = \frac{1 - \dfrac{1}{\lambda_0}}{1 - \dfrac{1}{\lambda_1 \cdot \lambda_0}} \quad (13)$$

with

$$b = \frac{\dfrac{1}{\lambda_1} \cdot \left(1 - \dfrac{1}{\lambda_0}\right)}{1 - \dfrac{1}{\lambda_1 \cdot \lambda_0}} \quad (14)$$

and

$$a = \frac{\left(1 - \frac{1}{\lambda_1}\right) \cdot \left(1 - \frac{1}{\lambda_0}\right)}{1 - \frac{1}{\lambda_1 \cdot \lambda_0}}. \qquad (15)$$

We keep the double ratios on the right of (14) and (15) to find easily the limits for $\lambda_1 \to \infty$ and $\lambda_0 \to \infty$ respectively.

This means that for a given sampling design P the extent of the term $(b \cdot (1-b)/a^2) \cdot \sum_U (1/\pi_i) + (1 - 2 \cdot b - a/a) \cdot \sum_U (x_i/\pi_i)$ in the variance expression (6) does not depend on a single value of the design parameters, but on their aggregated effect on the loss of privacy measured by λ_1 and λ_0. Questioning designs with the same λ-values are equally efficient. Designs with larger λ_1 and/or λ_0 are less efficient than designs with lower λ's.

4. Optimum questioning designs

It does depend on the type of re-identification risk or sensitivity of the subject under study which of the special cases of the standardized randomized response strategy of Table 1 can be most efficient for given λ-measures. Strategies $ST5$ and $ST8$ can never perform best, because they do always protect a "no"-answer more than a "yes".

For a nonidentifying (or nonsensitive) variable (like for instance the season of birth), where $\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}} = \infty$ applies, only the direct questioning design ($ST1$ of Table 1) can achieve the variance-optimum performance (see Table 2, which shows these values of the design parameters, which guarantee the best performance of the estimator $\hat{\pi}_A$; to be able to use Table 2 properly the categorial variable under study has to be classified according to the following categories: C_1: The variable is not sensitive at all ($\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}} = \infty$); C_2: Only the membership of group U_A is sensitive, but not of U_A^c ($\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}} = \infty$); C_3: The membership of both groups U_A and U_A^c is sensitive, but not equally ($\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}} < \infty$); C_4: The membership of U_A and of U_A^c is equally sensitive ($\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}} < \infty$), which shows these values of the design parameters, which guarantee the best performance of the estimator $\hat{\pi}_A$). Although the other designs can be used for such variables, they do unnecessarily protect the privacy of the respondents in some way. This has to be paid by a loss of accuracy of the estimation of π_A. But for $p_1 = 1$ ($a = 1$ and $b = 0$) the variance of $\hat{\pi}_A$ (5) turns to the common formula of the direct questioning with the assumption of full response: $V_P(\hat{\pi}_A) = 1/N^2 \cdot V_P(\sum_s x_i/\pi_i)$.

For a variable, of which only the membership of U_A, but not of U_A^c is sensitive (for instance: U_A = set of drug users within the last year; $U_A^c = U - U_A$) there is $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}} = \infty$. Calculating (14) and (15) for $1 < \lambda_1 < \infty$ and $\lambda_0 \to \infty$ gives $a = 1 - b$ and inserting this into (6) leads to the following expression for the variance of the estimator:

$$V_P(\hat{\pi}_A) = \frac{1}{N^2} \cdot \left[V_P\left(\sum_s \frac{x_i}{\pi_i} \right) + \frac{b}{1-b} \cdot \left(\sum_U \frac{1}{\pi_i} - \sum_U \frac{x_i}{\pi_i} \right) \right]. \qquad (16)$$

Looking for those values of the design parameters, for which the standardized randomized response strategy can achieve this variance and for which equations (14) to (15) hold, we do find that in this case there is only one solution! The only questioning design, that is able to perform optimally, is $ST4$. Its variance-optimum design parameters are given by $p_1 = (\lambda_1 - 1)/\lambda_1$ and $p_4 = 1 - p_1$ (see Table 2). This means, that with probability $p_1 = (\lambda_1 - 1)/\lambda_1$ a respondent is asked the question on membership of U_A and with the remaining probability he or she is instructed to say "yes". In this way the data collector is only able to conclude from a "no"-answer directly on the nonsensitive *non*-possession of A but not from a "yes"-answer on the possession of this sensitive or identifying attribute.

Questioning design $ST1$ is not applicable for such subjects, because it does not protect the respondent's privacy in case of a "yes"-answer at all. All the other procedures protect a "no"-answer more than necessary. Therefore they may be used, but they cannot achieve the efficiency of $ST4$.

If the membership of both U_A and U_A^c is sensitive, so that the variable is sensitive as a whole (for instance: U_A = set of married people, who had at least one sexual intercourse with their partners last week; $U_A^c = U - U_A$), $\lambda_{1,\,\mathrm{opt}} \le \lambda_{0,\,\mathrm{opt}} < \infty$ applies. In this case neither the direct questioning on the subject nor design $ST4$ can be used because they are not able to protect both possible answers.

The other designs are applicable for such topics, but Warner's design cannot achieve the efficiency of the others, if $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}}$. The reason is that this design always protects the respondent's privacy with respect to a "yes"-answer equally to a "no"-answer. But if $\lambda_{1,\,\mathrm{opt}} = \lambda_{0,\,\mathrm{opt}}$ despite to the claims of some publications in the past (see for instance: Greenberg *et al.* 1969, page 526f, Mangat and Singh 1990, page 440, Singh *et al.* 2003, page 518f) there is *not one* randomized response technique that can perform *better* than Warner's technique $ST2$ with the optimum design parameters p_1 and p_2 according to Table 2. For $ST7$ this is only valid for $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}}$. Therefore $ST7$ is the perfect supplement of $ST2$, for which the very opposite is true.

Table 2
Optimum design parameters for given λ_1 and λ_0 and different types of sensitivity of the variable under study

Questioning design (Subject category)	Variance-optimum design parameters
$ST1 \ \ (C_1)$	$p_1 = 1$
$ST2 \ \ (C_4)$	$p_1 = \frac{\lambda_1}{\lambda_1+1},\ p_2 = 1 - p_1$
$ST3 \ \ (C_3, C_4)$	$\pi_B = \frac{\lambda_0-1}{\lambda_1+\lambda_0-2},\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3 = 1 - p_1$
$ST4 \ \ (C_2)$	$p_1 = \frac{\lambda_1-1}{\lambda_1},\ p_4 = 1 - p_1$
$ST6 \ \ (C_4)$	$\pi_B = 0.5,\ p_1\text{: } \frac{\lambda_1-1}{\lambda_1+1} < p_1 < \frac{\lambda_1}{\lambda_1+1},\ p_2 = p_1 - \frac{\lambda_1-1}{\lambda_1+1},$ $p_3 = 1 - p_1 - p_2$
$ST6 \ \ (C_3)$	$\pi_B\text{: } \frac{\lambda_0-1}{\lambda_1+\lambda_0-2} < \pi_B < 1,$ $p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} + \frac{(\lambda_1-1)\cdot\pi_B-(\lambda_0-1)\cdot(1-\pi_B)}{(\lambda_1\cdot\lambda_0-1)\cdot(2\pi_B-1)},$ $p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3 = 1 - p_1 - p_2$
$ST7 \ \ (C_3)$	$p_1 = \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = \frac{\lambda_1-1}{\lambda_1\cdot\lambda_0-1},\ p_4 = 1 - p_1 - p_2$
$ST9 \ \ (C_3, C_4)$	$\pi_B\text{: } 0 < \pi_B < \frac{\lambda_0-1}{\lambda_1+\lambda_0-2},\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3 = \frac{\lambda_1-1}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},\ p_4 = 1 - p_1 - p_3$
$ST10 \ \ (C_3, C_4)$	$\pi_B\text{: } \frac{\lambda_0-1}{\lambda_1+\lambda_0-2} < \pi_B < 1,\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3 = \frac{\lambda_0-1}{(\lambda_1\cdot\lambda_0-1)\cdot\pi_B},\ p_5 = 1 - p_1 - p_3$
$ST11 \ \ (C_3, C_4)$	$p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1},\ p_5 = 1 - p_1 - p_4$
$ST12 \ \ (C_3, C_4)$	$p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $\pi_B\text{: } 0 < \pi_B < \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{\lambda_1+\lambda_0-2-2p_2\cdot(\lambda_1\cdot\lambda_0-1)},\ p_3 = \frac{\lambda_1-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},$ $p_4 = 1 - \sum_{i=1}^{3} p_i$
$ST13 \ \ (C_3, C_4)$	$p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $\pi_B\text{: } \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{\lambda_1+\lambda_0-2-2p_2\cdot(\lambda_1\cdot\lambda_0-1)} < \pi_B < 1,\ p_3 = \frac{\lambda_0-1-p_2\cdot(\lambda_1\cdot\lambda_0-1)}{(\lambda_1\cdot\lambda_0-1)\cdot\pi_B},$ $p_5 = 1 - \sum_{i=1}^{3} p_i$
$ST14 \ \ (C_3, C_4)$	$p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},\ p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_2,\ p_5 = 1 - p_1 - p_2 - p_4$
$ST15 \ \ (C_3, C_4)$	$\pi_B\text{: } 0 < \pi_B < 1,\ p_1 = \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},$ $p_3\text{: } 0 < p_3 < \frac{\lambda_1-1}{(\lambda_1\cdot\lambda_0-1)\cdot(1-\pi_B)},\ p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_3\cdot\pi_B,$ $p_5 = 1 - p_1 - p_3 - p_4$
$ST16 \ \ (C_3, C_4)$	$\pi_B\text{: } 0 < \pi_B < 1,\ p_1\text{: } \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1} < p_1 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1},$ $p_2 = p_1 - \frac{(\lambda_1-1)\cdot(\lambda_0-1)}{\lambda_1\cdot\lambda_0-1},\ p_3\text{: } 0 < p_3 < \frac{\lambda_1\cdot\lambda_0-\lambda_0}{\lambda_1\cdot\lambda_0-1} - p_1,$ $p_4 = \frac{\lambda_0-1}{\lambda_1\cdot\lambda_0-1} - p_2 - p_3\cdot\pi_B,\ p_5 = 1 - \sum_{i=1}^{4} p_i$

All others of the designs of Table 1 like $ST11$ or $ST14$ can perform equally efficient for $\lambda_{1,\,\mathrm{opt}} \le \lambda_{0,\,\mathrm{opt}} < \infty$, if the design parameters are chosen according to the restrictions (14) to (15). Among them Greenberg *et al.*'s strategy with known π_B ($ST3$) has on the one hand the advantage over Warner's design to be able to perform optimally also if $\lambda_{1,\,\mathrm{opt}} < \lambda_{0,\,\mathrm{opt}}$. On the other hand, however, it has the disadvantage (like $ST6$), that the size π_B of subpopulation U_B is completely predetermined (or at least bounded by an interval), if we want to achieve the optimum efficiency. This means in practice, that we have to find a subpopulation not related to the possession and nonpossession of attribute A *and* of appropriate relative size to be able to achieve the estimator's optimum accuracy. In principle this also applies to $ST9$, $ST10$, $ST12$ and $ST13$, but looking at the presettings of design parameter π_B, it turns out that $ST9$ and $ST10$ as well as $ST12$ and $ST13$ perfectly complement each other so that in fact any subset U_B of the population can be used. Finally the most complex special cases, $ST15$ and $ST16$, of our standardized randomized response strategy can both be used with any subpopulation $U_B \subset U$ to achieve the best performance.

5. A real-data example

An empirical study was carried out to demonstrate the applicability of the strategy as a questioning design. For this purpose the population of 80 students, who attended the author's course on "Statistics II" at the Johannes Kepler University in Linz (Austria) during the spring term of 2009, volunteered for the survey. The subject under study was academic cheating behaviour. To this end cheating was defined as any behaviour, that was not allowed in the written exams (including just looking at the test scripts of other students or the use of forbidden documents). It is beyond doubt that this subject is sensitive for such a population. Moreover during the survey all of the students were sitting in one lecture room. The parameter of interest was the proportion of the population of students, that fudged on at least one of the exams of the previous semester (including the exam of the author's course on "Statistics I"). Therefore it is beyond reasonable doubt to assume, that direct questioning on the subject would have resulted into a substantial underestimation of this proportion. An empirical study of Scheers and Dayton (1987) for instance showed very small proportions for almost all different cheating behaviours asked, when the subject in question was asked directly. The use of Greenberg's randomized response strategy $ST3$ lead to a significant increase of these proportions (ibd., page 68).

Apparently, for the variable of interest the membership of group U_A, formed by the "cheaters", is sensitive, but not the membership of the complementary set U_A^c. Therefore in accordance with the recommendations of Section 4 we decided to use questioning design $ST4$ for our survey and to compare it with Warner's strategy $ST2$. The λ-values of loss of privacy were fixed at $\lambda_1 = 4$ and $\lambda_0 = \infty$. From Table 2 we calculated $p_1 = 0.75$ and $p_4 = 0.25$ as the variance-optimum design parameters of $ST4$. To achieve these probabilities the students were asked to throw two dice without showing the result to somebody else and answer in a questionnaire the question "Did you cheat at the exams at least one time?" only if the sum of the numbers on the dice was 5 to 10. Otherwise they should just respond "yes".

Previous to the survey some effort was made to explain the consequences of this randomization strategy on the privacy protection. After giving the answer on the first sheet of the questionnaire, only these sheets were collected. 63 out of the 80 persons answered "yes". 20 of 80 students were expected to do so, because they received the "say yes-instruction". Therefore expected 43 of 60 other students should have answered "yes" on the sensitive question. The estimator for π_A is given by

$$\hat{\pi}_A^{ST4} = \frac{\hat{\pi}_y^{ST4} - p_4}{p_1} = \frac{0.7875 - 0.25}{0.75} = 0.71\dot{6}.$$

For this population survey the estimated variance of $\hat{\pi}_A$ is then

$$\hat{V}(\hat{\pi}_A^{ST4}) = \frac{1 - p_1}{n \cdot p_1} \cdot (1 - \hat{\pi}_A^{ST4}) = 1.181 \cdot 10^{-3}.$$

After this questioning design was completed, the students were asked directly on the second sheet of the questionnaire, whether they had truthfully answered the first question or not. Only four students said that this was not the case. This means, that – if that's true – it is likely that 4 more students did actually cheat. The next question to answer was, if they would still cooperate, if p_1 (of $ST4$) would be higher than 0.75. 32 of 80 students agreed to do so, but the others did not. Obviously (at least) four of them did not cooperate when p_1 was 0.75.

Finally, Warner's technique was applied with the same sensitive question as $ST4$ before. To come close to a λ_1-level of 4 – indicating the same loss of privacy as to a "yes"-answer for both questioning designs –, the sum of the numbers of two dice had to be 3 to 9 to apply a design parameter $p_1 = 0.80\dot{5}$. The λ-measures of loss of privacy for this choice are given by $\lambda_1 = \lambda_0 = 4.143$, indicating a slightly higher loss of privacy compared to $ST4$. With a probability of $0.80\dot{5}$ the students had to answer "Are you a member of U_A? " and with the remaining probability the alternative "Are you a member of U_A^c? ".

Now only 38 of 80 persons gave a "yes"-answer. This results in an estimated proportion of "cheaters" of

$$\hat{\pi}_A^{ST2} = \frac{\hat{\pi}_y^{ST2} - p_2}{p_1 - p_2} = \frac{0.475 - 0.19\dot{4}}{0.6\dot{1}} = 0.45\dot{9}0.$$

Additionally to the slight increase of the objective loss of privacy there is another reasonable explanation for this significantly lower result. Although λ_1 did not change that much, some test persons must have been irritated by the raise of p_1 up to $0.80\dot{5}$ after being asked for $ST4$, if they would still cooperate, if p_1 would be higher than 0.75. Not being able to distinguish between the loss of privacy caused by different design parameters in different questioning designs, some of the "cheaters" did not want to answer truthfully again. Just to demonstrate the effect of the different questioning designs on the efficiency of the estimation process we calculate the estimator of the variance of $\hat{\pi}_A^{ST2}$:

$$\hat{V}(\hat{\pi}_A^{ST2}) = \frac{p_1 \cdot (1 - p_1)}{n \cdot (2p_1 - 1)^2} = 5.243 \cdot 10^{-3}.$$

The reason for this considerable increase of the estimated variance is, that Warner's strategy does protect a "no"-answer always in the same way as a "yes". Since in our case a "no"-answer does not have to be protected at all, this unnecessary protection has to be paid in terms of accuracy.

6. Summary

Randomized response strategies have originally been developed to reduce the nonreponse as well as the untruthful answering rate for sensitive subjects in sample surveys, but they can be applied as masking techniques for public use microdata files as well. The standardization of these techniques for the estimation of proportions developed in this paper provides an opportunity to derive a general formula for the variance of the estimator under probability sampling. Different questioning designs, partly published, partly – to the best of our knowledge – unpublished up to now, can be regarded as special cases of the standardized strategy (see Table 1). For the purpose of a comparison of the accuracy of these designs it is essential to include the levels of privacy protection offered by them in our considerations. Doing this by means of the "λ-measures" of loss of privacy explicated in Section 3 a completely new picture has to be painted in comparison to almost all publications in the past as far as the author knows them. It turns out that the identifying or sensitive subjects have to be classified into different categories in order to find the variance-minimum questioning designs for a given privacy protection (see Table 2). The first category consists of subjects, which are not sensitive at all. The second comprises topics, where only the possession but not the nonpossession of a certain attribute is embarrassing to the respondents. The last category is formed by subjects, which are sensitive as a whole.

For subjects out of the first category it is clear enough that no strategy can be more efficient than the direct questioning on the subject ($ST1$ of Table 1).

Concerning topics of the second category there is just one design available, that can achieve the minimum variance of the estimator. This is the questioning design in which each respondent either with probability p_1 has to answer the question on membership of the sensitive group or with probability $1 - p_1$ is instructed to answer "yes" ($ST4$). All the other special cases of the standardized strategy protect the interviewee's privacy not only in case of a "yes"-answer like $ST4$ does, but also in case of a "no"-answer. Therefore their performances cannot reach the minimum achievable level.

For subjects out of the third category it is shown, that contrary to the claim of other publications, there is not one single strategy available that can perform better than Warner's of 1965 as long as the membership of the subgroup under investigation is equally sensitive to the membership of its complement. A lot of other designs are *equally* efficient as Warner's but not a single one is *more* efficient.

For the variables of this category, where the membership of one group is sensitive, but not equally sensitive as the membership of the complementary one, the situation changes dramatically: Compared under the same levels of privacy protection Warner's technique is not able to achieve the best achievable performance of the standardized randomized design anymore, whereas many other strategies can. For some of the designs including the question on membership of a nonsensitive subpopulation not related to the attribute under study, it is required to find an adequate subpopulation of predetermined relative size. Other designs can be used with subpopulations of any size and are therefore more practicable. Therefore a data collector or publisher could select that one of the equally efficient designs, that seems to be more easily applicable than the others.

Acknowledgements

The author is very grateful to the Associate Editor and two referees for their valuable comments and suggestions.

Appendix

Proofs of theorems 1 and 2

Proof of Theorem 1:

$$E(\hat{\pi}_A) = \frac{1}{N} \cdot E_P\left(E_R\left(\sum_s \frac{\hat{x}_i}{\pi_i} \mid s\right)\right)$$
$$= \frac{1}{N} \cdot E_P\left(\sum_s \frac{x_i}{\pi_i}\right) = \frac{1}{N} \cdot \sum_U x_i = \pi_A.$$

The variance of estimator (5) is given by

$$V(\hat{\pi}_A) = V_P(E_R(\hat{\pi}_A \mid s)) + E_P(V_R(\hat{\pi}_A \mid s)).$$

Then

$$V_P(E_R(\hat{\pi}_A \mid s)) = \frac{1}{N^2} \cdot V_P\left(\sum_s \frac{x_i}{\pi_i}\right).$$

Let the sample inclusion indicator

$$I_i = \begin{cases} 1 & \text{if unit } i \in s, \\ 0 & \text{otherwise.} \end{cases}$$

Because the covariance $C_R(\hat{x}_i, \hat{x}_j \mid s) = 0 \ \forall \ i \neq j$, for the second summand of $V(\hat{\pi}_A)$ applies

$$E_P(V_R(\hat{\pi}_A \mid s)) = E_P\left(\frac{1}{N^2} \cdot V_R\left(\sum_U I_i \cdot \frac{\hat{x}_i}{\pi_i} \mid s\right)\right)$$
$$= E_P\left(\frac{1}{N^2} \cdot \sum_U \frac{I_i^2}{\pi_i^2} \cdot V_R(\hat{x}_i)\right)$$
$$= \frac{1}{N^2} \cdot \sum_U \frac{V_R(\hat{x}_i)}{\pi_i}.$$

For $V_R(\hat{x}_i)$ we have

$$V_R(\hat{x}_i) = \frac{1}{a^2} \cdot V_R(y_i)$$

and

$$V_R(y_i) = b + a \cdot x_i - (b + a \cdot x_i)^2$$
$$= (b + a \cdot x_i) \cdot (1 - b - a \cdot x_i)$$
$$= b \cdot (1 - b) + a \cdot (1 - 2 \cdot b - a) \cdot x_i.$$

Then

$$E_P(V_R(\hat{\pi}_A \mid s)) =$$
$$\frac{1}{N^2} \cdot \left(\frac{b \cdot (1 - b)}{a^2} \cdot \sum_U \frac{1}{\pi_i} + \frac{1 - 2 \cdot b - a}{a} \cdot \sum_U \frac{x_i}{\pi_i}\right).$$

This completes the proof of Theorem 2.

References

Chaudhuri, A., and Mukerjee, R. (1987). *Randomized Response*. New York: Marcel Dekker.

Dalenius, T. (1977). Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1, 73-86.

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

Defays, D., and Anwar, M.N. (1998). Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14 (4), 449-461.

Domingo-Ferrer, J., and Mateo-Sanz, J.M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189-201.

Fidler, D.S., and Kleinknecht, R.E. (1977). Randomized response versus direct questioning: Two data collection methods for sensitive information. *Psychological Bulletin*, 84 (5), 1045-1049.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9 (2), 383-406.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14 (4), 463-478.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. Hoboken: John Wiley & Sons, Inc.

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 65-72.

Hua, M., and Pei, J. (2008). A survey of utility-based privacy-preserving data transformation methods. In: *Privacy-preserving Data Mining*: *Models and Algorithms*, (Eds., C.C. Aggarwal and P.S. Yu), New York: Springer, 207-238.

Kim, J. (1987). A further development of the randomized response technique for masking dichotomous variables. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 239-244.

Kim, J.M., and Warde, W.D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, 133, 211-221.

Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77 (2), 436-438.

Landsheer, J.A., van der Heijden, P. and van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, 33, 1-12.

Leysieffer, F.W., and Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.

Mangat, N.S. (1992). Two stage randomized response sampling procedure using unrelated question. *Journal of the Indian Society of Agricultural Statistics*, 44, 82-87.

Mangat, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society*, Series B, 56, 93-95.

Mangat, N.S., and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.

Mangat, N.S., Singh, S. and Singh, R. (1993). On the use of a modified randomization device in randomized response inquiries. *Metron*, 51, 211-216.

Nathan, G. (1988). A bibliography of randomized response: 1965-1987. *Survey Methodology*, 14, 331-346.

Quatember, A. (2007). Comparing the efficiency of randomized response techniques under uniform conditions. *IFAS Research Paper Series*, 23, www.ifas.jku.at/e2550/e2756/index_ger.html.

Rosenberg, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 311-316.

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

Scheers, N.J., and Dayton, C.M. (1987). Improved estimation of academic cheating behaviour using the randomized response technique. *Research in Higher Education*, 26 (1), 61-69.

Singer, E., Mathiowetz, N.A. and Couper, M.P. (1993). The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. Census. *The Public Opinion Quarterly*, 57 (4), 465-482.

Singer, E., van Hoewyk, J. and Neugebauer, R.J. (2003). Attitudes and behavior: The impact of privacy and confidentiality concerns on participation in the 2000 Census. *The Public Opinion Quarterly*, 67 (3), 368-384.

Singh, R., Singh, S., Mangat, N.S. and Tracy, D.S. (1995). An improved two stage randomized response strategy. *Statistical Papers*, 36, 265-271.

Singh, S., Horn, S., Singh, R. and Mangat, N.S. (2003). On the use of modified randomization device for estimating the prevalence of a sensitive attribute. *Statistics in Transition*, 6 (4), 515-522.

Singh, S., Singh, R., Mangat, N.S. and Tracy, D.S. (1994). An alternative device for randomized responses. *Statistica*, 54, 233-243.

Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10 (1), 31-51.

Soeken, K.L., and Macready, G.B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92 (2), 487-489.

Tracy, D.S., and Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade – A follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4 (2/3), 147-158.

van den Hout, A., and van der Heijden, P.G.M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review*, 70 (2), 269-288.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

Willenborg, L., and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. New York: Springer.

Winkler, W.E. (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. *Research Report Series of the Statistical Research Division of the U.S. Bureau of the Census*, #2004-06.

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 153-164
Statistics Canada, Catalogue No. 12-001-X

Treatments for link nonresponse in indirect sampling

Xiaojian Xu and Pierre Lavallée [1]

Abstract

We examine overcoming the overestimation in using generalized weight share method (GWSM) caused by link nonresponse in indirect sampling. A few adjustment methods incorporating link nonresponse in using GWSM have been constructed for situations both with and without the availability of auxiliary variables. A simulation study on a longitudinal survey is presented using some of the adjustment methods we recommend. The simulation results show that these adjusted GWSMs perform well in reducing both estimation bias and variance. The advancement in bias reduction is significant.

Key Words: Weight share method; Nonresponse; Indirect sampling; Longitudinal survey.

1. Introduction

Indirect sampling refers to selecting samples from the population which is not, but it is related to, the target population of interest. Such a sampling scheme is often carried out when we do not have sampling frames for the target population, but have sampling frames for another population which is related to it. We call the latter sampling population. For an example in Lavallée (2007), we consider the situation where the estimate is concerned with young children belonging to families, but we only have a list of parents' names as our sampling frame. Consequently, we must first select a sample of parents before we can select the sample of children. In this typical indirect sampling situation. The sampling population is that of parents while the target population is that of children. We note that the children of a particular family can be selected through either the father or the mother. Figure 1 provides a simple illustration for this indirect sampling scheme (Figure 1.2, Lavallée 2007).

There is a sizeable amount of literature concerning estimation problems that are associated with indirect sampling, a few of which we name here. Initially, estimation methods for production of cross-sectional estimates using longitudinal household survey are discussed in Ernst (1989). This study presents weight share method in the context of longitudinal survey and also shows that this method provides an unbiased estimator for the total for any characteristic in the population of interest. Kalton and Brick (1995) conclude that such a method also provides minimal variance of estimated population total for some simple sampling schemes for the longitudinal household panel survey. Lavallée (1995) extends weight share method in a completely general context of indirect sampling which includes longitudinal survey as its particular example, called generalized weight share method (GWSM). This work justifies that this weighting scheme provides unbiased estimates irrespective of sampling schemes in obtaining a sample in the sampling population. As with any other weighting scheme, in the process of GWSM implementation an adjustment for a variety of nonresponse problems must be made. Lavallée (2001) provides adjusted GWSM incorporating possible total nonresponse problems in indirect sampling. In indirect sampling there is another type of nonresponse called link nonresponse, termed by Lavallée (2001) as "relationship nonresponse," which is associated with a situation where it is impossible to determine, or where one has failed to determine, whether or not a unit in the sampling population is related to a unit in the target population. Lavallée (2001) points out the problem of overestimation in using GWSM when link nonresponse occurs and leaves finding suitable adjustment of GWSM for link nonresponse as a rather open question. This present study focuses on developing treatments of estimation bias caused by such link nonresponse.

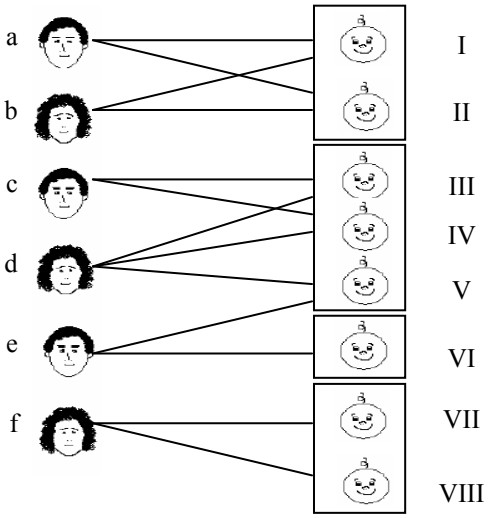

Figure 1 Indirect sampling of children

1. Xiaojian Xu, Department of Mathematics, Brock University, St. Catharines, Ontario, Canada, L2S 3A1. E-mail: xxu@brocku.ca; Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: Pierre.lavallée@statcan.gc.ca.

The rest of this work has been arranged in the following sections. Notation and the problem defined are described in Section 2. We propose a few modification methods in using GWSM incorporating link nonresponse in Section 3. A simulation study using a real life data set is presented in Section 4 with a few closing remarks in Section 5. We note that we show the advances of the new methods provided in this paper through a simulation study while other theoretical contributions relevant to this problem can be found in Lavallée (2002), Deville and Lavallée (2006), and Lavallée (2007).

2. Notation and problem

We use U^A and U^B to denote sampling population and target population respectively. Then, U^A is the population related to U^B with a known sampling frame. We let s^A, M^A, and m^A be a selected sample from U^A, the number of units in U^A, and the number of units in s^A respectively. We use π_j^A to represent the selection probability of j^{th} unit in U^A with $\pi_j^A > 0$ and $\sum_{j=1}^{M^A} \pi_j^A = m^A$. We also make use of the notation: M^B, N, U_i^B, and M_i^B to be the number of units in U^B, the number of clusters in U^B, the i^{th} cluster of U^B with $\cup_{i=1}^{N} U_i^B = U^B$, and the number of units in i^{th} cluster U_i^B.

We define $l_{j,ik}$ as 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. We also define $L_{j,i}^B$ as the total number of links existing between unit j of U^A and units of U_i^B, *i.e.*, $L_{j,i}^B = \sum_{k=1}^{M_i^B} l_{j,ik}$. Let L_i^B be the total number of links existing between units of U^A and units of U_i^B, *i.e.*, $L_i^B = \sum_{j=1}^{M^A} L_{j,i}^B$. We denote the value of the characteristics for the k^{th} unit of i^{th} cluster in population U^B by y_{ik}, and the total of all $y_{ik}'s$ by Y^B. Then, we have $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$.

We let Ω^B denote the clusters in U^B where there is at least one unit ik such that $l_{j,ik} = 1$ for some j^{th} unit in s^A, and we say that it can be identified by units j in s^A, *i.e.*, such i satisfies $L_i^B = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} l_{j,ik} > 0$. The number of clusters in Ω^B is n. After sampling we relabeled the clusters in Ω^B as $i = 1, 2, \ldots, n$. We let w_{ik} refer to the estimation weight assigned to k^{th} unit of i^{th} cluster, Ω_i^A refer to the set of units in U^A that have links to some units in U_i^B with $i \in \Omega^B$, and Ω^A refer to the set of units in U^A that have links to some units in Ω^B, *i.e.*, $\Omega^A = \{j | \sum_{i \in \Omega^B} L_{j,i}^B \neq 0\}$. We use s_i^A to indicate the set of units in s^A that have links to some units in U_i^B with $i \in \Omega^B$. We let T^A, T_i^A, and m_i^A denote the number of units in Ω^A, the number of units in Ω_i^A, and the number of units in s_i^A respectively. Finally, we make use of the following three indicators: let t_j be the indicator variable of being selected in s^A: $t_j = 1$ indicates that j^{th} unit in U^A is in s^A and $t_j = 0$ indicates otherwise; let t_j^L be the indicator variable of being included in s^A for units in Ω^A: $t_j^L = 1$ indicates that j^{th} unit in Ω^A is in s^A and $t_j^L = 0$ indicates otherwise; and let $t_{j,i}^L$ be the indicator variable of being included in s_i^A for units in Ω_i^A: $t_{j,i}^L = 1$ indicates that j^{th} unit in Ω_i^A is in s_i^A and $t_{j,i}^L = 1$ indicates otherwise.

Our goal is to estimate the total Y^B, the parameter of our interest, for target population U^B which is divided into N clusters. In order to do so, we select a sample s^A from U^A with selection probability π_j^A. Then we identify Ω^B using $l_{j,ik} \neq 0$. All units of the clusters in Ω^B are surveyed where y_{ik} and the set of $l_{j,ik}$ are measured.

By applying the GWSM, an estimation weight w_{ik} will be assigned to each unit k of surveyed cluster i's. Such weights can be chosen in an appropriate manner so that the estimator of Y^B:

$$\hat{Y}^B = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \tag{1}$$

performs well in estimating Y^B.

We are interested in estimating the quantity Y^B using $\hat{Y}^B$. According to Horvitz and Thompson (1952), let w_{ik} be inverse of selection probability, π_{ik}, of the k^{th} individual of U_i^B in the target population. Then $\hat{Y}^B$ gives an unbiased estimator for Y^B. However, the computation for π_{ik} is difficult or even impossible in the present case, due to the complication in the indirect sampling scheme. Therefore, GWSM is introduced to address this issue. For readers' convenience, here we outline the GWSM in computing the weights for each cluster that has been observed.

Step 1: Provide the initial weights w_{ik}'

$$w_{ik}' = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}; \tag{2}$$

Step 2: Compute L_i^B

$$L_i^B = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} l_{j,ik}; \tag{3}$$

Step 3: Obtain final weight w_i

$$w_i = \frac{\sum_{k=1}^{M_i^B} w_{ik}'}{L_i^B}; \tag{4}$$

Step 4: Set $w_{ik} = w_i$ for all k in i^{th} cluster.

It follows Theorem in Section 3 of Lavallée (2001) that

$$\hat{Y}^B = \sum_{i=1}^{n} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B} \sum_{k=1}^{M_i^B} y_{ik} \qquad (5)$$

offers an unbiased estimator for Y^B provided all links $l_{j,ik}$ can be correctly identified. The estimation weights assigned in (5) are

$$w_{ik} = \begin{cases} \dfrac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{L_i^B}, & \text{for all units } k \text{ in cluster } i \text{ when } i \text{ in } \Omega^B; \\ 0, & \text{when } i \text{ is not in } \Omega^B. \end{cases} \qquad (6)$$

A simple example is illustrated in Figure 2. We aim to estimate the total Y^B linked to the target population U^B. Suppose that we select the units $j = 1,$ and 2 from U^A. By selecting the unit $j = 1,$ we survey the units of cluster $i = 1$. Likewise, by selecting the unit $j = 2,$ we survey the units of clusters $i = 1,$ and 2. We therefore have $\Omega^B = \{1, 2\}$. For each unit k of clusters i of Ω^B, we calculate the initial weights w'_{ik} in (2), the total number of links existing between units of U^A and units of $U_i^B, L_i^B,$ and the final weights w_{ik}. Then, according to (5) the resulting estimator for Y^B is as below (see Lavallée 2007, pages 17-18 for more details):

$$\hat{Y}^B = \frac{1}{2}\left[\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right] y_{11} + \frac{1}{2}\left[\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right] y_{12} + \frac{1}{3\pi_2^A} y_{21} + \frac{1}{3\pi_2^A} y_{22} + \frac{1}{3\pi_2^A} y_{23}. \qquad (7)$$

We note that for the estimator with known $l_{j,ik}$, the only assumption for unbiasedness is to have $L_i^B > 0$ for all clusters i's in U^B. That is, every cluster of the target population must have at least one link from U^A. We know that if some links were missing, then the estimator (5) would be biased. When link nonresponse occurs, as indicated in Lavallée (2001), L_i^B can not be determined. Traditionally, using total links observed to replace this unknown quantity results in overestimation on Y^B since some link components are actually missing in summation L_i^B. Our proposed study focus is on just such a problem, and we attempt to adjust the estimation weights w_{ik} by estimating L_i^B so as to obtain a better performance of estimation on Y^B.

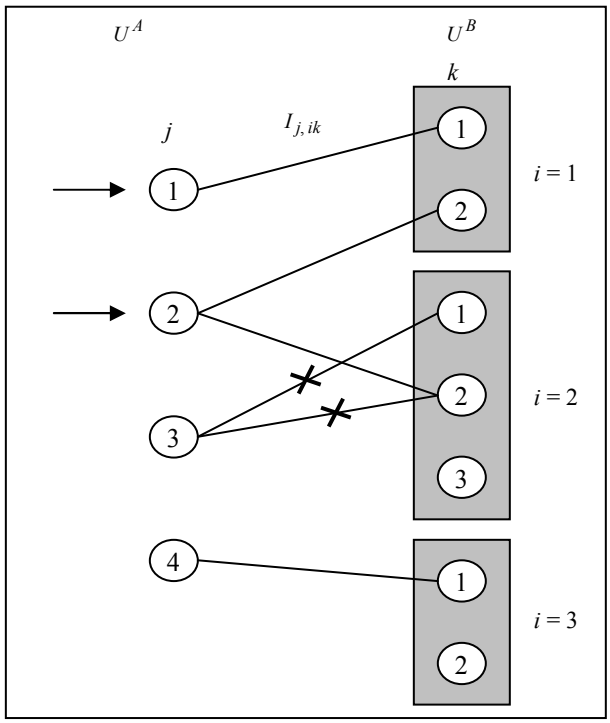

Figure 2 Example of links in indirect sampling

3. Treatments of biased estimation problems

As indicated in Section 1, the biased estimation using GWSM occurs due to link nonresponse problems. In this situation, not all of the composition in L_i^B can be identified or observed. Although the links between units in s^A and units in U^B can normally be determined in practice, the parts of links outside s^A are often difficult or even impossible to identify. We say that such units have missing links with U^B. Let $\Delta^A = \Omega^A \setminus s^A$ be the set of units with possible missing links. Then,

$$L_i^B = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta^A} \sum_{k=1}^{M_i^B} l_{j,ik}. \qquad (8)$$

If we carry out the GWSM without taking these missing links into account, we use the total of observed $l_{j,ik}$ as L_i^{B*} instead to compute $\hat{Y}^B$ using

$$L_i^{B*} = \sum_{j \in s^A} \sum_{k=1}^{M_i^B} l_{j,ik} + \sum_{j \in \Delta_0^A} \sum_{k=1}^{M_i^B} l_{j,ik}, \qquad (9)$$

where Δ_0^A is a subset of Δ^A and only contains the units whose links are observed. The cost is overestimation of Y^B in using (5) since

$$L_i^B \geq L_i^{B*}.$$

We suggest a few methods for applying GWSM under consideration of link nonresponse by estimating L_i^B.

3.1 Estimating L_i^B without availability of auxiliary variables

3.1.1 Estimating L_i^B by proportional adjustment for each individual cluster (Method 1)

To address the link nonresponse problem, we focus on estimating L_i^B using the known information about the links within s^A. To compute the weights in (6) using GWSM, we only need to estimate L_i^B for those $i \in \Omega^B$. For any $i \in \Omega^B$,

$$L_i^B = \sum_{j=1}^{T_i^A} L_{j,i}^B. \tag{10}$$

A general estimator for this total can be expressed as

$$\hat{L}_i^B = \sum_{j=1}^{T_i^A} w_{j,i}^L L_{j,i}^B, \tag{11}$$

where $w_{j,i}^L$ is a random weight that takes the value $w_{j,i}^L = 0$ if j is not in the sample s_i^A. For each $i \in \Omega^B$, we use the known link information between s_i^A and U_i^B to estimate the link information between Ω_i^A and U_i^B. The expectation of $\hat{L}_i^B$ is

$$E(\hat{L}_i^B) = \sum_{j=1}^{T_i^A} E(w_{j,i}^L) L_{j,i}^B. \tag{12}$$

By comparing (10) and (12), it can be observed that $\hat{L}_i^B$ is unbiased for L_i^B for any weighting scheme with $E(w_{j,i}^L) = 1$ for all j.

First of all, we adopt the Horvitz-Thompson estimator (Horvitz & Thompson 1952), also called π estimator (Särndal, Swensson, and Wretman 1991). Note that, by the definition of $\Omega_i^A, \Omega_i^A \supset s_i^A$ for all i. We *imitate* a procedure for estimating the number of links in Ω_i^A using that in s_i^A. The procedure is to select a "sample" s_i^A from the "population" Ω_i^A. Let $\pi_{j,i}^L$ be the probability of j (which is in Ω_i^A) being included in s_i^A. Then, let

$$w_{j,i}^L = \begin{cases} 1/\pi_{j,i}^L, & j \text{ is in } s_i^A, \\ 0, & j \text{ is in } \Omega_i^A \setminus s_i^A. \end{cases} \tag{13}$$

According to Corollary 3.1 in Cassel, Särndal, and Wretman (1977), this weighting scheme provides an unbiased estimator for L_i^B. We have

$$\hat{L}_i^B = \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_j^L}{\pi_{j,i}^L}. \tag{14}$$

It provides us with an asymptotically unbiased (proof follows) estimator of Y^B:

$$\tilde{Y}^B = \sum_{i=1}^{n} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_{j,i}^L}{\pi_{j,i}^L}} \sum_{k=1}^{M_i^B} y_{ik}. \tag{15}$$

In order to show its unbiasedness, we employ Taylor's expansion. According to Corollary 5.1.5 (Fuller 1996), we obtain

$$\frac{1}{\hat{L}_i^B} = \frac{1}{L_i^B} - \frac{1}{(L_i^B)^2}(\hat{L}_i^B - L_i^B) + O([\hat{L}_i^B - L_i^B]^2)$$

$$= \frac{1}{(L_i^B)^2}(2L_i^B - \hat{L}_i^B) + O_p(n^{-1}).$$

It follows that

$$p\lim\left\{ n^{1/2}\left[\frac{1}{\hat{L}_i^B} - \frac{1}{(L_i^B)^2}(2L_i^B - \hat{L}_i^B) \right] \right\} = 0.$$

Therefore, by Theorem 5.2.1 (Fuller 1996), the limiting distribution of $n^{1/2}[1/\hat{L}_i^B]$ is the limiting distribution of $n^{1/2}[1/(L_i^B)^2(2L_i^B - \hat{L}_i^B)]$. We note that $\tilde{Y}^B$ is a function of both random variable: t_j, and random variable: $t_{j,i}^L$; therefore we denote the expectation of $\tilde{Y}^B$ with respect to t_j by $E_{t_j}(\cdot)$ and that with respect to $t_{j,i}^L$ by $E_{t_{j,i}^L}(\cdot)$. Hence, asymptotically we have

$$\begin{aligned} & E(\tilde{Y}^B) \\ & \approx \sum_{i=1}^{n} E_{t_j}\left[E_{t_{j,i}^L}\left(\frac{1}{(L_i^B)^2}\left(2L_i^B - \sum_{j=1}^{T_i^A} \frac{L_{j,i}^B t_{j,i}^L}{\pi_{j,i}^L} \right) \right. \right. \\ & \qquad \left. \left. \sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \middle| \Omega^B \right] \sum_{k=1}^{M_i^B} y_{ik} \\ & = \sum_{i=1}^{n} E_{t_j}\left(\frac{1}{L_i^B} \sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \sum_{k=1}^{M_i^B} y_{ik} \qquad (16) \\ & = E_{t_j}\left(\sum_{i=1}^{n} \left(\frac{1}{L_i^B} \sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A} \right) \sum_{k=1}^{M_i^B} y_{ik} \right) \qquad (17) \\ & = E_{t_j}(\hat{Y}^B). \end{aligned}$$

According to Lavallée (1995), $E_{t_j}(\hat{Y}^B) = Y^B$. Therefore, $\tilde{Y}^B$ is an approximately unbiased estimator of Y^B.

Now we need to compute $\pi^L_{j,i}$. It is a function of π^A_j yet it depends on how s^A_i affects on U^B_i, therefore on Ω^A_i. Such an effect is difficult to track and varies from case to case; however, we can give a general estimate of it. The first approach we propose in this paper is to estimate selection probability, $\pi^L_{j,i}$ using the proportion of the units in s^A which take in Ω^A. Namely

$$\hat{\pi}^{L(1)}_{j,i} = \frac{m^A_i}{T^A_i}. \tag{18}$$

Therefore,

$$\hat{L}^{B(1)}_i = \sum_{j=1}^{T^A_i} \frac{L^B_{j,i}\, t^L_j}{\hat{\pi}^{L(1)}_{j,i}} = \frac{T^A_i}{m^A_i} \sum_{j=1}^{m^A_i} L^B_{j,i}. \tag{19}$$

and

$$\hat{Y}^{B(1)} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{M^A} L^B_{j,i} \frac{t_j}{\pi^A_j}}{\frac{T^A_i}{m^A_i} \sum_{j=1}^{m^A_i} L^B_{j,i}} \sum_{k=1}^{M^B_i} y_{ik} = \sum_{i=1}^{n} w^{(1)}_i \sum_{k=1}^{M^B_i} y_{ik}, \tag{20}$$

with

$$w^{(1)}_i = \frac{m^A_i}{T^A_i} \frac{\sum_{j=1}^{m^A} \frac{L^B_{j,i}}{\pi^A_j}}{\sum_{j=1}^{m^A_i} L^B_{j,i}}. \tag{21}$$

We revisit the example in Figure 2, assuming that there are two link nonresponses that happened between the unit $j=3$ in U^A and the units $k=1,2$ of cluster $i=2$ in U^B. If we use the GWSM without adjustment in (5), the resulting estimator for Y^B is no longer (7). We have instead

$$\hat{Y}^B = \frac{1}{2}\left(\frac{1}{\pi^A_1} + \frac{1}{\pi^A_2}\right) y_{11} + \frac{1}{2}\left(\frac{1}{\pi^A_1} + \frac{1}{\pi^A_2}\right) y_{12} + \frac{1}{\pi^A_2} y_{21} + \frac{1}{\pi^A_2} y_{22} + \frac{1}{\pi^A_2} y_{23}, \tag{22}$$

which is biased. In order to apply (20), we first compute m^A_i / T^A_i. Then the resulting weights using Method (1) in (21) for this example is shown in Table 1. Therefore, this modified method provides the estimator:

$$\hat{Y}^B = \frac{1}{2}\left(\frac{1}{\pi^A_1} + \frac{1}{\pi^A_2}\right) y_{11} + \frac{1}{2}\left(\frac{1}{\pi^A_1} + \frac{1}{\pi^A_2}\right) y_{12} + \frac{1}{2\pi^A_2} y_{21} + \frac{1}{2\pi^A_2} y_{22} + \frac{1}{2\pi^A_2} y_{23}, \tag{23}$$

which is less biased than (22).

Table 1
Initial weights, total number of responded links, and final weights from (21)

i	k	w'_{ik}	L^B_i	m^A_i	T^A_i	m^A_i/T^A_i	$w^{(1)}_i$
1	1	$1/\pi^A_1$	1	2	2	1	$1/2(1/\pi^A_1 + 1/\pi^A_2)$
1	2	$1/\pi^A_2$	1	2	2	1	$1/2(1/\pi^A_1 + 1/\pi^A_2)$
2	1	0	0 (missing)	1	2	1/2	$1/2\pi^A_2$
2	2	$1/\pi^A_2$	1 (one link is missing)	1	2	1/2	$1/2\pi^A_2$
2	3	0	0	1	2	1/2	$1/2\pi^A_2$

3.1.2 Estimating L^B_i by overall proportional adjustment (Method 2)

In the previous approach, the information regarding m^A_i and T^A_i is needed for every i. Suppose we ignore the variation of Ω^A_i among all i, then we simply propose that

$$L^{B*}_i = \sum_{j=1}^{T^A} \frac{L^B_{j,i}\, t^L_j}{\pi^L_j} \tag{24}$$

using link information in s^A to estimate the link information in T^A, where t^L_j being the indicator variable for being in s^A from Ω^A. Now we need to compute π^L_j. Again it is a function of π^A_j and yet it depends on the complexity of effects of s^A on Ω^B, hence to Ω^A. While the computation is difficult and varies from case to case without a general form, we can usually give a rough estimate of it.

The second approach we propose in this paper is to estimate π^L_j using the proportion of the units in s^A which appear in Ω^A, *i.e.*, $\pi^{L*}_j = m^A / T^A$. It informs us that

$$\hat{L}^{B(2)}_i = \frac{T^A}{m^A} \sum_{j=1}^{m^A} L^B_{j,i}. \tag{25}$$

For simple random designs with or without stratification, $\hat{L}^{B(2)}_j$ provides an unbiased estimator for L^B_i. For more complex designs, it provides a model-based unbiased estimator under assumption (A) as follows:

(A) Suppose that for any cluster i, the average of total existing links associated with all units in the sample s^A is the same as that of existing links associated with all units in U^A, *i.e.*,

$$\frac{\sum_{j=1}^{m^A} L^B_{j,i}}{m^A} = \frac{\sum_{j=1}^{M^A} L^B_{j,i}}{T^A}. \tag{26}$$

So, the estimation weights are provided by

$$w_{ik}^{(2)} = w_i^{(2)} = \frac{m^A}{T^A} \frac{\sum_{j=1}^{M^A} L_{j,i}^B \frac{t_j}{\pi_j^A}}{\sum_{j=1}^{M^A} L_{j,i}^B t_j}, \text{ for all units } k \text{ in cluster } i. \tag{27}$$

It follows that Y^B can be estimated by

$$\hat{Y}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^{n} \frac{\sum_{j=1}^{m^A} \frac{L_{j,i}^B}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B} \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{n} w_i^{(2)} \sum_{k=1}^{M_i^B} y_{ik}, \tag{28}$$

We recall the example in Figure 2 with two link nonresponses that happened between the unit $j=3$ in U^A and the units $k=1,2$ of cluster $i=2$ in U^B. In order to apply (28), we first compute m^A / T^A. For this example, we have $m^A = 2,$ and $T^A = 3$. Then the resulting estimator for Y^B using the adjustment Method (2) for this example is

$$\hat{Y}^B = \frac{2}{3}\left[\frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right) y_{11} + \frac{1}{2}\left(\frac{1}{\pi_1^A} + \frac{1}{\pi_2^A}\right) y_{12} + \frac{1}{\pi_2^A} y_{21} + \frac{1}{\pi_2^A} y_{22} + \frac{1}{\pi_2^A} y_{23}\right]. \tag{29}$$

Therefore, this adjustment made in (28) is different from Method (1) for this example.

We know that $\text{var}(\hat{Y}^{B(1 \text{ or } 2)}) = \text{var}\{E(\hat{Y}^{B(1 \text{ or } 2)} \mid s^A)\} + E\{\text{var}(\hat{Y}^{B(1 \text{ or } 2)} \mid s^A)\}$. The inner expectation and variance (conditional on s^A) are taken over all possible sets of "responding" $l_{j,ik}$, given the sample s^A while the outer expectation and variance are taken over all possible sample s^A. Generally, the adjustments made above will not eliminate the second term which depends on the randomness of $l_{j,ik}$.

3.2 Estimating L_i^B with availability of auxiliary variables

3.2.1 Estimating $l_{j,ik}$ using logistic model

The estimation methods for L_i^B proposed in Section 3.1 are simple to apply and do not need additional information. However, sometimes the assumption can be violated which results in an undesirable estimate. For instance, $L_{j,i}^B$ may depend on some characteristics of unit *j* and cluster *i*.

We assume that the probability of a link between a unit in sampling population and a unit in target population depends on some auxiliary variables through a logistic regression model. We may estimate this probability function so that the estimation of the quantity of interest in the target population is desirable. Let $P_{j,ik} = P(l_{j,ik} = 1)$ which is affected by some variable vector $\mathbf{x}_j^A$ in U^A and $\mathbf{x}_{ik}^B$ in U^B.

We may fit the logistic model

$$\log\left(\frac{P_{j,ik}}{1 - P_{j,ik}}\right) = \mathbf{a}'\mathbf{x}_j^A + \mathbf{b}'\mathbf{x}_{ik}^B \tag{30}$$

using the observed links and their corresponding characteristic variables. The unknown parameter vectors $\mathbf{a}$ and $\mathbf{b}$ can be estimated. Then, for those $l'_{j,ik}s$ which can not be identified we suggest to impute them with their probability estimates:

$$\hat{P}_{j,ik} = \frac{e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}{1 + e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}, \tag{31}$$

where $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$ is an estimator for $(\mathbf{a}, \mathbf{b})$, for instance, we use the weighted maximum likelihood (pseudolikelihood) estimator. We then have

$$\hat{L}_i^{B(3)} = \sum_{j \in s^A \cup \Delta_0^A} L_{j,i} + \sum_{j \in \Omega^A \setminus (s^A \cup \Delta_0^A)} \hat{L}_{j,i} = \sum_{j \in s^A \cup \Delta_0^A} L_{j,i} + \sum_{j \in \Omega^A \setminus (s^A \cup \Delta_0^A)} \sum_{k=1}^{M_i^B} \frac{e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}{1 + e^{\hat{\mathbf{a}}'\mathbf{x}_j^A + \hat{\mathbf{b}}'\mathbf{x}_{ik}^B}}. \tag{32}$$

After replacing L_i^B with $\hat{L}_i^{B(3)}$ in (5), (5) provides us with a consistent estimator for Y^B when the model specified in (30) is correct and $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$ is consistent. Note that there are alternatives for the logistic model, such as logit and complementary log-log models. See Draper and Smith (1998) for details. Their research also states that the choice of which model should be employed is not always clear in practice.

3.2.2 Directly estimating L_i^B use log-linear model

We consider that there is a variable vector $\mathbf{x}_i^B$ which affects the value of L_i^B. This indicates that the total number of links in a cluster only varies according to the characteristics of the cluster itself. Using the log-linear model, we can propose (33) below:

$$\log(L_i^B) = \theta^T \mathbf{x}_i^B. \tag{33}$$

If the fit is reasonable, L_i^B can be estimated directly by

$$\hat{L}_i^{B(4)} = e^{\hat{\theta}^T \mathbf{x}_i^B}, \tag{34}$$

where $\hat{\theta}$ is an estimator for θ. When $\hat{\theta}$ is consistent then after replacing L_i^B with $\hat{L}_i^{B(4)}$ in (5), (5) provides a consistent estimator for Y^B. We note that $\hat{L}_i^{B(4)}$ might be non-integer valued, and therefore might have to be rounded to the nearest integer value.

4. Simulation study

When the production of cross-sectional estimates at a particular point in time after the initial point is also of interest in a longitudinal survey design, it becomes a practical example of an indirect sampling problem. Since the population changes over time, the target population is not the same as the initial population which the longitudinal sample is selected from. In this section we will use Survey of Labour and Income Dynamics (SLID) as an example to demonstrate the performance of one of the estimators we introduced in Section 3.1.

The sample design for SLID is detailed in Lavallée (1993). Some terminologies we use in this report - such as cohabitants, initially-present individuals, and initially-absent individuals - follow Lavallée (1995). Initially-absent individuals in the population are individuals who were not part of the population in the year the longitudinal sample was selected, but are considered in the later sample; included among these are newborns and immigrants. After the initial year of selection, the population contains longitudinal individuals, initially-present individuals and initially-absent individuals. Focusing on the households containing at least one longitudinal individual (*i.e.*, longitudinal households), initially-present and initially-absent individuals who join these households are referred to as cohabitants.

In this specific example, U^A is the population at the initial year, say yr_0, of the longitudinal survey, and U^B is the population at any of the following years, say year yr_t, after the initial year. The sample s^A is all the longitudinal individuals. $L_{j,i}$ is a binary variable; it values 1 if individual j lives in i^{th} household at yr_t; 0 otherwise. L_i^B is the total number of longitudinal persons and initially-present cohabitants at yr_0 who lives in i^{th} household at yr_t.

For a longitudinal individual the link would be one to one. For cohabitants there is a significant possibility that this link will be impossible to identify a few years past the initial year, for reasons such as new birth and immigration; further, the greater proportion of cohabitants occupying the target population, the larger this possibility becomes. For instance, in survey panel 3 in SLID, cohabitants represent 7.8 percents out of 47,377 individuals in the year of 2000 which is one year after the initial year. This increases to 13.87 percent in the year 2002 (3 years later), and 15.22 percent in 2003 (4 years later). We can see that the link nonresponses can not be overlooked in such a significant proportion of cohabitants. Due to the availability of observed information, we implement the approach of estimating L_i^B by two kinds of proportional adjustments, which we proposed in Section 3.1.1 and 3.1.2. In order to test the performance of the estimates obtained by these approaches, we carry out a simulation study using SLID data. Cross-sectional estimations for four income variables are of interest for the year of 2003. These four variables are: total income before taxes; total income after taxes; earnings (includes wages and salaries before deductions and self-employment income); and wages and salaries before deductions (also called employment income). We are interested in the total of the population incomes for all these variables. These four quantities of interest have been estimated at both the national level and the provincial level.

For a longitudinal survey, the total number of links in cluster i are generally not more than the total number of individuals in this cluster and not less than the number of longitudinal individuals in this cluster. Since T_i^B is unknown, we replace T_i^B by M_i^B in (5) in our simulation study.

First, we assume that the links between all units selected in the initial year (1999) and all units in the whole population in 2003 are correctly specified. Then we compute the totals using GWSM. We use it as our estimation target, the "truth."

Second, we randomly take away 50 percent of the links associated with initially-present individuals by setting up at random some initially present cohabitants as initially absent ones. The number of links taken makes up approximately 6.3 percent of the total population with which we are interested, with a size of 30,224. Without any adjustment, we recalculate the estimates using GWSM. We use it as our estimation benchmark, the "placebo."

Third, we estimate the same quantities using GWSM with proportional adjustment approaches, Method (1) and (2) in Section 3.1, to see whether the estimates are close enough to the "truth" and how much improvement these adjustments make.

This simulation study using SLID data demonstrates that the proposed method performs very well in overcoming the overestimation problems that arise from link nonresponse.

We denote

$$w_i^{\text{mean}} = \frac{\sum_{j=1}^{m^A} L_{j,i}^B \frac{1}{\pi_j^A}}{\sum_{j=1}^{m^A} L_{j,i}^B} \tag{35}$$

Then, using Method (1) and (2) in Section 3.1 we estimate Y^B by

$$\hat{Y}_{\text{mean}}^{B(1)} = \sum_{i=1}^{n} \frac{m_i^A}{T_i^A} w_i^{\text{mean}} \sum_{k=1}^{M_i^B} y_{ik}, \tag{36}$$

and

$$\hat{Y}_{\text{mean}}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^{n} w_i^{\text{mean}} \sum_{k=1}^{M_i^B} y_{ik}, \tag{37}$$

respectively.

We note that w_i^{mean} is the average weight of longitudinal persons who live in i^{th} household at yr_t. Therefore, it is also reasonable to use median weight:

$$w_i^{\text{median}} = \text{the median of } \frac{1}{\pi_j^A}, j = 1, 2, \ldots, m^A. \tag{38}$$

instead to enhance the robustness of the estimates. Namely, we estimate Y^B as well by

$$\hat{Y}_{\text{median}}^{B(1)} = \sum_{i=1}^{n} \frac{m_i^A}{T_i^A} w_i^{\text{median}} \sum_{k=1}^{M_i^B} y_{ik}, \tag{39}$$

and

$$\hat{Y}_{\text{median}}^{B(2)} = \frac{m^A}{T^A} \sum_{i=1}^{n} w_i^{\text{median}} \sum_{k=1}^{M_i^B} y_{ik}. \tag{40}$$

The comparison for these proposed methods with and without incorporation in nonresponse problems both using mean and median weight within each household are presented in Tables 2-5.

The next four tables give the result for the performance of our estimate using relative error defined as:

$$\left| \frac{\text{estimate - “truth”}}{\text{“truth”}} \right| \times 100\%.$$

Table 2
Total income before taxes (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	9,261,958,108	9,788,749,735	9,317,420,236	9,304,530,248
PEI	2,720,448,008	2,858,506,466	2,735,943,043	2,734,922,451
NS	18,277,017,251	19,573,546,299	18,140,076,618	18,067,144,557
NB	15,297,155,323	16,281,178,934	15,291,696,585	15,236,482,035
QC	1.57839E+11	1.69664E+11	1.56533E+11	1.56405E+11
ON	2.895E+11	3.07642E+11	2.85409E+11	2.85599E+11
MA	23,436,397,548	25,043,168,032	23,632,717,226	23,553,543,216
SK	20,185,285,649	21,595,804,296	20,163,683,598	20,095,359,071
AB	69,063,402,292	74,576,351,600	68,716,661,193	68,582,541,733
BC	81,749,374,346	86,593,614,506	81,387,640,982	81,248,680,715
National	6.8733E+11	7.33617E+11	6.8286E+11	6.82356E+11

Table 3
Total income after taxes (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	7,846,587,557	8,287,351,908	7,892,754,014	7,882,437,105
PEI	2,300,092,795	2,416,503,441	2,314,256,124	2,313,544,320
NS	15,154,508,564	16,257,679,161	15,080,155,194	15,020,088,623
NB	12,878,350,198	13,718,260,686	12,894,700,593	12,849,252,205
QC	1.27632E+11	1.37514E+11	1.27118E+11	1.26999E+11
ON	2.3788E+11	2.53073E+11	2.35192E+11	2.3534E+11
MA	19,541,510,220	20,877,377,918	19,713,628,649	19,649,142,217
SK	16,894,929,025	18,073,635,883	16,890,410,993	16,834,787,407
AB	57,466,974,767	62,055,315,246	57,183,814,491	57,073,904,623
BC	68,710,569,670	72,770,595,462	68,431,531,373	68,309,055,749
National	5.66306E+11	6.05044E+11	5.63958E+11	5.63518E+11

Table 4
Earnings (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	6,433,112,169	6,837,522,157	6,541,306,193	6,530,174,122
PEI	1,898,192,704	2,019,341,995	1,964,066,449	1,962,669,664
NS	12,772,667,160	13,809,197,160	12,999,111,234	12,939,785,579
NB	11,250,688,811	12,030,378,710	11,411,530,716	11,370,222,533
QC	1.18878E+11	1.28949E+11	1.19797E+11	1.19717E+11
ON	2.27577E+11	2.43404E+11	2.26812E+11	2.27092E+11
MA	17,560,695,670	18,995,682,322	18,066,353,153	18,001,882,362
SK	15,159,319,031	16,340,668,148	15,381,733,004	15,319,210,228
AB	56,152,023,359	61,059,244,608	56,540,145,524	56,418,889,147
BC	60,532,655,979	64,499,398,960	61,192,920,832	61,085,986,951
National	5.28214E+11	5.67945E+11	5.3199E+11	5.31722E+11

Table 5
Wages and salaries before deductions (in Canadian dollars)

Province	Estimates by GWSM without missing links	Estimates by GWSM with missing links	Estimates by adjusted GWSM using mean	Estimates by adjusted GWSM using median
NFL	6,180,713,343	6,572,345,010	6,283,079,555	6,272,429,515
PEI	1,636,344,440	1,747,755,878	1,713,809,312	1,713,157,676
NS	12,327,220,137	13,341,912,666	12,579,519,733	12,521,159,025
NB	10,742,381,379	11,508,445,078	10,961,105,589	10,921,102,477
QC	1.08636E+11	1.18092E+11	1.10024E+11	1.09898E+11
ON	2.07331E+11	2.22043E+11	2.07265E+11	2.07495E+11
MA	16,146,993,217	17,504,024,442	16,701,823,718	16,641,840,086
SK	13,982,423,360	15,129,217,320	14,311,467,435	14,255,519,224
AB	52,594,490,290	57,359,188,114	53,195,227,508	53,077,388,907
BC	56,206,787,033	59,886,429,369	56,875,663,895	56,764,297,512
National	4.85784E+11	5.23184E+11	4.91116E+11	4.90763E+11

Table 6
Comparison of relative errors in estimating income before taxes (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	5.688	0.599	0.460	1.059	2.397
PEI	5.075	0.570	0.532	2.859	4.063
NS	7.094	0.749	1.148	3.549	2.459
NB	6.433	0.037	0.397	2.693	2.987
QC	7.492	0.828	0.909	4.372	2.896
ON	6.267	1.413	1.348	4.691	1.771
MA	6.856	0.838	0.500	1.644	3.654
SK	6.988	0.107	0.446	2.480	2.598
AB	7.982	0.502	0.696	3.185	2.407
BC	5.926	0.442	0.612	3.995	3.343
National	6.734	0.650	0.724	3.868	2.662

Table 7
Comparison of relative errors in estimating income after taxes (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	5.617	0.588	0.457	1.101	2.409
PEI	5.061	0.616	0.585	2.832	4.121
NS	7.279	0.491	0.887	3.338	2.765
NB	6.522	0.127	0.226	2.539	3.150
QC	7.742	0.403	0.496	3.991	3.375
ON	6.387	1.130	1.068	4.432	2.081
MA	6.836	0.881	0.551	1.645	3.733
SK	6.977	0.027	0.356	2.406	2.675
AB	7.984	0.493	0.684	3.180	2.415
BC	5.909	0.406	0.584	3.989	3.419
National	6.841	0.415	0.492	3.657	2.927

Table 8
Comparison of relative errors in estimating earnings (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	6.286	1.682	1.509	0.041	3.585
PEI	6.382	3.470	3.397	0.0739	7.115
NS	8.115	1.773	1.308	1.265	5.281
NB	6.930	1.430	1.062	1.279	4.512
QC	8.472	0.773	0.706	2.827	4.560
ON	6.955	0.336	0.213	3.760	2.920
MA	8.172	2.879	2.512	0.291	5.835
SK	7.793	1.467	1.055	0.979	4.324
AB	8.739	0.691	0.475	2.140	3.777
BC	6.553	1.091	0.914	2.643	5.081
National	7.522	0.715	0.664	2.628	4.131

They show that our estimates using both method (1) and method (2) perform very well in terms of reducing bias. Method (1) does work better than Method (2) overall, yet the improvement from Method (1) to Method (2) is much less compared to that made by moving from without adjustment to method (2). Since Method (2) provides us with high quality and involves much less information than Method (1), Method (2) is recommended.

Now, we focus on Method (2) using mean, which gives the estimate $\hat{Y}^{B(2)}_{\text{mean}}$, to analyze how its variance performs in terms of estimating Y^B. We use the bootstrap technique to estimate the variance of $\hat{Y}^{B(2)}_{\text{mean}}$ at both the national level and the provincial level. The bootstrap used for our simulation in this paper is the classical Bootstrap with replacement, where bootstrapping is performed at the first stage of sampling. The bootstrap weights taken here are provided with the SLID data, and incorporate all the necessary adjustments. See Lévesque (2001), and LaRoche (2003) for details on the use of the Bootstrap for SLID. The improvement in reducing the variance is not as large as in reducing bias; however, it is revealed in this simulation study that the proposed method provides a smaller variance as well compared to applying GWSM without an adjustment for missing links. See Table 10 for the results.

The simulation results presented here are based on a single sample of SLID and a single random removal of the links of initially-present individuals. For a complete assessment of the properties of the above estimators, a Monte-Carlo process would have been suitable. Such simulations have been performed by Hurand (2006) based on agricultural data. In these simulations, 1,000 samples have been selected and for each selected sample, the worst-case-scenario has been used, *i.e.*, all links from the non-sample units have been removed. The results of these simulations showed that proportional adjustment and global proportional adjustment are the two methods whose estimates are, on average, the closest to the real total, and whose biases are negligible.

Table 9
Comparison of relative errors in estimating wages and salaries before deductions (%)

Province	GWSM with missing links	Method (1) using mean	Method (1) using median	Method (2) using mean	Method (2) using median
NFL	6.336	1.656	1.484	0.1012	3.593
PEI	6.809	4.734	4.694	1.056	8.424
NS	8.231	2.047	1.573	0.939	5.509
NB	7.131	2.036	1.664	0.685	5.133
QC	8.704	1.278	1.162	2.294	5.070
ON	7.096	0.0317	0.0791	3.473	3.265
MA	8.404	3.436	3.065	0.787	6.469
SK	8.202	2.353	1.953	0.107	5.213
AB	9.059	1.142	0.918	1.713	4.247
BC	6.547	1.190	0.992	2.565	5.234
National	7.699	1.098	1.025	2.251	4.541

Table 10
Comparison of standard deviation estimates

	Variables	Total income before taxes	Total income after taxes	Earnings	Wages and salaries before deductions
National	GWSM with missing links	9,677,258,789	7,343,792,762	8,850,202,075	8,468,718,449
level	Method (2) using mean	9,471,103,083	7,238,715,323	8,593,015,854	8,232,428,642
Ontario	GWSM with missing links	7,888,106,377	6,101,001,739	7,245,688,373	7,149,203,530
	Method (2) using mean	7,601,169,501	5,939,509,894	6,952,217,872	6,831,300,511
Quebec	GWSM with missing links	4,341,215,711	3,113,247,130	3,772,369,180	3,162,277,660
	Method (2) using mean	4,160,251,472	2,974,248,451	3,668,996,929	3,100,868,366

5. Closing remarks

We have constructed four estimation methods to address the link nonresponse problem in indirect sampling. The simulation results in this article show that the adjustments methods we have presented in the example for using GWSM incorporating the link nonresponse performs well in terms of both reducing the estimation bias and providing an overall improvement in variance. The advancement in bias reduction seems significant. The implementation of the methods proposed in Section 3.2 for real data sets will be studied in the near future.

The following significant observations emerged from our study:

1. Adjustment methods are simple to apply.
2. In a more general situation, such as $L_{j,i} > 1$ for some j's, (35) represents the weighted mean weighted by $L^B_{j,i}$. Accordingly the median approach delivered by (39) and (40) can be modified using a generalized version of median – "weighted" median. Namely, we replace (38) by

$$w_i^{\text{median}} = \text{the median of } \frac{1}{\pi_j^A}$$

where $j = 1, 2, \ldots, L^B_{1,i}; 1, 2, \ldots, L^B_{2,i}; \ldots; 1, 2, \ldots, L^B_{m^A,i}$.

3. Some valid link responses outside s^A can not be used in estimating L_i^B by the methods proposed in Section 3.1. However, this valid information would be beneficial to the approaches by predicting $l_{j,ik}$ using auxiliary variables, as can be seen in Section 3.2.1.

Acknowledgements

The authors would like to thank the Associate Editor and the two referees for their helpful suggestions and comments on the previous versions of this paper. This research is funded by the Natural Sciences and Engineering Research Council of Canada, and Mathematics of Information Technology and Complex Systems.

References

Cassel, C.-M., Särndal, C.-E. and Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons, Inc.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: Foundations of the generalised weight share method. *Survey Methodology*, 32, 165-176.

Draper, N.R., and Smith, H. (1998). *Applied Regression Analysis*, 3rd Ed. New York: John Wiley & Sons, Inc.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc., 135-159.

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, 663-685.

Hurand, C. (2006). La méthode généralisée du partage des poids et le problème d'identification des liens. Internal report of the Social Survey Methods Division, Statistics Canada, July 2006.

Kalton, G., and Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.

LaRoche, S. (2003). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics. *Income Research Paper Series*, Catalogue no. 75F0002MIE - No. 007, Statistics Canada.

Lavallée, P. (1993). Sample representativity for the Survey of Labour and Income Dynamics. *Statistics Canada, Research Paper of the Survey of Labour and Income Dynamics*, Catalogue No. 93-19, December 1993.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using weight share method. *Survey Methodology*, 21, 25-32.

Lavallée, P. (2001). Correcting for non-response in indirect sampling. *Proceedings of Statistics Canada's Symposium 2001*.

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles and Éditions Ellipse.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lévesque, I. (2001). Enquête sur la dynamique du travail et du revenu - Estimation de la variance. Internal document from Statistics Canada, July 2, 2001.

Särndal, C.-E., Swensson, B. and Wretman, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 165-176
Statistics Canada, Catalogue No. 12-001-X

Nonparametric propensity weighting for survey nonresponse through local polynomial regression

Damião N. da Silva and Jean D. Opsomer [1]

Abstract

Propensity weighting is a procedure to adjust for unit nonresponse in surveys. A form of implementing this procedure consists of dividing the sampling weights by estimates of the probabilities that the sampled units respond to the survey. Typically, these estimates are obtained by fitting parametric models, such as logistic regression. The resulting adjusted estimators may become biased when the specified parametric models are incorrect. To avoid misspecifying such a model, we consider nonparametric estimation of the response probabilities by local polynomial regression. We study the asymptotic properties of the resulting estimator under quasi-randomization. The practical behavior of the proposed nonresponse adjustment approach is evaluated on NHANES data.

Key Words: Kernel regression; Missing data; Propensity scores; Unit nonresponse; Weighting adjustment.

1. Introduction

Propensity weighting is a procedure that is often applied in sampling surveys to compensate for unit nonresponse. Under this type of nonresponse, complete data collection is accomplished at only a part of the units selected to the sample, which are termed as the respondents. The propensity weighting procedure operates by increasing the sampling weights of the respondents in the sample using estimates of the probabilities that they responded to the survey. These probabilities are also referred to as response propensities in virtue of their analogy with the propensity score theory of Rosenbaum and Rubin (1983) for observational studies, incorporated into survey nonresponse problems by David, Little, Samuhel and Triest (1983).

General descriptions of propensity weighting to adjust classical survey estimators for nonresponse can be seen, for example, in Nargundkar and Joshi (1975), Cassel, Särndal and Wretman (1983) and Groves, Dillman, Eltinge and Little (2002). Traditionally, the way the procedure is implemented estimates the response probabilities with parametric regression curves, such as logistic, probit or exponential models. See Alho (1990), Folsom (1991), Ekholm and Laaksonen (1991) and Iannacchione, Milne and Folsom (1991) for earlier references. A recent theoretical account of the statistical properties of the procedure is given in Kim and Kim (2007). These parametric models are readily fitted as generalized linear models. However, an important and sometimes overlooked part of this procedure is the specification of the form of the link function to relate the response propensities and a linear predictor of the auxiliary information. If this function, which we shall refer to as the response propensity function, is misspecified, the resulting adjusted estimators of the population quantities are likely to be biased.

Another approach to estimate the response propensities is through nonparametric methods. The main motivation to use such methods is that the parametric form for the response propensity function need not be specified. In this sense, these methods offer an appealing alternative to the choice of a link function, as raised by Laaksonen (2006), or when a parametric model is difficult to specify a priori. In this context, Giommi (1984) proposed using kernel smoothing, in the form of the Nadaraya-Watson estimator, to estimate the response probabilities. Da Silva and Opsomer (2006) established the consistency of Giommi's estimator for the population mean and derived rates for the asymptotic bias and the variance. Theoretical properties of a Jackknife variance estimator were also studied.

In this article, we extend the results of Da Silva and Opsomer (2006) in two directions. First, we consider the estimation of the response propensities by local polynomial regression, a nonparametric technique described, for instance, in Wand and Jones (1995). Compared to kernel smoothing, local polynomial regression improves the local approximation to the unknown propensity function, which results in better practical and theoretical properties. It is also much more prevalent as a smoothing method in practice, with implementations available in most major statistical programs. Second, we apply the nonparametric propensity score estimation approach to data from the National Health and Nutrition Examination Survey (NHANES), which makes it possible to compare several nonresponse adjustment methods, both parametric and nonparametric, in a realistic setting.

In Section 2, we introduce the weighting procedure and the estimation of the response propensities. The theoretical properties of the adjusted estimators are discussed in Section 3. In section 4, we describe how to adapt a replication

1. Damião N. da Silva, Departamento de Estatística, Campus Universitário, Natal, RN 59078-970, Brazil. E-mail: damiao@ccet.ufrn.br; Jean D. Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, U.S.A. E-mail: jopsomer@stat.colostate.edu.

variance procedure to estimate the variance of the proposed adjusted estimators. Finally, in Section 5, we demonstrate the finite sample properties of the estimators by means of a simulation experiment using data from NHANES.

2. Weighting by local polynomial regression

Consider a population of N_v units, denoted by $U_v = \{1, 2, ..., N_v\}$. Suppose that a sample s_v is drawn from U_v, according to some probabilistic sampling design $p(s_v)$. Let n_v be the size of s_v and $\pi_i = \pi_{iv} = \Pr\{i \in s_v\} = \sum_{s_v:\, i \in s_v} p(s_v)$ be the inclusion probability of unit i, for all $i \in U_v$. It is of interest to estimate the population mean of a study variable y, namely $\bar{y}_{N_v} = N_v^{-1} \sum_{i \in U_v} y_i$, where y_i denotes the value of y for the i^{th} unit of U_v. We assume that the values x_i of an auxiliary variable x are fully observed throughout the sample. Let $\boldsymbol{y}_v = (y_1, ..., y_{N_v})$, and similarly for $\boldsymbol{x}_v$.

When the sample contains unit nonresponse, we only observe the values of the study variables for the units in a subset $r_v \subset s_v$. To account for the information lost in the estimation of the parameters of interest, it becomes necessary to model the response process. To define this response model, let R_i be an indicator variable assuming the value one if the unit i respond to the survey, and the value zero otherwise, for all $i \in s_v$. We assume that, given the sample, the response indicators are independent Bernouilli random variables with

$$\Pr\{R_i = 1 \mid i \in s_v, \boldsymbol{y}_v, \boldsymbol{x}_v\} = \phi(x_i) \equiv \phi_i, \text{ for all } i \in s_v, \quad (1)$$

where the exact form of the *response propensity function* $\phi(\cdot)$ is unspecified, but it is assumed to be a smooth function of x_i with $\phi(\cdot) \in (0, 1]$. The relationship in (1) defines a nonresponse process said to be ignorable, in the sense that the response propensities are independent of the values of any study variable, conditional on the covariate x (see Lohr 1999, page 265). The theory developed here, therefore, does not intend to handle non-ignorable response mechanisms.

If all response propensities were known, resulting weighting adjustments could be obtained by applying a two-phase estimation approach. For instance, two possible estimators of the population mean $\bar{y}_{N_v}$ would be given by

$$\bar{y}_{\pi\phi v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} y_i R_i \quad (2)$$

and

$$\bar{y}_{\text{rat},\, \pi\phi v} = \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} y_i R_i \Big/ \sum_{i \in s_v} \pi_i^{-1} \phi_i^{-1} R_i, \quad (3)$$

which are forms of adjustments for the Horvitz-Thompson and the Hájek estimators to compensate for the unit nonresponse. The same ideas can be used to obtain propensity weighting adjustments for the generalized regression estimator for estimation in the presence of nonresponse (Cassel *et al*. 1983).

Estimators (2) and (3) are unbiased and nearly unbiased for $\bar{y}_{N_v}$ respectively, under the quasi-randomization approach of Oh and Scheuren (1983), where the statistical properties are evaluated using the joint distribution of the sampling design and the response model. However, the response propensities are usually unknown in practice and we need to replace the ϕ_i in (2) and (3) by estimates $\hat{\phi}_i$, satisfying $0 < \hat{\phi}_i \leq 1$. The resulting propensity weighting estimators are therefore

$$\bar{y}_{\pi\hat{\phi}v} = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} y_i R_i \quad (4)$$

and

$$\bar{y}_{\text{rat},\, \pi\hat{\phi}v} = \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} y_i R_i \Big/ \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} R_i. \quad (5)$$

The latter formula has the advantage of being location-scale invariant, because the summation of its adjusted weights $\pi_i^{-1} \hat{\phi}_i^{-1} R_i / \sum_{i \in s_v} \pi_i^{-1} \hat{\phi}_i^{-1} R_i$ is equal to one, and does not require the population size N_v to be known.

In order to implement the propensity weighting estimators (4) and (5), it is necessary to estimate the response propensities $\hat{\phi}_i$. Da Silva and Opsomer (2006) used kernel regression for this purpose. The procedure we consider here is local polynomial regression, which can be described as follows. Let $K(\cdot)$ be a continuous and positive kernel function and h_v be its bandwidth. Define the $N_v \times (k + 1)$ matrix

$$\mathbf{X}_{Ui} = \begin{bmatrix} 1 & (x_1 - x_i) & \cdots & (x_1 - x_i)^k \\ \vdots & \vdots & & \vdots \\ 1 & (x_{N_v} - x_i) & \cdots & (x_{N_v} - x_i)^k \end{bmatrix},$$

the $N_v \times N_v$ matrix

$$\mathbf{W}_{Ui} = \text{diag}\left\{ \frac{1}{h_v} K\left(\frac{x_j - x_i}{h_v} \right) : 1 \leq j \leq N_v \right\}.$$

and population vector of response indicators $\mathbf{R}_U = (R_1, R_2, ..., R_{N_v})'$. The vector $\mathbf{R}_U$ would be known if, instead of the sample s_v, a census was considered from the population U_v. In that case, the local polynomial regression estimator of degree k of $\phi_i = \phi(x_i)$, based on the whole population, would be given by the fit

$$\hat{\phi}_{Ui} = \boldsymbol{e}_1'(\mathbf{X}_{Ui}'\mathbf{W}_{Ui}\mathbf{X}_{Ui})^{-1}\mathbf{X}_{Ui}'\mathbf{W}_{Ui}\mathbf{R}_U, \quad (6)$$

where e_j denotes the j^{th} column of the identity matrix of order $k+1$ and it is assumed that $\mathbf{X}_{Ui}'\mathbf{W}_{Ui}\mathbf{X}_{Ui}$ is non-singular.

Since the values of the response indicators are only observed for those units selected into the sample, the population fit (6) is unfeasible. However, defining $\mathbf{X}_{si}$ as the $n_\nu \times (k+1)$ matrix formed with the rows of $\mathbf{X}_{Ui}$ corresponding to the units $j \in s_\nu$,

$$\mathbf{W}_{si} = \text{diag}\left\{\frac{1}{\pi_j h_\nu} K\left(\frac{x_j - x_i}{h_\nu}\right):\ j \in s_\nu\right\}$$

and $\mathbf{R}_s = (R_j:\ j \in s_\nu)'$, then a sample-based local polynomial regression estimator of degree k of $\phi_i = \phi(x_i)$ is given by

$$\hat{\phi}_i^o = e_1'\hat{\mathbf{T}}_{si}^{-1}\hat{\mathbf{t}}_{si} \quad (7)$$

where

$$(\hat{\mathbf{T}}_{si},\ \hat{\mathbf{t}}_{si}) \equiv \left(\{\hat{T}_{si,\,pq}\}_{p,\,q=1}^{k+1},\ (\hat{t}_{si,\,p})_{p=1}^{k+1}\right) = (\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{X}_{si},\ \mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{R}_s)$$

and it is assumed that $\hat{\mathbf{T}}_{si}$ is invertible. An special case of (7) is obtained by considering $k = 0$, which corresponds to the kernel regression estimator of Da Silva and Opsomer (2006). Other special cases from (7) are the local linear, the local quadratic and the local cubic response propensity estimators, which result from the local fit of polynomials of degree one, two and three, respectively.

In practice, when $\hat{\mathbf{T}}_{si}$ happens to be singular, a simple procedure to insure that $\hat{\phi}_i^o$ is well defined is choosing a bandwidth large enough to guarantee at least $k+1$ values of R_j in the window $[x_i - h_\nu,\ x_i + h_\nu]$, for all $i \in s_\nu$. If this window does not contain enough responses indicators and the bandwidth has to remain fixed, another approach has to be considered. To this purpose, we adopt here the adjustment made by Breidt and Opsomer (2000) and define the sample-based local polynomial regression estimator of degree k of $\phi_i = \phi(x_i)$ by

$$\hat{\phi}(x_i,\ k,\ h_\nu) = \boldsymbol{e}_1'\left(\hat{\mathbf{T}}_{si} + \text{diag}\left\{\frac{\delta_1}{N_\nu}\right\}\right)^{-1}\hat{\mathbf{t}}_{si},\ i \in s_\nu. \quad (8)$$

where δ_1 is some small positive constant. The smaller order terms δ_1/N_ν added to the main diagonal of $\hat{\mathbf{T}}_{si}$ are sufficient to make the resulting adjusted matrix invertible for any h_ν. As a consequence, $\hat{\phi}(x_i,\ k,\ h_\nu)$ will be well defined, for all $i \in s_\nu$. However, another technical difficulty to use $\hat{\phi}(x_i,\ k,\ h_\nu)$ as a propensity weighting adjustment arises because the response propensity estimator (8) can indeed become arbitrarily close to zero. To tackle this problem, we bound $\hat{\phi}(x_i,\ k,\ h_\nu)$ away from zero by considering the estimator

$$\hat{\phi}_i = \max\{\hat{\phi}(x_i,\ k,\ h_\nu),\ \delta_2(N_\nu h_\nu)^{-1}\}, \quad (9)$$

for some constant $\delta_2 > 0$. This idea is related to the adjustment made by Da Silva and Opsomer (2006) for the kernel regression estimator.

3. Asymptotic properties

In this section, we present the properties of the propensity weighting estimators (4) and (5) under estimation of the response propensities by the local polynomial estimator (9). The assumptions, lemmas and outlines of the proofs for the following results are given in the Appendix, and a complete theoretical investigation can be found in Da Silva and Opsomer (2008). The full derivations are not reported in this article, because they follow the general approach described in Da Silva and Opsomer (2006). We consider an asymptotic framework by which the population U_ν is embedded into the increasing sequence of populations $\{U_\nu:\ N_\nu < N_{\nu+1}\}_{\nu=1}^{\infty}$. From each U_ν, a sample s_ν of size $n_\nu (n_\nu \geq n_{\nu-1})$ is selected according to a sampling design $p_\nu(\cdot)$. This framework is commonly adopted in asymptotic studies of survey estimators. See Isaki and Fuller (1982) for an early reference.

As a population-based approximation for $\phi_i \equiv \phi(x_i)$, we shall consider in the derivation of most results in this section the population fit by local polynomial regression

$$\tilde{\phi}_i \equiv \tilde{\phi}(x_i,\ k,\ h_\nu) = \boldsymbol{e}_1'\mathbf{B}_i \equiv \boldsymbol{e}_1'\mathbf{T}_i^{-1}\mathbf{t}_i,\ i \in U_\nu, \quad (10)$$

where

$$(\mathbf{T}_i,\ \mathbf{t}_i) \equiv (\{T_{i,\,pq}\}_{p,\,q=1}^{k+1},\ (t_{i,\,p})_{p=1}^{k+1}) \equiv \text{E}(\hat{\mathbf{T}}_{si},\ \hat{\mathbf{t}}_{si}) = (\mathbf{X}_{Ui}'\mathbf{W}_{Ui}\mathbf{X}_{Ui},\ \mathbf{X}_{Ui}'\mathbf{W}_{Ui}\boldsymbol{\phi}_U),$$

the matrices $\mathbf{X}_{Ui}$ and $\mathbf{W}_{Ui}$ are as in (6) and $\boldsymbol{\phi}_U = (\phi(x_1),\ \phi(x_2),\ ...,\ \phi(x_{N_\nu}))'$. The following theorem states the asymptotic properties of $\bar{y}_{\pi\hat{\phi}\nu}$ under a set of assumptions in the Appendix. These assumptions are regularity conditions on the sampling design and the finite population, both of which are standard infinite population asymptotics, ignorability conditions on the nonresponse mechanism, and a set of standard regularity conditions related to the local polynomial regression of the response propensity function.

Theorem 1. *Assume the assumptions* (*A*1)-(*A*4), (*B*1)-(*B*3) *and* (*C*1)-(*C*5) *in the Appendix hold. Consider the estimation of the population mean* $\bar{y}_{N_\nu}$ *by the propensity weighting estimator* $\bar{y}_{\pi\hat{\phi}\nu}$ *defined in* (4), *and suppose the response propensities are estimated by* $\hat{\phi}_i$, *the local polynomial regression estimator of degree* k *in* (9). *Let*

$$\overline{y}_{\pi\hat{\psi}\nu} = \frac{1}{N_\nu} \sum_{i \in s_\nu} \pi_i^{-1} \hat{\psi}_i^{-1} y_i R_i, \tag{11}$$

where

$$\hat{\psi}_i^{-1} = \tilde{\phi}_i^{-1} - \tilde{\phi}_i^{-2} \boldsymbol{e}_1' \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si} \mathbf{B}_i),$$

$\hat{\mathbf{t}}_{si}$ *and* $\hat{\mathbf{T}}_{si}$ *are given in* (7) *and* $\tilde{\phi}_i$, $\mathbf{B}_i$, $\mathbf{T}_i$ *are defined in* (10). *Then*,

$$\mathrm{E}[(\overline{y}_{\pi\hat{\phi}\nu} - \overline{y}_{\pi\hat{\psi}\nu})^2] = O\left(\frac{1}{n_\nu^2 h_\nu^2}\right) \tag{12}$$

and the bias and variance of $\overline{y}_{\pi\hat{\psi}\nu}$ *satisfy*

$$\begin{aligned} B_\nu &\equiv \mathrm{E}[\overline{y}_{\pi\hat{\psi}\nu} - \overline{y}_{N_\nu}] \\ &= \begin{cases} O\left(h_\nu^{k+(3/2)}\right) + O\left(\dfrac{1}{n_\nu h_\nu}\right) & k \text{ even}, \\ O\left(h_\nu^{k+1}\right) + O\left(\dfrac{1}{n_\nu h_\nu}\right) & k \text{ odd}, \end{cases} \end{aligned} \tag{13}$$

and

$$\mathrm{Var}[\overline{y}_{\pi\hat{\psi}\nu}] = O\left(\frac{1}{n_\nu h_\nu}\right). \tag{14}$$

Results (12) and (13) imply that the propensity weighting estimator $\overline{y}_{\pi\hat{\phi}\nu}$, using a response propensity estimator based on local polynomial regression, is asymptotically unbiased for the population mean $\overline{y}_{N_\nu}$ under the joint distribution of the sampling design and the response model (1). Combining this result with (14), then we obtain that

$$\hat{y}_{\pi\hat{\phi}\nu} = \overline{y}_{N_\nu} + O_p\left(\frac{1}{\sqrt{n_\nu h_\nu}}\right), \tag{15}$$

when the bandwidth satisfies

$$h_\nu = \begin{cases} O\left(n_\nu^{-\frac{1}{2k+4}}\right), & k \text{ even}, \\ O\left(n_\nu^{-\frac{1}{2k+3}}\right), & k \text{ odd}. \end{cases} \tag{16}$$

Hence, without assuming a parametric form for the response propensity function $\phi(\cdot)$, $\overline{y}_{\pi\hat{\phi}\nu}$ is consistent for the population mean with respect to the sampling design and the response model, as long as the response propensities are a smooth function of the covariate x. As a price paid for this robustness, the rate of convergence is of order $\sqrt{n_\nu h_\nu}$ instead of the usual parametric rate $\sqrt{n_\nu}$. However, as the degree of the local polynomial k increases, the rate of convergence improves. Since the kernel regression estimator in Da Silva and Opsomer (2006) is equivalent to the case $k = 0$, local polynomial regression with higher degree is asymptotically superior to kernel regression in the context of a nonresponse adjustment. This theoretical finding is consistent with that in other contexts (see *e.g.*, Wand and Jones 1995, page 130).

Expression (11) on Theorem 1 generalizes another finding from Da Silva and Opsomer (2006) to the case of local polynomial regression, which is that the asymptotic weights $\hat{\psi}_i^{-1}$ cannot be approximated by the inverse of response propensities ϕ_i^{-1} (or their population-level estimators $\tilde{\phi}_i^{-1}$). One immediate consequence is that the estimator $\overline{y}_{\pi\hat{\phi}\nu}$ is *not* asymptotically equivalent to $\overline{y}_{\pi\hat{\phi}\nu}$ in (2).

The following corollary provides an asymptotic distribution for $\overline{y}_{\pi\hat{\phi}\nu}$, assuming the asymptotic normality of $\overline{y}_{\pi\hat{\psi}\nu}$.

Corollary 1. *Assume the conditions of Theorem* 1 *hold. Suppose that the sampling design and the response model are such that*

$$\frac{\overline{y}_{\pi\hat{\psi}\nu} - \overline{y}_{N_\nu} - B_\nu}{[\mathrm{Var}(\overline{y}_{\pi\hat{\psi}\nu})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } \nu \to \infty,$$

where B_ν *is defined in* (13). *If additionally*

$$\lim_{\nu \to \infty} (n_\nu h_\nu) \mathrm{Var}(\overline{y}_{\pi\hat{\psi}\nu}) \in (0, \infty),$$

then

$$\frac{\overline{y}_{\pi\hat{\phi}\nu} - \overline{y}_{N_\nu} - B_\nu}{[\mathrm{Var}(\overline{y}_{\pi\hat{\psi}\nu})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

We now discuss the properties of the ratio-based version of propensity weighting estimator given in (5). Based on the results for $\overline{y}_{\pi\hat{\phi}\nu}$, standard ratio estimation theory can be used to derive asymptotic results for $\overline{y}_{\mathrm{rat},\,\pi\hat{\phi}\nu}$. In particular, under the same assumptions the asymptotic rates for the approximate bias and variance of $\overline{y}_{\mathrm{rat},\,\pi\hat{\phi}\nu}$ are the same as those in Theorem 1, and the asymptotic distribution of $\overline{y}_{\mathrm{rat},\,\pi\hat{\phi}\nu}$ is given in the following result.

Theorem 2. *Assume the conditions of Theorem* 1 *hold. Suppose the population mean is to be estimated by the propensity weighted estimator* $\overline{y}_{\mathrm{rat},\,\pi\hat{\phi}\nu}$ *of* (5) *and the response propensities are estimated by* $\hat{\phi}_i$, *the local polynomial regression estimator of degree* k *defined in* (8). *Let*

$$\overline{e}_{\pi\hat{\psi}\nu} = \frac{1}{N_\nu} \sum_{i \in s_\nu} \pi_i^{-1} \hat{\psi}_i^{-1} (y_i - \overline{y}_{N_\nu}) R_i,$$

where the weights $\hat{\psi}_i^{-1}$ *are given in Theorem* 1. *Suppose that*

$$\frac{\overline{e}_{\pi\hat{\psi}\nu} - \mathrm{E}(\overline{e}_{\pi\hat{\psi}\nu})}{[\mathrm{Var}(\overline{e}_{\pi\hat{\psi}\nu})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } \nu \to \infty,$$

and

$$\lim_{\nu\to\infty}(n_\nu\, h_\nu)\mathrm{Var}(\overline{e}_{\pi\hat{\psi}\nu}) \in (0,\ \infty).$$

Then,

$$\frac{\overline{y}_{\mathrm{rat},\,\pi\hat{\phi}\nu} - \overline{y}_{N_\nu} - B_{\mathrm{rat},\,\nu}}{[\mathrm{Var}(\overline{e}_{\pi\hat{\psi}\nu})]^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,\ 1)$$

as $\nu \to \infty$, *where* $B_{\mathrm{rat},\,\nu} = O(h_\nu^{k+1})$, *if* k *is odd, and* $B_{\mathrm{rat},\,\nu} = O(h_\nu^{k+(3/2)})$, *if* k *is even.*

4. Variance estimation

As noted in Section 3, the estimator $\overline{y}_{\pi\hat{\phi}\nu}$ is not asymptotically equivalent to $\overline{y}_{\pi\phi\nu}$, so that approximating the asymptotic variance of the former by that of the latter is typically incorrect. In fact, a proof that the asymptotic variance of $\overline{y}_{\pi\phi\nu}$ overestimates the variance of $\overline{y}_{\pi\hat{\phi}\nu}$ is given by Kim and Kim (2007) when the response propensities are assumed to follow a parametric model. In the present context, the asymptotic variance of $\overline{y}_{\pi\hat{\phi}\nu}$ is

$$\mathrm{Var}[\,\overline{y}_{\pi\hat{\psi}\nu}] = \mathrm{Var}\left(\frac{1}{N_\nu}\sum_{i\in s_\nu}\pi_i^{-1}\,\hat{\psi}_i^{-1}R_i\,y_i\right),$$

with $\hat{\psi}_i^{-1}$ given in Theorem 1. As was previously noted in Da Silva and Opsomer (2006) for the simpler case of a zero degree polynomial, the high level of complexity in the expression makes direct estimation of this variance impractical, and a replication method was proposed instead. We briefly outline the procedure here, which is extended to local polynomials of degree k. We omit the theoretical derivations.

We start from a set of replicate weights in the absence of nonresponse, defined for estimating the variance of a linear estimator

$$\hat{\theta} = \frac{1}{N_\nu}\sum_{i\in s_\nu} w_i\, y_i.$$

The replicate variance estimator for $\hat{\theta}$ is defined as

$$\hat{\mathrm{V}}(\hat{\theta}) = \sum_{\ell=1}^{L_\nu} c_\ell\,(\hat{\theta}^{(\ell)} - \hat{\theta})^2, \qquad (17)$$

where

$$\hat{\theta}^{(\ell)} = \frac{1}{N_\nu}\sum_{i\in s_\nu} w_i^{(\ell)}\, y_i,\ \ell = 1,\ 2,\ ...,\ L_\nu,$$

denotes a set of L_ν replicates for $\hat{\theta}$, $w_i^{(\ell)}$ are sampling weights associated with the ℓ^{th} replicate and c_ℓ is factor that depends on the replication procedure. Examples of replication procedures satisfying (17) use variants of the Jackknife method or the Balanced Repeated Replication technique. The process to adapt the replication procedure to estimating the variance of $\overline{y}_{\pi\hat{\phi}\nu}$ and $\overline{y}_{\mathrm{rat},\,\pi\hat{\phi}\nu}$ is straightforward. The needed replicates of these adjusted estimators, namely $\overline{y}^{(\ell)}_{\pi\hat{\phi}\nu}$ and $\overline{y}^{(\ell)}_{\mathrm{rat},\,\pi\hat{\phi}\nu}$, are obtained by replacing the $w_i = \pi_i^{-1}$ by $w_i^{(\ell)}$ in (4) and (5), respectively, and also in the computations needed to produce the $\hat{\phi}_i$ in (9). In section 5.4 below, we evaluate the practical performance of the replication variance procedure on NHANES data.

5. Application to NHANES data

5.1 The NHANES design

We evaluate the performance of the local polynomial adjusted estimators on real data. We consider the 2005-2006 release of the National Health and Nutrition Examination Survey (NHANES), which is conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention (NCHS/CDC), of the U.S. Department of Health and Human Services. This survey consists of a stratified, multistage sample of the U.S. civilian non-institutionalized population. A general overview of the sample formation is as follows:

(i) within each stratum, primary sampling units (PSUs) consisting of counties or grouped smaller counties are selected by sampling with probabilities proportional to a measure of size;
(ii) from the sampled PSUs, groups of city blocks (segments) containing clusters of households are selected also by sampling with probability proportional to size;
(iii) in the selected segments, clusters of households are randomly selected with varying selection probabilities to oversample groups of age, ethnic, or income in certain geographic areas; and
(iv) in the selected households, one or more participants are selected randomly.

The public release of NHANES data has two important aspects. First, to reduce disclosure risks, the stratified, four-stage survey is condensed in a stratified one-stage design, with neither the new stratum variable nor the new PSU variable corresponding to the same variables in the original design. Secondly, the base sampling weights, obtained by reciprocal of the inclusion probabilities of the survey participants, are not released. The weights provided reflect adjustments made to the base weights to account for unit nonresponse, in the interview and exam portions of the survey, and to produce estimates satisfying known population controls.

5.2 The simulation experiment

In order to empirically evaluate the local polynomial estimators as adjustments for nonresponse in complex surveys, we will apply an artificially generated source of unit nonresponse to the public-release NHANES dataset. The nonresponse mechanism will be taken as a smooth function of the age in years of the survey participant (AGE). For this comparison, we chose as study variables four characteristics related to heart diseases, namely the systolic blood pressure (SBP), the diastolic blood pressure (DBP), the indicator of hypertension (HTN) and the indicator of high serum total cholesterol (HTC). All of these were measured on survey participants who were 18 years or older. The systolic and diastolic variables were obtained as the average of the corresponding measurements in a set of up to four readings. Hypertension was defined for individuals having systolic blood pressure of 140 mm Hg or higher or a mean diastolic blood pressure of 90 mm Hg or higher or currently taking medication to lower high blood pressure. High serum total cholesterol was considered when the individual had a total serum cholesterol greater than or equal to 240 mg/dL. The unweighted sample correlations among these and the AGE variable are 0.481 (SBP), 0.118 (DBP), 0.552 (HTN) and 0.060 (HTC), respectively. Hence, it is reasonable to postulate that unit nonresponse related to age is likely to have different effects on survey estimators for these four variables.

The total number of eligible individuals in the NHANES dataset is 4,727. We generated unit nonresponse for the four variables of interest according to two logistic response propensity functions of the auxiliary variable x taken by the age (in years) of the survey participant minus 18. These functions consider a linear and a nonlinear predictor of x as follows

Linear predictor:

$$\phi_I(x) = \{1 + \exp[-(\beta_0 + \beta_1 x)]\}^{-1}$$

Nonlinear predictor:

$$\phi_{II}(x) = \{1+\exp[-(\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 \cos(\beta_4 x^2/\pi)\sin(\beta_5 x/\pi))]\}^{-1},$$

where the regression coefficients $\beta_0, ..., \beta_5$ were chosen so that the response propensity functions give an overall nonresponse rate of about 30% when applied to the sample values of x. In both cases, we kept the NHANES sample fixed and generated $B = 1{,}000$ independent response indicator vectors by Poisson sampling.

The following six nonresponse adjustments were evaluated on these data. Note that in all cases we reported the ratio versions (5) of the estimators, because they were found to be much more precise than the Horvitz-Thompson versions.

1. True response probabilities: $\hat{\phi}_i = \phi(x_i)$, $i \in s_v$.
2. Logistic regression adjustment: $\hat{\phi}_i$ obtained as the estimated probabilities from a logistic regression of each response vector on x, using a polynomial in x of degree one as the linear predictor.
3. Weighted local polynomial regression of degree k and bandwidth h_v: $\hat{\phi}_i = \hat{\phi}(x_i, k, h_v)$ given by (8), with $i \in s_v$, $k = 0, 1, 2, 3$, $h_v = 0.15, 0.25, 0.50$ and the Epanechnikov kernel function
$$K(t) = (3/4)(1 - x^2) I\{|x| \le 1\}.$$
4. Unweighted local polynomial regression of degree k and bandwidth h_v: the same as above but not including the sampling weights in (8) to obtain the $\hat{\phi}_i = \hat{\phi}(x_i, k, h_v)$. This might be somewhat easier to compute in practice and should lead to similar results, even if it does not, strictly speaking, follow the pseudo-randomization theory of Section 3.
5. Weighting within cell: within each stratum, respondents and nonrespondents were classified into four classes of age based on the sample quartiles of this variable. This procedure subdivided the sample in a total of 60 cells. Let s_g and s_{rg} denote respectively the set of sampled elements and the set of responding elements in the g^{th} cell. Then, the WC adjustment is defined by taking
$$\hat{\phi}_i = \frac{\sum_{i \in s_{rg}} w_i}{\sum_{i \in s_g} w_i},$$
for all respondents $i \in s_{rg}$.
6. Naive: $\hat{\phi}_i = 1$, $i \in s_v$.

5.3 Bias and robustness against a misspecified response propensity function

When the full sample without artificial nonresponse is used, the Hájek estimated means for the four study variables are respectively SBP = 122.19 mm Hg, DBP = 70.29 mm Hg, HTN = 29.04% and HTC = 15.76%. Table 1 gives the percentage bias relative to those means across response sets obtained for every adjustment procedure in this simulation experiment. For both weighted and unweighted Local Polynomial Regression adjusted estimators, we only display the results for the bandwidth $h_v = 0.25$, but those for other bandwidth values are similar. We instead show the results for different degrees of the local polynomial, so that the effect of moving from local constant to higher order polynomials can be evaluated.

Table 1
Relative biases (%) of nonresponse-adjusted estimators for mean systolic blood pressure (SBP), diastolic blood pressure (DBP), indicator of hypertension (HTN) and indicator of high serum total cholesterol (HTC), based on 1,000 response sets for two propensity functions of the age of the survey participant in NHANES 2005-2006

	Logistic propensity function (linear predictor)				Logistic propensity function (nonlinear predictor)			
Type of adjustment	**SBP**	**DBP**	**HTN**	**HTC**	**SBP**	**DBP**	**HTN**	**HTC**
True Response Propensities	0.01	0.01	-0.01	0.04	-0.00	-0.00	0.01	-0.22
Logistic Regression	0.01	0.00	-0.03	0.03	0.47	-1.67	6.49	-6.76
Weighted Local Polynomial Regression:								
Degree 0	0.27	0.34	3.39	2.41	-0.20	-0.39	-1.20	-2.27
Degree 1	0.00	0.04	-0.03	0.20	-0.01	-0.49	0.34	-2.36
Degree 2	0.01	0.01	0.03	0.07	0.03	-0.05	0.51	-0.27
Degree 3	0.01	0.01	-0.02	0.04	-0.03	-0.05	-0.24	-0.44
Unweighted Local Polynomial Regression:								
Degree 0	0.11	0.24	1.34	1.53	-0.17	-0.47	-0.98	-2.70
Degree 1	0.01	0.05	-0.00	0.25	-0.01	-0.57	0.34	-2.69
Degree 2	0.01	0.01	-0.00	0.07	0.01	-0.07	0.26	-0.40
Degree 3	0.00	0.01	-0.06	0.03	-0.03	-0.06	-0.29	-0.48
Weighting Within Cell	0.08	0.08	0.84	0.69	-0.11	-0.07	-0.84	-0.48
Naive	1.62	0.80	20.49	8.04	-1.30	-1.60	-15.61	-10.77

Among the estimators affected by the generated nonresponse, the worst bias performances are clearly for the unadjusted "Naive" estimator. As displayed in the last row of Table 1, the biases are higher in the estimation of the prevalence of hypertension and the mean systolic blood pressure, as these are the characteristics of the study variables with higher correlations with the AGE variable, and also for the prevalence of high serum total cholesterol. The biases of the Naive estimator can be successfully reduced with the true response propensity estimator, any of the local polynomial regression adjusted estimators, the weighting-within cell estimator or with the logistic adjusted estimator, if the model for the propensity function is correctly specified. The best performances in terms of small bias are obtained using the estimator adjusted by the true response propensities, because it is conditionally unbiased for the full sample estimates. The logistic adjustment when it is applied under the correct model, given by the propensity function with a linear predictor, also gives nearly unbiased estimates. For the second propensity function, where the form of the predictor is not well captured by the logistic regression fit of a regression line, this adjustment yields a conditionally biased estimator.

The averages of the local polynomial regression estimates become generally closer to the full sample estimates by increasing the degree of the polynomial fitted, with the largest jump when moving from a local constant to a local linear estimator. Hence, it seems that local polynomial regression is indeed superior to kernel regression in this context. There is very little difference between the weighted and unweighted forms of this adjustment and both procedures have overall smaller conditional biases than the biases of the weighting-within cell estimator, when they are implemented by fitting locally a polynomial of order greater than zero to estimate the response propensities. The zero degree propensity weighted and unweighted adjusted estimators have smaller biases at smaller bandwidths, as we observed with the bandwidth 0.15, for instance, but smaller bandwidths tend to increase the variance of the estimators. Overall, both weighted and unweighted local polynomial regression adjustments outperform the parametric logistic adjustment when the response model is misspecified. By implementing the local polynomial adjustments with degrees above one, their performances are similar to the one of the logistic adjustment under the correct specification of the response model.

5.4 Variance and variance estimation

Table 2 shows the variance of the adjustment methods considered here across the nonresponse replicates, and we normalized them by the variance for the true response propensity adjustment for clarity. Interestingly, there appears to be an inverse relationship between the magnitude of the relative biases in Table 1 and the variances in this table. In those cases where the relative bias was small (the weighted and unweighted local polynomial regression, the weighting within cell as well as the logistic regression adjustment for the linear propensity function), all the methods appear to result in roughly similar variances. There is a tendency for higher degree local polynomials to be more variable than lower degree ones, and this is particularly noticeable for the nonlinear propensity function, where a clear jump is seen when one moves from degree 1 (local linear) to 2 (local quadratic). Overall, it seems that local linear regression, either weighted or unweighted, offers a good compromise between the bias and the variance of the nonresponse adjustment procedure.

Table 2
Normalized Monte Carlo variances of nonresponse-adjusted estimators for mean systolic blood pressure (SBP), diastolic blood pressure (DBP), indicator of hypertension (HTN) and indicator of high serum total cholesterol (HTC), based on 1,000 response sets for two propensity functions of the age of the survey participant in NHANES 2005-2006

	Logistic propensity function (linear predictor)				Logistic propensity function (nonlinear predictor)			
Type of adjustment	**SBP**	**DBP**	**HTN**	**HTC**	**SBP**	**DBP**	**HTN**	**HTC**
True Response Propensities	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Logistic Regression	85.9	92.4	79.5	96.5	63.9	61.5	54.1	52.0
Weighted Local Polynomial Regression:								
Degree 0	74.9	81.2	65.7	92.1	70.3	67.0	67.4	75.0
Degree 1	81.8	89.5	66.2	92.7	73.6	69.8	68.9	76.0
Degree 2	81.3	89.8	65.5	94.0	90.3	81.7	88.0	96.1
Degree 3	82.3	90.2	65.8	93.1	90.1	82.2	87.7	96.2
Unweighted Local Polynomial Regression:								
Degree 0	82.2	85.8	77.6	95.8	71.9	69.2	70.7	74.7
Degree 1	85.6	90.1	79.4	95.7	74.4	71.1	71.2	74.6
Degree 2	86.6	91.3	79.3	96.1	91.8	84.5	91.8	96.8
Degree 3	87.3	91.5	78.5	95.0	91.2	84.7	91.2	96.9
Weighting Within Cell	79.7	89.1	62.1	91.6	82.5	77.0	81.1	92.3
Naive	71.3	58.0	81.7	74.6	48.6	48.7	45.5	45.1

The above simulation results showed the behavior of several nonresponse adjustments in the NHANES setting. We now consider the replication variance estimation approach of Section 4 and evaluate its usefulness as a sample-based measure of uncertainty for the nonresponse-adjusted estimators in the same setting. We implemented (17) with the Jackknife method. Since NHANES does not provide information on the joint sample inclusion probabilities, we could not apply a full Jackknife variance estimator as in, for instance, Berger and Skinner (2005), as a means to account for the selection of units with varying probabilities in the survey. Because of this, we assumed the within-stratum designs in NHANES could be approximated by cluster sampling with replacement and rewrite (17) in the form proposed by Rust (1985),

$$\hat{\mathrm{V}}_{JK}(\hat{\theta}) = \sum_{t=1}^{T} c_t \sum_{j \in s_t} (\hat{\theta}^{(tj)} - \hat{\theta})^2, \qquad (18)$$

where s_t denote the set of units in sample from the t^{th} NHANES stratum, $t = 1, 2, ..., T$, n_t be the number of units selected to s_t, $c_t = (n_t - 1)/n_t$ and $\hat{\theta}^{(tj)}$ is obtained from (5) by replacing the w_i with the replication weights

$$w_{i(tj)} = \begin{cases} 0, & \text{for a survey participant } i \in \text{PSU } j,\ j \in s_t \\ n_t/(n_t - 1)\, w_i, & \text{for a survey participant } i \in \text{PSU } j',\ j' \in s_t\ (j' \neq j) \\ w_i, & \text{for a survey participant } i \notin s_t. \end{cases}$$

These weights were also applied in the estimation of the response propensities for the weighted local polynomial regression procedure adjustment procedure.

The Jackknife variance estimator (18) was applied to each response vector from the two propensity functions, yielding estimates $\hat{v}_{JK}(\hat{\theta}(b))$, $b = 1, 2, ..., B$, for all adjusted estimators in the experiment. For the sake of comparison, it would be informative to produce estimates of the corresponding variances by the Monte Carlo method. However, as the NHANES sample is fixed, the Monte Carlo variance of the point estimates $\hat{\theta}(b)$ across response vectors estimates only the conditional variance $\mathrm{Var}(\hat{\theta}|s_{\mathrm{v}})$ with respect to the response model. Since

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\mathrm{E}(\hat{\theta}|s_{\mathrm{v}})) + \mathrm{E}(\mathrm{Var}(\hat{\theta}|s_{\mathrm{v}})),$$

where the "inner" moments are taken with respect to the response model given the sample and the "outer" moments are with respect to the sampling design, the design variance of $\mathrm{E}(\hat{\theta}|s_{\mathrm{v}})$ needs to be accounted for in order to have a valid estimation target for $\hat{\mathrm{V}}_{JK}(\hat{\theta})$. Using the fact that weighted and unweighted local polynomial regression and weighting within cell all produce approximately conditionally unbiased estimators of the full sample estimator, $\bar{y}_{\pi,\,\mathrm{rat}} = \sum_{i \in s_{\mathrm{v}}} w_i y_i / \sum_{i \in s_{\mathrm{v}}} w_i$, for the two response propensities functions, we decided to use the Jackknife variance estimator of $\bar{y}_{\pi,\,\mathrm{rat}}$ as a "proxy" for $\mathrm{Var}(\mathrm{E}(\hat{\theta}|s_{\mathrm{v}}))$. Hence, our "comparison variance" will be defined as

$$\hat{v}_C(\hat{\theta}) = \hat{v}_{JK}(\bar{y}_{\pi,\,\mathrm{rat}}) + \frac{1}{B-1}\sum_{b=1}^{B} (\hat{\theta}(b) - \hat{\theta})^2.$$

Using $\hat{v}_C(\hat{\theta})$ instead of the true variance will tend to underestimate any bias issues associated with the use of the jackknife variance estimator for the full sample estimator. However, it will show how well the replication procedure manages to capture the nonresponse variability.

Table 3 gives relative biases of the Jackknife variance estimators obtained in this experiment. The results show that the jackknife variance estimator performs reasonably well for both nonresponse mechanisms and all estimators considered. The weighted local polynomial regression adjusted procedure appears to yield estimated variances in greater consonance with the comparison variance than when the procedure is implemented by its unweighted version. The results for the nonlinear predictor function exhibit more bias than those for the linear predictor, with more pronounced positive and negative biases present for the former for all the variables. As discussed in Da Silva and Opsomer (2006), replication methods for nonresponse-adjusted estimators often ignore a component of the total variance, which includes the effect of both sampling and the response mechanism. We therefore conjecture that the different bias behaviors exhibited for the different variables could be due to this missing variance component.

6. Concluding remarks

In this article, we studied properties of nonparametric propensity weighting as an adjustment procedure for survey nonresponse. The local polynomial regression technique is seen to offer a flexible way of constructing new survey adjustments for nonresponse. The results in the article extend those in Da Silva and Opsomer (2006) by allowing the use of local polynomials of arbitrary degree, which offers both theoretical and practical advantages over zero-degree kernel regression.

In addition to its good theoretical properties, the estimator was shown in the simulation experiment to be competitive with an estimator based on a correctly specified parametric model in terms of bias and variance, while protecting against a potentially misspecified model. The weighting-cell estimator is similarly robust against model misspecification, but a particular advantage of nonparametric regression methods over weighting cell approaches is the connection to broad classes of modeling techniques available in the non-survey literature. Extensions of the methodology we described here to semiparametric and (generalized) additive models (Hastie and Tibshirani 1986) are readily formulated and should work well in a wide range of potential response model scenarios, including situations with multiple covariates that are both categorical and continuous. A detailed discussion of these extensions is beyond the scope of the current paper, however.

In Section 5, we applied the nonparametric nonresponse adjustment to NHANES data by modeling the response probability as a smooth function of the age of the respondents, and weighting the data by the inverses of the estimated response probabilities. The same approach can be used in other survey datasets whenever continuous covariates related to the response probability are available for all elements in the original sample. This provides a viable alternative to the commonly used weighting-within-cell approach for situations in which cells are constructed by "binning" one or several continuous variables.

Table 3
Relative biases (%) of the Jackknife variance estimators of estimators of the mean systolic blood pressure (SBP), diastolic blood pressure (DBP), indicator of hypertension (HTN) and indicator of high serum total cholesterol (HTC), based on 1,000 response sets for two propensity functions of the age of the survey participant in NHANES 2005-2006

	Logistic propensity function (linear predictor)				Logistic propensity function (nonlinear predictor)			
Type of adjustment	**SBP**	**DBP**	**HTN**	**HTC**	**SBP**	**DBP**	**HTN**	**HTC**
True Response Propensities	0.55	-0.47	-0.06	0.16	0.92	-0.26	-1.03	-2.76
Weighted Local Polynomial Regression:								
Degree 0	-0.66	2.33	2.74	4.44	1.63	-2.27	-5.12	-9.44
Degree 1	-0.31	-1.03	0.31	1.87	5.27	4.03	2.60	-9.95
Degree 2	-0.14	-0.76	0.41	0.49	0.25	0.65	-2.60	-3.60
Degree 3	-0.27	-1.03	0.39	0.48	0.19	0.45	-2.19	-3.02
Unweighted Local Polynomial Regression:								
Degree 0	2.00	2.77	3.57	5.56	5.73	0.31	1.83	-10.22
Degree 1	2.02	1.06	2.63	2.61	7.46	5.57	4.33	-10.43
Degree 2	2.26	1.07	2.88	1.36	4.16	3.81	1.62	-2.94
Degree 3	2.21	1.01	2.94	1.46	3.45	3.65	0.96	-2.63
Weighting Within Cell	-1.15	1.70	-0.47	5.16	2.69	-6.91	3.06	-5.88

There are still a number of open issues that need to be further investigated with respect to implementation of the method in actual surveys, whether in the univariate case described in detail here or in the various model extensions just mentioned. An important practical issue is the selection of estimator settings such as the degree of the local polynomial and the bandwidth. As noted in the non-parametric literature (*e.g.*, Fan and Gijbels 1996, page 77) and also confirmed in the simulations, higher degree polynomials reduce the bias but increase the variance, so that polynomials of degree $k = 1$ or 2 are generally recommended as a good compromise. More critical is the choice of bandwidth parameter. In our simulations, the results were only modestly sensitive to the choice of bandwidth within a "reasonable" range of values, *i.e.*, ones ensuring that the number of observations used for estimating $\phi(x)$ at any x does not become too small (see discussion at the end of Section 2), or that is so large that the fit cannot capture changes in $\phi(\cdot)$ over the range of x. As a rule of thumb, we would recommend considering values for h that are within 20% and 50% of the range of x as a good place to start, and making a final determination by looking at both model diagnostics for the model fit $\hat{\phi}(x)$ and weight diagnostics for the adjusted survey weights $(\pi_i \hat{\phi}_i)^{-1}$, similarly as would be done when constructing cell-based weights.

Acknowledgements

We thank the Associate Editor and two referees for their useful comments. The first author was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, under the grant Projeto Universal 480518/2004-1.

Appendix

A.1 Assumptions

We now state the assumptions needed to derive our main results. A detailed discussion of these assumptions is provided in Da Silva and Opsomer (2008). Consider the asymptotic framework of Section 3. Let $\boldsymbol{I}_\nu = (I_1, I_2, ..., I_{N_\nu})'$ be the sample inclusion indicator vector for the ν^{th} population. Suppressing the ν for ease of notation, let $\pi_i = \Pr(I_i = 1)$, and let

$$\Delta_{j_1, ..., j_k} \equiv \mathrm{E_d}\left(\prod_{\ell=1}^{k} (I_{j_\ell} - \pi_{j_\ell})\right) \qquad (19)$$

denote higher moments for the sample inclusion indicators $I_{j_1}, I_{j_2}, ..., I_{j_k}$ with respect to the sampling design. We assume that there are positive constants $\lambda_1, \lambda_2, ..., \lambda_6$ such that:

(A1) $\lambda_1 < N_\nu n_\nu^{-1} \pi_i < \lambda_2 < \infty, \forall\, i \in U_\nu$;

(A2) $N_\nu^{-1} n_\nu \to \pi$, for some $0 < \pi < 1$, as $\nu \to \infty$;

(A3) For distinct $j_1, j_2, ..., j_k \in U_\nu$, where $k = 2, 3, ..., 8$,

$$|\Delta_{j_1, ..., j_k}| \le \begin{cases} \left[\prod_{\ell=1}^{k} (N - \ell + 1)\right]^{-1} n_\nu^{\frac{k}{2}} \lambda_3 & \text{if } k \text{ is even,} \\ \left[\prod_{\ell=1}^{k} (N - \ell + 1)\right]^{-1} n_\nu^{\frac{k-1}{2}} \lambda_4 & \text{if } k \text{ is odd} \end{cases}$$

(A4) $\lim_{\nu \to \infty} N_\nu^{-1} \sum_{i \in U_\nu} y_i = \mu \in (-\infty, \infty)$ and $N_\nu^{-1} \sum_{i \in U_\nu} |y_i|^4 \le \lambda_5$, for all $\nu \ge 1$.

Let $\boldsymbol{R}_\nu = (R_1, R_2, ..., R_{N_\nu})'$ denote the response indicator vector for the ν-th population. In addition to the assumptions on the sampling design and the population distribution of the variable Y, we will also need the following assumptions on the response mechanism:

(B1) $R_1, R_2, ..., R_{N_\nu}$ are independent random variables;

(B2) $\Pr\{R_i = 1 \mid \boldsymbol{I}_\nu, \boldsymbol{y}_\nu, \boldsymbol{x}_\nu\} = \Pr\{R_i = 1 \mid \boldsymbol{x}_\nu\} \equiv \phi_i, \forall\, i \in U_\nu$;

(B3) $\phi_i = \phi(x_i), \forall\, i \in U_\nu$, where $\phi(\cdot)$ is a $(k+2)^{\text{th}}$ continuously differentiable function with $\lambda_6 < \phi(\cdot) \le 1$. The first derivative $\phi'(\cdot)$ has a finite number of sign changes.

Regarding the distribution of the x_i and the kernel estimator, we assume that:

(C1) For all $\nu \ge 1$, $x_1, x_2, ..., x_{N_\nu}$ are realizations of random variables $X_1, X_2, ..., X_{N_\nu}$ independent and identically distributed with distribution $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$, where $f_X(\cdot)$ is a continuous and positive probability density function on a compact set $[a_X, b_X]$;

(C2) The kernel function $K(\cdot)$ is a bounded and continuous probability density, which is symmetric around zero and supported on [-1, 1];

(C3) $\int_{-1}^{1} |z|^{k+4} K(z)\, dz < \infty$;

(C4) For all $\nu \ge 1$, $\{h_\nu\}$ is a sequence of bandwidths satisfying $0 < h_\nu \le 1$, $h_\nu \to 0$, $n_\nu h_\nu^2 \to \infty$ and $N_\nu h_\nu / \log N_\nu \to \infty$, as $\nu \to \infty$;

(C5) The first derivative $f_X'(\cdot)$ is continuously differentiable and contains a finite number of sign changes on $\mathrm{supp}(f_X)$. The first derivative $K'(\cdot)$ has a finite number of sign changes on $\mathrm{supp}(K)$;

(C6) The matrix $N_\nu \mathbf{T}_i^{-1}$ is non-singular for all $i \in U_\nu$ and all $\nu \geq 1$.

A.2 Technical derivations

Complete proofs are in Da Silva and Opsomer (2008). The proof of Theorem 1 relies on bounding the moments of the difference $\overline{y}_{\pi\hat{\psi}\nu} - \overline{y}_{\pi\tilde{\phi}\nu}$ under the combined design and response model probability mechanism, followed by deriving the rates of convergence for the bias and variance of the linearized estimator $\overline{y}_{\pi\hat{\psi}\nu}$. This is done in a series of six lemmas, which are stated here without proof. The proof of Theorem 2 is based on the result of Theorem 1, followed by an additional linearization of the ratio form.

For notational simplicity in what follows, we suppressed the fact that the results are conditional on the sequences $\boldsymbol{x}_\nu = (x_1, ..., x_{N_\nu})$ in the populations U_ν. However, the results in these lemmas are shown to hold with probability one over these sequences in Da Silva and Opsomer (2008), as was also done in Da Silva and Opsomer (2006). Hence, the results can be interpreted to hold for all population sequences, except on a set of probability 0 with respect to the distribution of the $\boldsymbol{x}_\nu$.

Lemma 1. *Assume that assumptions* (*C*1)-(*C*5) *hold. Consider* $\mu_\ell(K, x) = \int_{D_{x,h}} z^\ell K(z)\,dz$, *where* $D_{x,h_\nu} = \{t: (x + ht) \in \text{supp}(f_X)\} \cap \text{supp}(K)$. *Then, for all* $\ell = 0, 1, ..., k + 2$,

$$\sup_{x \in \text{supp}(f_X)} \left| \frac{1}{N_\nu h_\nu} \sum_{j \in U_\nu} K\left(\frac{X_j - x}{h_\nu}\right)(X_j - x)^\ell - E_\nu(x, \ell) \right| \underset{\nu \to \infty}{\to} 0,$$

where

$$E_\nu(x, \ell) = f_X(x)\mu_\ell(K, x)h_\nu^\ell + f_X'(x)\mu_{\ell+1}(K, x)h_\nu^{\ell+1} + o(h_\nu^{\ell+1}).$$

Lemma 2. *Assume that assumptions* (*C*1)-(*C*5) *hold. Consider the population fit* $\tilde{\phi}_i = \tilde{\phi}(x_i, k, h_\nu)$, $i \in U_\nu$, *defined in* (10). *Hence, for all* $i \in U_\nu$, *there exists positive bounded terms* $c_1(x_i)$, $c_2(x_i)$ *and* $c_3(x_i)$, *such that if* x_i *in an interior point of* $\text{supp}(f_X)$

$$\tilde{\phi}_i - \phi(x_i) = \begin{cases} c_1(x_i)h_\nu^{k+2} + o(h_\nu^{k+2}) & k \text{ is even} \\ c_2(x_i)h_\nu^{k+1} + o(h_\nu^{k+1}) & k \text{ is odd} \end{cases}$$

and if x_i *in a boundary point of* $\text{supp}(f_X)$

$$\tilde{\phi}_i - \phi(x_i) = c_3(x_i)h_\nu^{k+1} + o(h_\nu^{k+1}),$$

where all the smaller order terms hold uniformly in $i \in U_\nu$.

Lemma 3. *Assume that assumptions* (*C*1) *and* (*C*4) *hold. Then*,

i) *For* $p \in [0, \infty)$ *fixed*,

$$\limsup_{\nu \to \infty} \left(\frac{1}{N_\nu h_\nu} \sum_{j \in U_\nu} I_{\{x - h_\nu \leq x_j \leq x + h_\nu\}} \right)^p < \infty,$$

uniformly in x;

ii) $\limsup_{\nu \to \infty} \frac{1}{2N_\nu h_\nu} \sum_{j \in U_\nu} I_{\{x_j \in [0, h_\nu] \cup (1 - h_\nu, 1]\}} < \infty$;

iii) $\limsup_{\nu \to \infty} \frac{1}{N_\nu} \sum_{j \in U_\nu} I_{\{x_j \in (h_\nu, 1 - h_\nu]\}} < \infty$.

iv) *there exists* ν^*, *independent of* x, *such that whenever* $\nu \geq \nu^*$,

$$\sum_{j \in U_\nu} I_{\{|x_j - x| \leq h_\nu\}} \geq k + 1;$$

Lemma 4. *Suppose the assumptions of Theorem 1 hold. Consider the matrices* $\hat{\mathbf{T}}_{si} = \{\hat{T}_{si,\,pq}\}$ *and* $\mathbf{T}_i = \{T_{si,\,pq}\}$ *and the vectors* $\hat{\mathbf{t}}_{si} = \{\hat{t}_{si,\,p}\}$, $\mathbf{t}_i = \{t_{i,\,p}\}$ *and* $\mathbf{B}_i = \{B_{i,\,p}\}$ *given in* (7) *and* (10). *Then*,

i) *the* $N_\nu^{-1} T_{i,\,pq}$ *and* $N_\nu^{-1} t_{i,\,p}$ *are uniformly bounded in* $i \in U_\nu$, *for all* $p, q = 1, ..., k + 1$;

ii) *the* $\hat{T}_{si,\,pq}$ *and* $\hat{t}_{si,\,p}$ *satisfy*

$$\max_{1 \leq p,\, q \leq k+1} \text{E}\left(\frac{\hat{T}_{si,\,pq} - T_{i,\,pq}}{N_\nu} \right)^8 = O\left(\frac{1}{n_\nu^4 h_\nu^4} \right) \text{ and}$$

$$\max_{1 \leq p \leq k+1} \text{E}\left(\frac{\hat{t}_{si,\,p} - t_{i,\,p}}{N_\nu} \right)^8 = O\left(\frac{1}{n_\nu^4 h_\nu^4} \right),$$

uniformly in $i \in U_\nu$;

iii) *the random variable* $\boldsymbol{e}_1' \mathbf{T}_i^{-1}(\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si}\mathbf{B}_i)$ *satisfies*

$$\max_{i \in U_\nu} \text{E}\,(\boldsymbol{e}_1' \mathbf{T}_i^{-1}(\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si}\mathbf{B}_i) I_i R_i) = O\left(\frac{1}{n_\nu h_\nu} \right) \quad (20)$$

and

$$\max_{i \in U_\nu} \text{E}\,(\boldsymbol{e}_1' \mathbf{T}_i^{-1}(\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si}\mathbf{B}_i))^4 = O\left(\frac{1}{n_\nu^2 h_\nu^2} \right). \quad (21)$$

Lemma 5. *Suppose the assumptions of Theorem 1 hold. Then, for all* $\nu \geq 1$

i) *the reciprocal of* $\tilde{\phi}_i$ *is uniformly bounded in* $i \in U_\nu$;

ii) *the partial derivatives of* $\hat{\phi}_i^{-1}$ *of orders one up to four, when evaluated at* $\hat{\mathbf{T}}_{si} = \mathbf{T}_i$, $\hat{\mathbf{t}}_{si} = \mathbf{t}_i$, $\delta_1 = 0$ *and* $\delta_2 = 0$, *are uniformly bounded in* $i \in U_\nu$;

iii) $\text{E}(\hat{\phi}_i^{-4})$ *is uniformly bounded in* $i \in U_\nu$;

iv) *the reciprocal of* $\hat{\phi}_i$ *satisfies*

$$\hat{\phi}_i^{-1} = \tilde{\phi}_i^{-1} - \tilde{\phi}_i^{-2}\, \boldsymbol{e}_1'\, \mathbf{T}_i^{-1}(\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si}\, \mathbf{B}_i) + \varepsilon_{iv} + O\left(\frac{1}{N_v^2\, h_v^2}\right), \tag{22}$$

uniformly in $i \in U_v$, *where the* ε_{iv} *are random variables such that*

$$\max_{i \in U_v} \mathrm{E}\left(\varepsilon_{iv}^2\right) = O\left(\frac{1}{n_v^2 h_v^2}\right).$$

Lemma 6. *Suppose the assumptions of Theorem 1 hold. Define the random variables* $\overline{y}_{\pi\tilde{\phi}v}$, $\overline{d}_{\pi\tilde{\phi}v}$ *and* $\overline{\varepsilon}_{\pi\tilde{\phi}v}$ *as*

$$(\overline{y}_{\pi\tilde{\phi}v},\ \overline{d}_{\pi\tilde{\phi}v},\ \overline{\varepsilon}_{\pi\tilde{\phi}v})' = \frac{1}{N_v} \sum_{i \in s_v} \pi_i^{-1} \tilde{\phi}_i^{-1} (1,\ \tilde{\phi}_i^{-1} e_1' \mathbf{T}_i^{-1} (\hat{\mathbf{t}}_{si} - \hat{\mathbf{T}}_{si} \mathbf{B}_i),\ \varepsilon_{iv})' y_i R_i.$$

Then,

$$\mathrm{E}(\overline{y}_{\pi\tilde{\phi}v} - \overline{y}_{N_v}) = \begin{cases} O(h_v^{k+(3/2)}) & k \text{ even}, \\ O(h_v^{k+1}) & k \text{ odd}, \end{cases} \tag{23}$$

$$\mathrm{Var}(\overline{y}_{\pi\tilde{\phi}v}) = O\left(\frac{1}{n_v}\right), \tag{24}$$

$$(\mathrm{E}[\overline{d}_{\pi\tilde{\phi}v}],\ \mathrm{E}[\overline{d}_{\pi\tilde{\phi}v}^2 \hat{A}])' = O\left(\frac{1}{n_v h_v}\right) \tag{25}$$

and

$$\mathrm{E}(\overline{\varepsilon}_{\pi\tilde{\phi}v}^2) = O\left(\frac{1}{n_v^2\, h_v^2}\right). \tag{26}$$

References

Alho, J.M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika*, 617-624.

Berger, Y.G., and Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society*, Series B: *Statistical Methodology*, 67(1), 79-89.

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 1026-1053.

Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete data in sample surveys*: *Theory and bibliographies*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin). Academic Press, New York: London, 3, 143-160.

Da Silva, D.N., and Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics*, 4, 563-579.

Da Silva, D.N., and Opsomer, J.D. (2008). Theoretical properties of propensity weighting for survey nonresponse through local polynomial regression. Technical Report #2008/6, Department of Statistics, Colorado State University.

David, M.H., Little, R., Samuhel, M. and Triest, R. (1983). Imputation models based on the propensity to respond. In *ASA Proceedings of the Business and Economic Statistics Section*, 168-173.

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 325-337.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. In *ASA Proceedings of the Social Statistics Section*, 197-202.

Giommi, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron*, 4, 185-200.

Groves, R., Dillman, D., Eltinge, J. and Little, R.J.A. (2002). *Survey Nonresponse*. New York: John Wiley & Sons, Inc.

Hastie, T.J., and Tibshirani, R.J. (1986). Generalized additive models. *Statistical Science*, 297-318.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. In *ASA Proceedings of the Section on Survey Research Methods*, 637-642.

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

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 4, 501-514.

Laaksonen, S. (2006). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications*, 2, 95-100.

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

Nargundkar, M., and Joshi, G.B. (1975). Non-response in sample surveys. In *40th Session of the ISI*, *Warsaw 1975*, *Contributed papers*, 626-628.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustments for unit non-response. In *Incomplete data in sample surveys*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin), *Theory and bibliographies*, Academic Press, New York: London. 2, 143-184.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41-55.

Rust, K. (1985). Variance estimation for complex estimators in sample survey. *Journal of Official Statistics*, 381-397.

Wand, M.P., and Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 177-190
Statistics Canada, Catalogue No. 12-001-X

Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design

Jan van den Brakel and Sabine Krieg [1]

Abstract

In this paper a multivariate structural time series model is described that accounts for the panel design of the Dutch Labour Force Survey and is applied to estimate monthly unemployment rates. Compared to the generalized regression estimator, this approach results in a substantial increase of the accuracy due to a reduction of the standard error and the explicit modelling of the bias between the subsequent waves.

Key Words: Small area estimation; Rotation group bias; Survey errors.

1. Introduction

The Dutch Labour Force Survey (LFS) is based on a rotating panel design. Each month a sample of addresses is drawn and data are collected by means of computer assisted personal interviewing of the residing households. The sampled households are re-interviewed by telephone four times at quarterly intervals. The estimation procedure of this survey is based on the generalized regression (GREG) estimator, developed by Särndal, Swensson and Wretman (1992).

Due to the following properties, GREG estimators are very attractive to produce official releases in a regular production environment and are therefore widely applied by national statistical institutes. First, GREG estimators are approximately design-unbiased, which provides a form of robustness in the case of large sample sizes. These estimators are derived from a linear regression model that specifies the relationship between the values of a certain target parameter and a set of auxiliary variables for which the totals in the finite target population are known. If this linear regression model explains the variation of the target variable reasonably well, then this might reduce the design variance as well as the bias due to selective nonresponse, Särndal and Swensson (1987), Bethlehem (1988), and Särndal and Lundström (2005). Model misspecification, on the other hand, might result in an increase of the design variance but the point estimates remain approximately design unbiased. Second, GREG estimators are often used to produce one set of weights for the estimation of all target parameters of a multi-purpose sample survey. This is not only convenient but also enforces consistency between the marginal totals of different publication tables.

There are two major problems with the rotating panel design of the LFS and the way that the GREG estimator is applied in the estimation procedure. First, there are substantial systematic differences between the subsequent waves of the panel due to mode- and panel effects. This is a well-known problem for rotating panel designs, and is in the literature referred to as rotation group bias (RGB), see Bailar (1975). In the LFS, the level of the unemployment rate in the subsequent waves is substantially smaller compared to the first wave. There are also systematic differences between the seasonal effects of the subsequent waves.

A second problem is that the monthly sample size of the LFS is too small to rely on the GREG estimator to produce official statistics about the monthly employment and unemployment. GREG estimators have a relatively large design variance in the case of small sample sizes. Therefore, in the LFS, each month the samples observed in the preceding three months are used to estimate quarterly figures about the labour market situation. The major drawback of this approach is that the real monthly seasonal pattern in the unemployment rate is smoothed out. Also structural changes in unemployment appear delayed in the series of quarterly figures.

Since the monthly sample size is too small to apply design-based or direct survey estimators, model-based estimation procedures might be used to produce sufficiently reliable statistics. In the case of continuously conducted surveys, a structural time series model can be applied to use information from preceding samples to improve the accuracy of the estimates. This model can be extended to account for the RGB and the autocorrelation (AC) between the different panels of the LFS. This approach makes efficient use of the rotating panel design of the LFS in estimating monthly figures about the labour market, and is originally proposed by Pfeffermann (1991) and Pfeffermann, Feder and Signorelli (1998). These techniques are applied in this paper to estimate the monthly unemployment rate of the LFS. Other references to authors that apply

1. Jan van den Brakel and Sabine Krieg, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401CZ Heerlen, The Netherlands. E-mail: jbrl@cbs.nl and skrg@cbs.nl.

time series models to develop estimates for periodic surveys are Scott and Smith (1974), Scott, Smith and Jones (1977), Tam (1987), Binder and Dick (1989, 1990), Bell and Hillmer (1990), Tiller (1992), Rao and Yu (1994), Pfeffermann and Burck (1990), Pfeffermann and Rubin-Bleuer (1993), Pfeffermann and Tiller (2006), Harvey and Chung (2000), and Feder (2001).

Composite estimators can be considered as an alternative to time series models. They are developed under the traditional design-based approach, to use information observed in previous periods from periodic surveys with a rotating panel design, to improve the precision of level and change estimates. Some key references to composite estimators are Hansen, Hurwitz and Meadow (1953), Rao and Graham (1964), Gurney and Daly (1965), Cantwell (1990), Singh (1996), Gambino, Kennedy and Singh (2001), Singh, Kennedy and Wu (2001) and Fuller and Rao (2001).

In Section 2, the survey design of the LFS is summarised. A structural time series model that accounts for the rotating panel design of the LFS is developed in Sections 3 and 4. The results are detailed in Section 5. Some general remarks are made in Section 6.

2. The dutch Labour Force Survey

2.1 Sample design

The objective of the Dutch LFS is to provide reliable information about the labour market. Each month a sample of addresses is selected from which households are identified that can be regarded as the ultimate sampling units. The target population of the LFS consists of the non-institutionalised population aged 15 years and over residing in the Netherlands. The sampling frame is a list of all known occupied addresses in the Netherlands, which is derived from the municipal basic registration of population data. The LFS is based on a stratified two-stage cluster design of addresses. Strata are formed by geographical regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. All households residing at an address, up to a maximum of three, are included in the sample (in the Netherlands, there is generally one household per address). Since most target parameters of the LFS concern people aged 15 through 64 years, addresses with only persons aged 65 years and over are undersampled.

In October 1999, the LFS changed from a continuous survey to a rotating panel design. In the first wave, data are collected by means of computer assisted personal interviewing (CAPI). For all members of the selected households, demographic variables are observed. For the target variables only persons aged 15 years and over are interviewed. When a household member cannot be contacted, proxy interviewing is allowed by members of the same household. Households, in which one or more of the selected persons do not respond for themselves or in a proxy interview, are treated as nonresponding households. The respondents aged 15 through 64 years are re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing (CATI). During these re-interviews a condensed questionnaire is applied to establish changes in the labour market position of the respondents. Proxy interviewing is also allowed during these re-interviews. The monthly gross sample size averaged about 8,000 addresses when the LFS first changed to a rotating panel design. The monthly sample size gradually declined to about 6,500 addresses in 2008. During this period about 65% completely responding households are obtained.

2.2 Rotation group bias

The rotating panel design, described in Section 2.1, results in systematic differences between the estimates of the unemployment rate of the successive waves in one time period. In the literature, this phenomenon is known as RGB, see *e.g.*, Bailar (1975), Kumar and Lee (1983) and Pfeffermann (1991). The RGB in the LFS results in a systematic underestimation of the level of the unemployment rate in the CATI waves but also in systematic differences between the seasonal patterns. The RGB is a consequence of the following strongly confounded factors:

- Selective nonresponse between the subsequent waves, *i.e.*, panel attrition.
- Systematic differences between the populations that are reached with the CAPI and CATI modes. It is anticipated that these differences are relatively small, since telephone numbers are asked during the first interview. As a result, secret numbers and cell-phone numbers are also called.
- Mode-effects, *i.e.*, systematic differences in the data due to the fact that the interviews are conducted by telephone instead of face to face. Under the CAPI mode the interview speed is lower, respondents are more engaged with the interview and are more likely to exert the required cognitive effort to answer questions carefully. Also less socially desirable answers are obtained under the CAPI mode due to the personal contact with the interviewer. As a result, less measurement errors are expected under the CAPI mode (Holbrook, Green and Krosnick 2003, and Roberts 2007). Van den Brakel (2008) describes an experiment where the CAPI and CATI data collection modes are compared in the first wave of the LFS. It follows that the estimated unemployment rate is significantly smaller under the CATI mode.

– The fraction of proxy interviews is larger under the CATI mode (Van den Brakel 2008). This might result in an increased amount of measurement errors.
– Effects due to differences between the CAPI questionnaire and the CATI questionnaire. The CATI questionnaire is a strongly condensed version of the CAPI questionnaire since the re-interviews focus on changes in the labour market position of the respondents.
– Panel effects, *i.e.*, systematic changes in the behaviour of the respondents in the panel. For example, questions about activities to find a job in the first wave might increase the search activities of the unemployed respondents in the panel. Respondents might also adjust their answers in the subsequent waves systematically, since they learn how to keep the routing through the questionnaire as short as possible.

It is assumed that the estimates based on the first wave are the most reliable, since CAPI generally results in a higher data quality and the first wave does not suffer from the panel effects mentioned above. In order to minimize the effects of the RGB, the second, third, fourth and fifth waves are currently calibrated to the first wave as will be described in Section 2.3.

2.3 Regular estimation procedure

Target parameters about the employment and unemployment are defined as population totals or as ratios of two population totals. The unemployment rate, which is investigated in this paper, is defined as the ratio of the total unemployment to the total labour force. This population parameter is estimated as the ratio of the GREG estimate for the total unemployed labour force to the estimated total labour force. Each month estimates about the employment and unemployment for the preceding three months are published.

In an attempt to correct for the RGB, a rather laborious weighting procedure is used in the regular estimation procedure. The most important steps are summarized here. First, the inclusion probabilities are derived, which reflect the sampling design described above as well as the different response rates between geographical regions. Subsequently, the inclusion weights of each CATI wave are calibrated with the GREG estimator to the labour force status observed in the first wave. In the next step, the calibrated weights of the CATI waves and the inclusion weights of the CAPI wave are used as the design or starting weights of the GREG estimator, using a weighting scheme that is based on a combination of different socio-demographic classifications. The integrated method for weighting persons and families of Lemaître and Dufour (1987) is applied to obtain equal weights for persons belonging to the same household. Finally, a bounding algorithm proposed by Huang and Fuller (1978) is applied to avoid negative weights. This estimation procedure is conducted with the software package Bascula, Nieuwenbroek and Boonstra (2002).

Since this weighting procedure hardly corrects for the RGB, an additional rigid correction is applied. For the most important parameters the ratio between the estimates based on CAPI only and the estimates based on all waves is computed using the data of 12 preceding quarters. Estimates for the preceding three months are multiplied by this ratio to correct for RGB.

2.4 Monthly GREG estimates based on monthly data

In Section 3, a structural time series model is developed to estimate the monthly unemployment rate. The input data for this time series model are the GREG estimates for the monthly unemployment rate using the monthly sample data of the separate waves. Let θ_t denote the true but unknown unemployment rate for month t. Now Y_t^{t-j} denotes the GREG estimate of the unemployment rate of month t, based on the sample which entered the panel in month $t-j$. For the period of January 2001 until December 2008 each month five independent GREG estimates for the same parameter θ_t are produced, using the five separate waves that are observed each month, *i.e.*, Y_t^{t-j} for $j = 0, 3, 6, 9, 12$. These estimates are defined as

$$Y_t^{t-j} = \frac{t_{y,t}^{t-j}}{t_{z,t}^{t-j}}, \qquad (2.1)$$

with $t_{y,t}^{t-j}$ and $t_{z,t}^{t-j}$ the GREG estimates for the unemployed labour force and the labour force at time t, based on the sample that entered the panel at $t-j$.

The separate monthly waves are weighted with a reduced version of the weighting scheme that is applied in the regular weighting procedure for the quarterly figures. The estimates based on the CATI data are not adjusted to correct for RGB, since a multivariate time series model is applied to correct for this bias.

The variance of (2.1) can be estimated with

$$\text{var}(Y_t^{t-j}) = \frac{1}{(t_{z,t}^{t-j})^2}\sum_{h=1}^{H}\frac{n_{h,t}^{t-j}}{n_{h,t}^{t-j}-1}\left(\sum_{k=1}^{n_{h,t}^{t-j}}(w_k e_{k,t}^{t-j})^2 - \frac{1}{n_{h,t}^{t-j}}\left(\sum_{k=1}^{n_{h,t}^{t-j}} w_k e_{k,t}^{t-j}\right)^2\right), \qquad (2.2)$$

with

$$e_{k,t}^{t-j} = \sum_{l=1}^{m_k}(y_{kl,t}^{t-j} - \mathbf{x}_{kl}^T\mathbf{b}_y) - Y_t^{t-j}(z_{kl,t}^{t-j} - \mathbf{x}_{kl}^T\mathbf{b}_z).$$

Here $y_{kl,t}^{t-j}$ is a binary variable taking value one if the l^{th} person belonging to the k^{th} household that entered the sample at time $t-j$ belongs to the unemployed labour force at time t and zero otherwise, $z_{kl,t}^{t-j}$ a binary variable taking value one if the l^{th} person of the k^{th} household

belongs to the labour force at time t and zero otherwise, $\mathbf{x}_{kl}$ a vector with the auxiliary information of the l^{th} person belonging to the k^{th} household used in the weighting scheme of the GREG estimator, $\mathbf{b}_y$ and $\mathbf{b}_z$ the regression coefficient of the regression function of $y_{kl,t}^{t-j}$ respectively $z_{kl,t}^{t-j}$ on $\mathbf{x}_{kl}$, w_k the regression weight of household k, $n_{h,t}^{t-j}$ the number of completely responding households of stratum $h = 1, ..., H$, at time t of the sample that entered the panel at $t - j$, and m_k the number of persons aged 15 years and over belonging to the k^{th} household. Recall from Section 2.3 that persons belonging to the same household have equal weights due to the application of the integrated method for weighting persons and families of Lemaître and Dufour (1987). Formula (2.2) is the variance estimation procedure implemented in Bascula to approximate the variance of the ratio of two GREG estimators.

The estimates for the monthly unemployment rate obtained with the structural time series approach will be compared in Section 5.3 with monthly estimates based on the GREG estimator using the data observed in the five waves. For this comparison a slightly simplified version of the procedure described in Section 2.3 is applied to combine the data observed in the different waves to obtain monthly GREG estimates. First, a GREG estimate Y_t is computed using the data observed in the five waves using the same weighting procedure used in the regular production process to estimate quarterly figures, see Section 2.3. The weighting scheme is slightly simplified because less data are available. Subsequently a correction factor based on the preceding three years is computed as:

$$c_t = \frac{\sum_{j=0}^{35} Y_{t-j}^{t-j}}{\sum_{j=0}^{35} Y_{t-j}}. \qquad (2.3)$$

Finally, the corrected estimate is computed:

$$Y_t^c = c_t Y_t. \qquad (2.4)$$

Because the series start at January 2001, c_t can be computed from December 2003. To get a corrected GREG estimate for all months, $c_{\text{December2003}}$ is used in formula (2.4) for the periods preceding December 2003. The variance of (2.4) is approximated by $\text{var}(Y_t^c) = c_t^2 \text{var}(Y_t)$, where $\text{var}(Y_t)$, is computed with formula (2.2), using the data of all waves accordingly.

3. Time series model

Direct estimators, like the Horvitz-Thompson estimator or the GREG estimator, assume that the monthly unemployment rate θ_t is a fixed but unknown population parameter. Under this design-based approach, an estimator for θ_t for cross-sectional surveys only uses the data observed at time t. Data from the past are only used in the case of partially overlapping samples in a panel design, but not in the case of repeatedly conducted cross-sectional designs. Scott and Smith (1974) proposed to consider the population parameter θ_t as a realization of a stochastic process that can be described with a time series model. Under this assumption, data observed in preceding periods $t - 1$, $t - 2$, ..., can be used to improve the estimator for θ_t, even in the case of non-overlapping sample surveys.

Recall from Section 2.4 that Y_t^{t-j} denotes the GREG estimator for θ_t based on the panel observed at time t, which entered the survey for the first time at $t - j$. Due to the applied rotation pattern, each month a vector $\mathbf{Y}_t = (Y_t^t \; Y_t^{t-3} \; Y_t^{t-6} \; Y_t^{t-9} \; Y_t^{t-12})^T$ is observed. According to Pfeffermann (1991), this vector can be modelled as

$$\mathbf{Y}_t = \mathbf{1}_5 \theta_t + \boldsymbol{\lambda}_t + \boldsymbol{\gamma}_t + \mathbf{e}_t, \qquad (3.1)$$

with $\mathbf{1}_5$ a five dimensional vector with each element equal to one, $\boldsymbol{\lambda}_t = (\lambda_t^0 \lambda_t^3 \lambda_t^6 \lambda_t^9 \lambda_t^{12})^T$ and $\boldsymbol{\gamma}_t = (\gamma_t^0 \gamma_t^3 \gamma_t^6 \gamma_t^9 \gamma_t^{12})^T$ vectors with time dependent components that account for the RGB in the trend and the RGB in the seasonal components respectively, and $\mathbf{e}_t = (e_t^t \; e_t^{t-3} \; e_t^{t-6} \; e_t^{t-9} \; e_t^{t-12})^T$ the corresponding survey errors for each panel estimate. Time series models for the different components in (3.1), *i.e.*, the population parameter θ_t, the RGB for the trend $\boldsymbol{\lambda}_t$, the RGB for the seasonal patterns $\boldsymbol{\gamma}_t$, and the survey errors $\mathbf{e}_t$, are developed in Sections 3.1 through 3.3.

3.1 Time series model for the population parameter

With a structural time series model, the population parameter θ_t in (3.1) can be decomposed in a trend component, a seasonal component, and an irregular component, *i.e.*:

$$\theta_t = L_t + S_t + \varepsilon_t, \qquad (3.2)$$

where L_t denotes a stochastic trend component, S_t a stochastic seasonal component, and ε_t the irregular component. For the stochastic trend component the so-called local linear trend model is used, which is defined by the following set of equations:

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1} + \eta_{L,t}, \\ R_t &= R_{t-1} + \eta_{R,t}, \\ E(\eta_{L,t}) = 0,\ \text{Cov}(\eta_{L,t}, \eta_{L,t'}) &= \begin{cases} \sigma_L^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases} \\ E(\eta_{R,t}) = 0,\ \text{Cov}(\eta_{R,t}, \eta_{R,t'}) &= \begin{cases} \sigma_R^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases} \end{aligned} \qquad (3.3)$$

The parameters L_t and R_t are referred to as the trend and the slope parameter respectively. The seasonal component is modelled with the trigonometric form

$$S_t = \sum_{l=1}^{6} S_{l,t}, \qquad (3.4)$$

where

$$\begin{aligned} S_{l,t} &= S_{l,t-1}\cos(h_l) + S^*_{l,t-1}\sin(h_l) + \omega_{l,t} \\ S^*_{l,t} &= S^*_{l,t-1}\cos(h_l) - S_{l,t-1}\sin(h_l) \\ &\quad + \omega^*_{l,t},\ l = 1,..., 6, \\ h_l &= \frac{\pi l}{6},\ l = 1,...,6, \\ E(\omega_{l,t}) &= E(\omega^*_{l,t}) = 0, \\ \mathrm{Cov}(\omega_{l,t}, \omega_{l',t'}) &= \mathrm{Cov}(\omega^*_{l,t}, \omega^*_{l',t'}) \\ &= \begin{cases} \sigma^2_\omega & \text{if } l = l' \text{ and } t = t' \\ 0 & \text{if } l \neq l' \text{ or } t \neq t' \end{cases}, \\ \mathrm{Cov}(\omega_{l,t}, \omega^*_{l,t}) &= 0 \text{ for all } l \text{ and } t. \end{aligned} \qquad (3.5)$$

The irregular component ε_t contains the unexplained variation and is modelled as a white noise process:

$$E(\varepsilon_t) = 0,\ \mathrm{Cov}(\varepsilon_t, \varepsilon_{t'}) = \begin{cases} \sigma^2_\varepsilon & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases} \qquad (3.6)$$

3.2 Time series model for rotation group bias

The systematic differences between the trend and the seasonal components of the subsequent waves are modelled in (3.1) with $\boldsymbol{\lambda}_t$ and $\boldsymbol{\gamma}_t$. Additional restrictions for the elements of both vectors are required to identify model (3.1). Here it is assumed that an unbiased estimate for θ_t is obtained with the first wave, which is observed by CAPI, *i.e.*, Y^t_t. This implies that the first component of $\boldsymbol{\lambda}_t$ and $\boldsymbol{\gamma}_t$ equals zero. Now $\boldsymbol{\lambda}_t$ measures the systematic differences in the trend of the second, third, fourth and fifth wave with respect to the first wave. The components of $\boldsymbol{\lambda}_t$ are defined as:

$$\lambda^0_t = 0,\ \lambda^j_t = \lambda^j_{t-1} + \eta_{\lambda,j,t},\ j = 3, 6, 9, 12, \qquad (3.7)$$

$$\begin{aligned} E(\eta_{\lambda,j,t}) &= 0, \\ \mathrm{Cov}(\eta_{\lambda,j,t}, \eta_{\lambda,j',t'}) &= \begin{cases} \sigma^2_\lambda & \text{if } t = t' \text{ and } j = j' \\ 0 & \text{if } t \neq t' \text{ or } j \neq j'. \end{cases} \end{aligned}$$

Furthermore $\boldsymbol{\gamma}_t$ measures the systematic differences in the seasonal components with respect to the first wave. This implies that $\gamma^0_t = 0$. The other components of $\boldsymbol{\gamma}_t$ are defined as trigonometric functions, which are of the form of (3.5). The variance of the disturbances of the seasonal components are assumed to be equal for all waves and is denoted by σ^2_γ.

To borrow information across the panel waves, the RGB for the trend as well as the RGB for the seasonal components are modelled as time invariant components, *i.e.*, $\sigma^2_\lambda = \sigma^2_\gamma = 0$. As a kind of model diagnostic, the model initially allows for time dependent components. The maximum likelihood estimates for σ^2_λ and σ^2_γ are close to zero in this application. If this is not the case, it might be possible to allow for separate time independent RGB components for different time intervals.

3.3 Time series model for the survey errors

Finally a time series model for the survey errors in (3.1) is developed, which uses the direct estimates for the variance and AC's for the survey errors of the different panels as prior information. From (3.1) it follows that the survey errors for the first wave are defined as $e^t_t = Y^t_t - \theta_t$. For the second, third, fourth and fifth wave, they are defined as $e^{t-j}_t = Y^{t-j}_t - \theta_t - \lambda^j_t - \gamma^j_t$, for $j = 3, 6, 9, 12$.

Direct estimates for the variances of the survey errors for the separate panels are obtained with (2.2). These estimates are smoothed by modelling the variance estimates for the separate panels with a linear regression model $\mathrm{Var}(Y^{t-j}_t) = b^j_0 + b^j_1 (Y^{t-j}_t / n^{t-j}_t) + \text{error}$, where n^{t-j}_t denotes the sample size at time t of the sample that entered the panel at $t - j$.

The rotating panel design implies sample overlap with panels observed in the past. The sample of the first wave enters the panel for the first time at time t, so there is no sample overlap with panels observed in the past. Consequently, the survey errors of the first wave, e^t_t, are not correlated with survey errors in the past. The survey error of the second wave, *i.e.*, e^{t-3}_t, is correlated with the survey error of the first wave that entered the panel three months earlier, *i.e.*, e^{t-3}_{t-3}. In a similar way, the survey error of the third wave, *i.e.*, e^{t-6}_t, is correlated with e^{t-6}_{t-3} and e^{t-6}_{t-6}. The survey error of the fourth wave, *i.e.*, e^{t-9}_t, is correlated with e^{t-9}_{t-3}, e^{t-9}_{t-6} and e^{t-9}_{t-9}. Finally, the survey error of the fifth wave, *i.e.*, e^{t-12}_t, is correlated with e^{t-12}_{t-3}, e^{t-12}_{t-6}, e^{t-12}_{t-9} and e^{t-12}_{t-12}.

The AC's between the survey errors of the subsequent waves are estimated using the approach proposed by Pfeffermann *et al.* (1998). Since the real survey errors cannot be observed directly, this approach starts with calculating the autocovariances for the pseudo survey errors, which are defined as $(Y^{t-j}_t - \overline{Y}_t)$, where $\overline{Y}_t$ denotes the average of the five panel estimates Y^{t-j}_t at time t. The autocovariances of the pseudo survey errors for a separate wave are influenced by the autocovariances of the real survey errors of the other waves, since the pseudo survey

errors are defined as the deviation of a panel estimate with the average of all panel estimates obtained at time t. Equation (4) of Pfeffermann *et al.* (1998) specifies the relation between the autocovariances of the pseudo survey errors and the real survey errors. From this equation, it follows that the autocovariances of the real survey errors can be derived from the autocovariances of the pseudo survey errors by $\boldsymbol{\varphi}_k = \mathbf{F}^{-1}\mathbf{C}_k$, with $\mathbf{C}_k$ a vector containing the five autocovariances of the pseudo survey errors at lag k, $\boldsymbol{\varphi}_k$ a vector containing the five autocovariances of the survey errors at lag k, and $\mathbf{F}$ a $M \times M$ dimensional matrix where the diagonal elements equal $(M-1/M)^2$ and the off-diagonal elements $(1/M)^2$. Here M denotes the number of waves of the panel design ($M = 5$ in this application). The AC's and the partial autocorrelations (PAC) of the survey errors of the subsequent waves are given in Table 3.1.

Table 3.1
Correlations and partial autocorrelations for the survey errors of the separate panels

wave		lag			
		1	2	3	4
1	AC	-0.029	0.264	0.022	0.230
	PAC	-0.029	0.263	0.038	0.175
2	AC	0.291	0.135	0.035	-0.250
	PAC	0.291	0.054	-0.020	-0.287
3	AC	0.240	0.120	0.087	0.219
	PAC	0.240	0.066	0.047	0.194
4	AC	0.442	0.253	0.122	0.156
	PAC	0.442	0.072	-0.016	0.115
5	AC	0.249	0.298	-0.183	0.127
	PAC	0.249	0.252	-0.344	0.218
Mean*	AC	0.306	0.224	-0.030	0.127
	PAC	0.306	0.144	-0.150	0.162

Underlined AC's and PAC's refer to waves with sample overlap
*): Means are based on the waves with sample overlap.

The standard errors of the estimated AC's equal $1/\sqrt{T}$, where T denotes the number of observations. This implies that correlations with an absolute value larger than 0.21 are significantly different from zero at a 5% significance level. The lags in Table 3.1 refer to three months periods, so lag one equals a time lag of three months, lag two a time lag of six months, *etc*.

The AC's in Table 3.1, which are based on overlapping samples, are underlined. The AC's for the overlapping samples are positive as might be expected. An exception is the AC at lag three for the fifth wave, which has a negative value. This correlation, however, is not significantly different from zero. The AC's for lag one of the overlapping samples are all significantly different from zero. For lag two, the AC's of the overlapping samples are significantly different from zero for the fourth and the fifth wave, but not for the third wave. The AC's that are based on non-overlapping samples are sometimes unexpectedly large, *e.g.*, lag two and four of the first wave and lag four of the third wave. The AC for lag four of the second wave, on the other hand, has a surprisingly large negative value.

Pfeffermann *et al.* (1998) also report large positive AC's for lags with non overlapping samples. In their case this can be explained since samples are replaced in small geographical regions. In the Dutch LFS sample replacement takes place at the national level. There is no good explanation why the AC's for the non overlapping samples are sometimes small and sometimes take significant positive as well as negative values. To obtain more stable estimates, the AC's are averaged over the waves which are based on overlapping samples. Thus the mean AC for lag one is the average of the AC for the second, third, fourth and fifth wave, *etc*. The values are reported in the last two rows of Table 3.1. The standard errors of the PAC's of order $p+1$ and higher for an AR(p) equal $1/\sqrt{T}$, Box and Jenkins (1970). This implies that the PAC's are not significantly different from zero for lags two and higher if an AR(1) model with a correlation coefficient of 0.306 is assumed to capture the AC of the survey errors for the second, third, fourth and fifth wave.

The direct estimates for the variance and covariance structure of the survey errors are combined in the time series model using the following general form of the survey error model $e_t^{t-j} = k_t^{t-j}\tilde{e}_t^{t-j}$ where $k_t^{t-j} = \sqrt{\mathrm{Var}(Y_t^{t-j})}$, see Binder and Dick (1990). This allows for non homogeneous variance in the survey errors, that arise *e.g.*, due to the gradually decreasing sample size over the last decade.

Since the first wave is uncorrelated with survey errors obtained in the past, it is assumed that $\tilde{e}_t^t$ is white noise with $E(\tilde{e}_t^t) = 0$ and $\mathrm{Var}(\tilde{e}_t^t) = 1$. As a result, the variance of the survey error equals $\mathrm{Var}(e_t^t) = (k_t^t)^2$, which is equal to the direct estimate of the variance of the GREG estimate for the first wave. For the second, third, fourth and fifth wave, it is assumed that $\tilde{e}_t^{t-j} = \rho\tilde{e}_{t-3}^{t-j} + \nu_t^{t-j}$, with $\rho = 0.306$, and

$$E(\nu_t^{t-j}) = 0,\ \mathrm{Cov}(\nu_t^{t-j},\ \nu_{t'}^{t'-j}) = \begin{cases} \sigma_\nu^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases}$$

Since $\tilde{e}_t^{t-j}$ is an AR(1) process, $\mathrm{Var}(\tilde{e}_t^{t-j}) = \sigma_\nu^2/(1-\rho^2)$. To enforce that $\mathrm{Var}(e_t^{t-j})$ equals the direct estimate for the variance of the GREG estimate, it follows that $\sigma_\nu^2 = (1-\rho^2)$.

3.4 Final time series model for the monthly unemployment rate

The time series model for the vector with GREG estimates $\mathbf{Y}_t$ is obtained by inserting the different components developed in Sections 3.1 through 3.3 into (3.1). This model

uses the five monthly GREG estimates as input data to obtain model-based estimates for the monthly unemployment rate. The component for the population parameter θ_t in (3.2), developed in Section 3.1, takes advantages of sample information observed in the past to improve the precision of the estimated monthly unemployment rate. The components for the RGB, developed in Section 3.2, account for the systematic differences between the five monthly GREG estimates to avoid that the estimated monthly unemployment rate is incurred with this bias. The component for the survey errors, developed in Section 3.3, accounts for the AC between the five GREG estimates that are based on the same sample, observed with quarterly intervals. Although this approach is model-based, it accounts for the complexity of the survey design of the LFS, since the GREG estimates are used as input data.

4. State space representation

The time series model for the five monthly GREG estimates developed in Section 3 can be expressed in the state space representation, see Harvey (1989) or Durbin and Koopman (2001). A state space model consists of a measurement equation and a transition equation. The measurement equation, which is sometimes also called the signal equation, specifies how the observations depend on a linear combination of the state vector that contains the unobserved state variables for the trend, seasonal, RGB and the survey errors. The transition equation, which is sometimes also referred to as the system equation, specifies how the state vector evolves in time. The state space representation of the model developed in Section 3 is given by Van den Brakel and Krieg (2009).

Under the assumption of normally distributed error terms, the Kalman filter can be applied to obtain optimal estimates for the state vector. Estimates for state variables for period t based on the information available up to and including period t are referred to as the filtered estimates. The filtered estimates of past state vectors can be updated, if new data become available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the completely observed time series. So the smoothed estimate for the state vector for period t also accounts for the information made available after time period t. In this paper, the Kalman filter estimates for the state variables are smoothed with the fixed interval smoother. See Harvey (1989), and Durbin and Koopman (2001) for technical details.

The analysis is conducted with software developed in Ox in combination with the subroutines of SsfPack 3.0, see Doornik (1998) and Koopman, Shephard and Doornik (2008). All state variables are non-stationary with the exception of the survey errors. The non-stationary variables are initialised with a diffuse prior, *i.e.*, the expectation of the initial states are equal to zero and the initial covariance matrix of the states is diagonal with large diagonal elements. The survey errors are stationary and therefore initialised with a proper prior. The initial values for the survey errors are equal to zero and the covariance matrix is available from the model developed for the survey errors in Section 3.3. In Ssfpack 3.0 an exact diffuse log-likelihood function is obtained with the procedure proposed by Koopman (1997).

5. Results

5.1 Preliminary analyses

With the GREG estimator monthly estimates for the unemployment rate are obtained for each wave as described in Section 2.4. In Figure 5.1 the unemployment rate based on the CAPI wave is compared with the average of the four CATI waves. The graph shows that the unemployment rate observed with the first wave is systematically higher than for the other four waves.

The five time series obtained with the different waves are modelled with the time series model proposed in Sections 3 and 4. Preliminary analyses indicate that the estimates for the RGB of the seasonal effects in the second wave are not significantly different from zero and the RGB for the seasonal effects of the third, fourth and fifth wave are not significantly different from each other. Therefore the model is simplified to one with a single RGB seasonal effect. See Van den Brakel and Krieg (2009) for the state space representation.

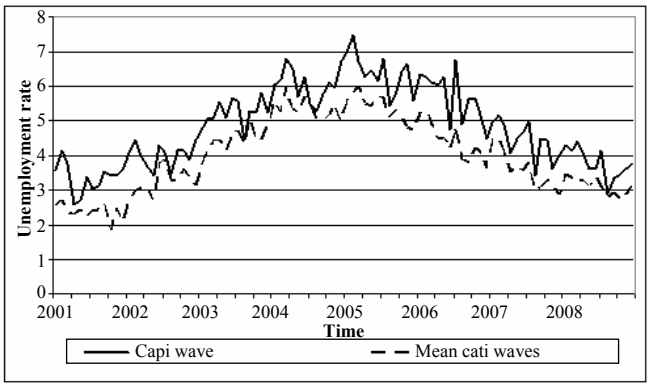

Figure 5.1 RGB monthly unemployment rate based on GREG estimates

5.2 Estimation results for the time series model

Maximum likelihood estimates for the hyperparameters, *i.e.*, the variance components of the stochastic processes for the state variables are obtained using a

numerical optimization procedure (BFGS algorithm, Doornik 1998). To avoid negative variance estimates, the log-transformed variances are estimated. The maximum likelihood estimates for the log-transformed variance of the level of the trend (σ_L^2), the seasonal component (σ_ω^2), the RGB of the trend (σ_λ^2) and the RGB of the seasonals tend (σ_λ^2) tend to large negative values with extremely large standard errors. These variance components are therefore put to zero in the final model. The estimation results for the remaining hyperparameters are presented in Table 5.1.

Table 5.1
Maximum likelihood estimates hyperparameters

Hyperparameter	Ln-transformed variance comp.		Variance components		
	Estimate	St. error	Estimate	95% conf. interval	
				Lower b.	Upper b.
Slope (σ_R^2)	-17.226	0.549	0.182E-3	0.106E-3	0.311E-3
Irregular comp. (σ_ε^2)	-13.480	0.482	1.183E-3	0.737E-3	1.897E-3

The smoothed Kalman filter estimates for the unemployment rate θ_t are given in Figure 5.2. These are the estimates for the monthly unemployment rate, based on the smooth trend model and a seasonal component, corrected for the RGB between the five GREG estimates. The local linear trend model simplified to a smooth trend model since $\sigma_L^2 = 0$. The trend component is time dependent since the maximum likelihood estimate of the hyperparameter for the slope is positive (see Table 5.1). The seasonal component is also time independent, since $\sigma_\omega^2 = 0$. Therefore the estimated seasonal effects obtained with the trigonometric form are exactly the same as the results obtained with the well known dummy variable seasonal model. The smoothed Kalman filter estimates for the trend and the seasonal component are plotted in Figures 5.3 and 5.4 respectively.

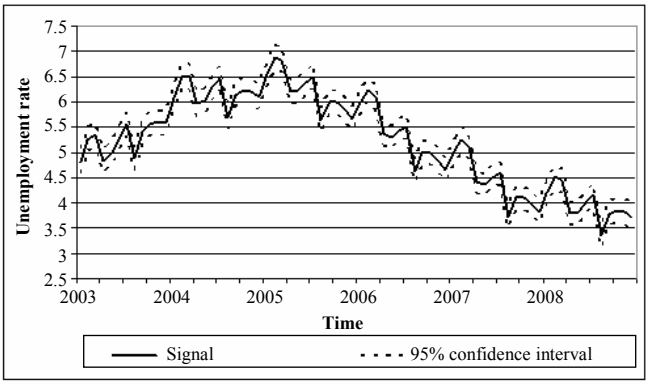

Figure 5.2 Smoothed Kalman filter estimates for the monthly unemployment rate

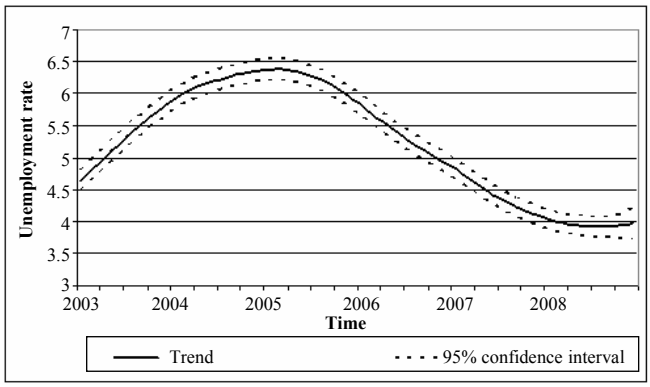

Figure 5.3 Smoothed Kalman filter estimates for the trend of the monthly unemployment rate

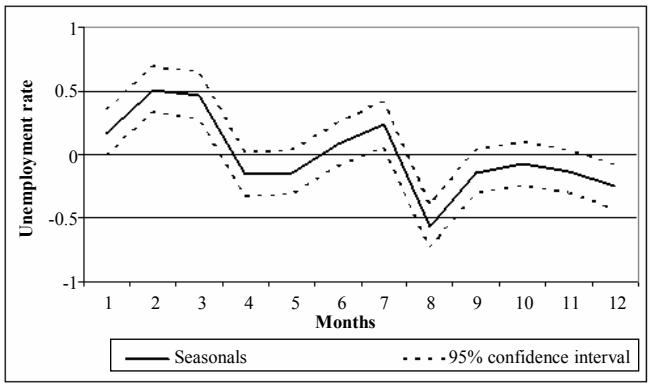

Figure 5.4 Smoothed Kalman filter estimates for the seasonal effect of the monthly unemployment rate

The Kalman filter estimates for the RGB of the trend are time independent. The smoothed Kalman filter estimates for the RGB are given in Table 5.2. The model beautifully detects a slightly increasing bias in the trend of the subsequent waves. The estimates for the RGB of the four CATI waves are significantly different from zero.

Table 5.2
Smoothed Kalman filter estimates RGB trend

Wave	RGB	St. error
2	-0.75	0.04
3	-0.86	0.04
4	-0.96	0.05
5	-1.10	0.05

An interesting empirical result of this application is the finding of the seasonality in the RGB. The Kalman filter estimates for the RGB of the seasonal effects are also time independent. Therefore, a sequence of likelihood ratio tests is conducted to reach the finally selected model and to test whether the seasonality effects in the RGB of this model are jointly significantly different form zero. Consider the following nested models:

- M1: separate and fixed RGB in the seasonality for wave two, three, four and five
- M2: equal to M1 where the RGB in the seasonality of wave two is equal to zero
- M3: equal to M2 with equal RGB in the seasonality of wave three, four and five
- M4: RGB in the seasonality of wave two, three, four and five is equal zero

The results of the likelihood ratio tests of this sequence of models are specified in Table 5.3.

Table 5.3
Likelihood-ratio tests for RGB in seasonality

Model	Log likelihood	Null hypothesis	Likh. ratio stat.	D.f.	p-value
M1	1,592.9				
M2	1,585.5	M2 = M1	14.7	11	0.19568
M3	1,573.7	M3 = M2	23.7	22	0.36422
M4	1,559.9	M4 = M3	27.6	11	0.00373

Testing the hypothesis that M2 equals M1 shows that the seasonality of the second wave is not significantly different from the first wave. Testing the hypothesis that M3 equals M2 shows that the RGB in the seasonality of the third, fourth and fifth wave are not significantly different. Testing the hypothesis that M4 equals M3 shows that the RGB of seasonal effects in last three waves are jointly significantly different from zero.

The smoothed Kalman filter estimates for the RGB of the seasonal effects for wave three, four and five are given in Figure 5.5. The smoothed Kalman filter estimates of the seasonal effects are compared with the smoothed estimates for the RGB of the seasonal effects in Figure 5.6.

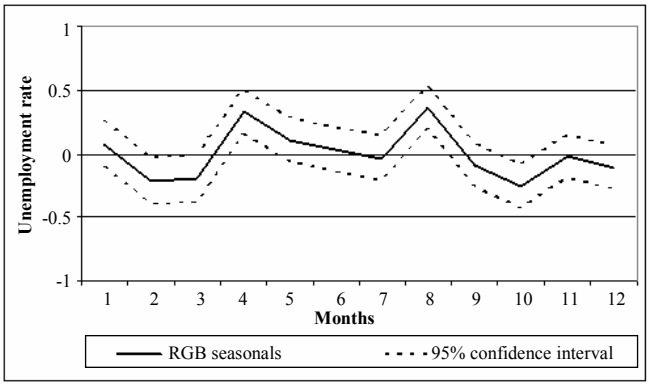

Figure 5.5 Smoothed Kalman filter estimates for the RGB of the seasonal effects in the third, fourth and fifth wave

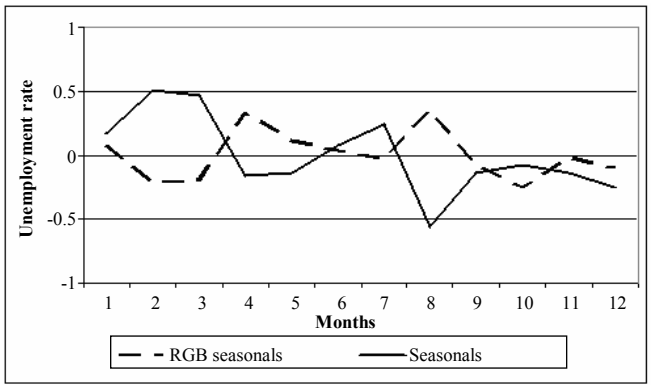

Figure 5.6 Comparison of smoothed Kalman filter estimates for the RGB of the seasonal effects in the third, fourth and fifth wave and the seasonal effects in 2008

It follows from Figure 5.5 that the seasonal effects in February, March, April, August and October in the third, fourth and fifth wave are significantly different from the first and the second wave. Figure 5.6 shows that the RGB in the seasonal effects largely nullifies the seasonal effects in these months. The seasonal effects in the last three waves are, apparently, less pronounced than in the first two waves. The different factors that contribute to the RGB in both the trend and the seasonal patterns are summarised in Section 2.2.

5.3 Comparison with GREG estimates

In this section the monthly GREG estimates for the unemployment rate and their standard errors are compared with the filtered model estimates. The filtered estimates are used since they are based on the complete set of information that would be available in the regular production process to produce a model-based estimate for the monthly unemployment rate for month t.

The GREG estimates based on the CAPI wave for the monthly unemployment rates are compared with the filtered model estimates in Figure 5.7. Some of the peaks and dips in the series of the GREG estimates are partially considered as survey errors under the structural time series model and flattened out in the filtered estimates for the series. Some of these peaks and dips are preserved since they are considered as seasonal effects under the time series model. It also follows that the filtered estimates are corrected for the RGB since the filtered series is at the same level as the series of the GREG estimates based on the CAPI wave. This is enforced with the assumption that the model parameters for the RGB for the first wave are zero (Section 3.2). This implies that the CATI waves are benchmarked to the outcomes of the first wave.

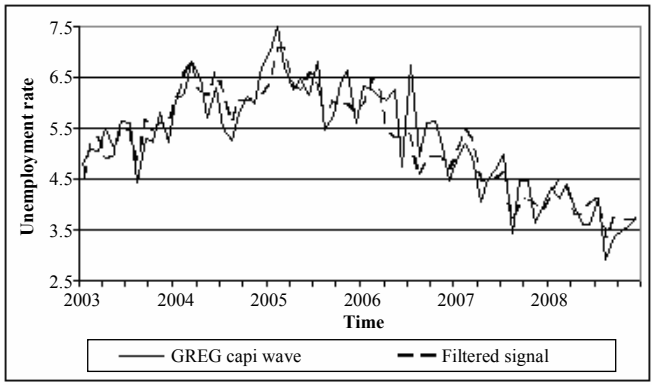

Figure 5.7 Filtered estimates and GREG estimates based on the CAPI wave for the monthly unemployment rate

The procedure applied in the regular estimation procedure of the LFS, to combine the CATI and the CAPI waves, is also used to estimate monthly unemployment figures. The GREG estimates for the monthly unemployment rates based on the five waves, using formula (2.4), are compared with the filtered estimates in Figure 5.8. Both estimates for the monthly unemployment rate follow the same level, since they are both benchmarked to the outcomes of the first wave. The GREG estimator is benchmarked in a rather rigid way using ratio (2.3), which is assumed to be constant in advance over a period of three years. The filtered estimates are benchmarked in a more subtle way through the explicit modelling of the trend and the seasonality in the RGB. The seasonality in the RGB indicates that the assumption of a constant RGB is not tenable. The monthly GREG estimates based on all waves are also compared with the GREG estimates based on the CAPI wave in Figure 5.9.

The ratio correction applied in formula (2.4) to the GREG estimates based on all waves removes the RGB in the trend, but does not correct for the RGB in the seasonal patterns. This follows from Figure 5.8 and 5.9. The series of the GREG estimates based on all waves follows the same level as the GREG estimates based on the CAPI wave (Figure 5.9). There are, however, subtle differences between the filtered estimates and the GREG estimates based on all waves (Figures 5.8). They partially arise because some of the dips and peaks in the GREG estimates are considered as survey errors by the time series model but they are also the result of systematic differences in the seasonal patterns between the subsequent waves. For example, the model estimates in February and March are larger in 2003, 2005 and 2006, and smaller in August in 2004, 2005 and 2006.

The standard errors of the monthly GREG estimates based on all waves, the CAPI wave and the filtered estimates are compared with each other in Figure 5.10. The standard errors for the GREG estimates are computed as described in Section 2.4. Standard errors of the filtered estimates are obtained by the standard recursion formulas of the Kalman filter, see Harvey (1989) or Durbin and Koopman (2001). The Kalman filter recursion assumes that the fitted state space model is the truth. As a result the standard errors for the filtered estimates do not reflect the additional variation induced by the use of likelihood estimates for the variance components in the state space model and are therefore too optimistic.

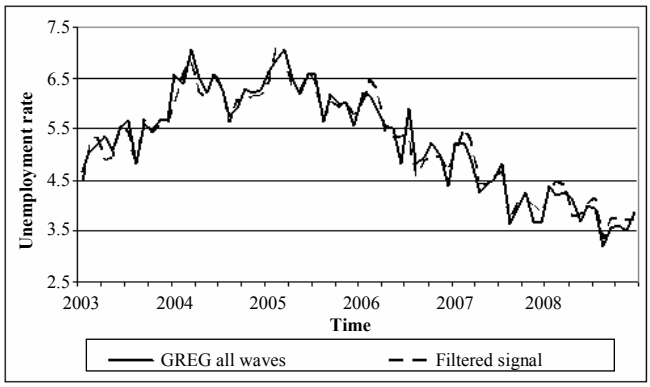

Figure 5.8 Filtered estimates and GREG estimates based on all waves for the monthly unemployment rate

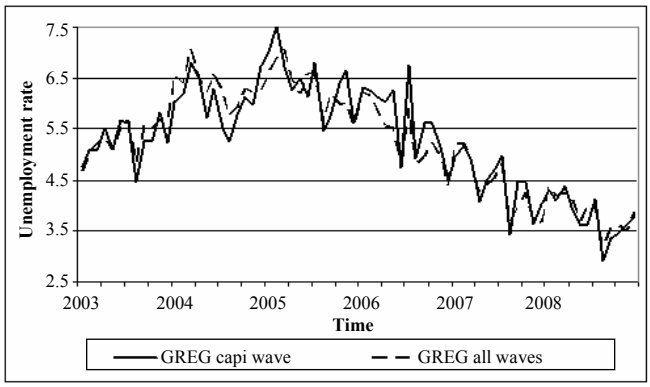

Figure 5.9 GREG estimates based on the CAPI wave and based on all waves for the monthly unemployment rate

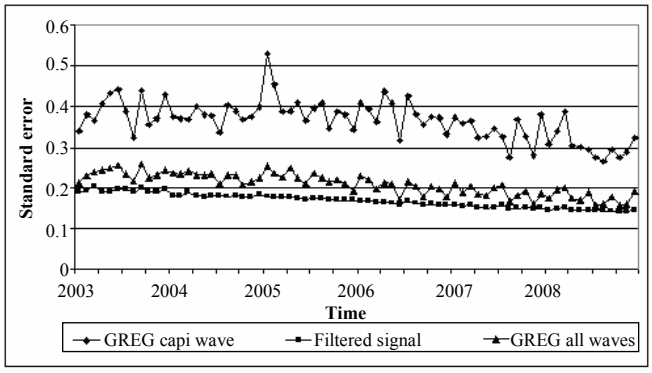

Figure 5.10 Standard errors of the GREG and filtered estimates for the monthly unemployment rate

As expected, the standard errors of the GREG estimates based on all waves are smaller than the standard errors of the GREG estimates based on the CAPI wave, since they are based on more data. The standard errors of the filtered estimates are smaller than the GREG estimates based on all waves, since the time series model uses additional sample information from preceding periods. The standard errors of the filtered estimates are slightly but continuously decreasing during the period 2003 to 2008.

The size and complexity of the applied time series model, is large compared to the length of the series available to fit the model. The final model that is applied to a five dimensional series which is monthly observed during a period of eight years contains 41 state variables. Therefore it is worthwhile to consider more parsimonious models, which might reduce the standard errors of the filtered estimates. Furthermore, the GREG estimate contains a bias since the RGB contains a seasonal effect, which is not reflected by its standard error. Therefore, the efficiency obtained by borrowing sample information from the past by relying on a time series model is illustrated more clearly if the standard error of the GREG estimates using all waves is compared with the standard error obtained with a time series model that accounts for the RGB in the trend only. Therefore a time series model without a component for the RGB in the seasonal pattern is applied to the data to illustrate the variance reduction by borrowing strength over time. The filtered estimates for the monthly unemployment rates based on a model with and without a component for the RGB in the seasonal pattern are compared in Figure 5.11.

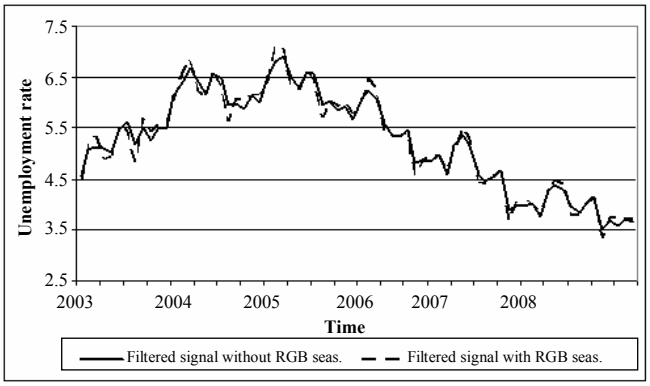

Figure 5.11 Filtered estimates of the monthly unemployment rate for two different time series models

The model without a component for the RGB of the seasonal effects assumes a seasonal effect for the population parameter θ_t that is based on an average of the seasonal effects of the five waves. The absolute values of the seasonal effects in February, March, and August are smaller under the simplified model, resulting in a lower estimate for the monthly unemployment rate in February and March and a larger estimate in August. This results in a more pronounced seasonal pattern in the filtered series obtained with the complete model.

The standard errors of the filtered estimates obtained with the two time series models and the standard errors of the GREG estimates using all waves are compared in Figure 5.12. The standard error of the filtered estimates of the simplified time series model is substantially smaller than the standard error of the GREG estimates using all waves. The simplification of the time series model by ignoring the RGB for the seasonal effects, results in a further reduction of the standard error at the cost of an increased bias in the seasonal effects. Under the model assumption that the estimates based on the first wave are unbiased, the time series model that accounts for the RGB in the seasonal effects is preferred, since it removes the bias in the seasonal pattern.

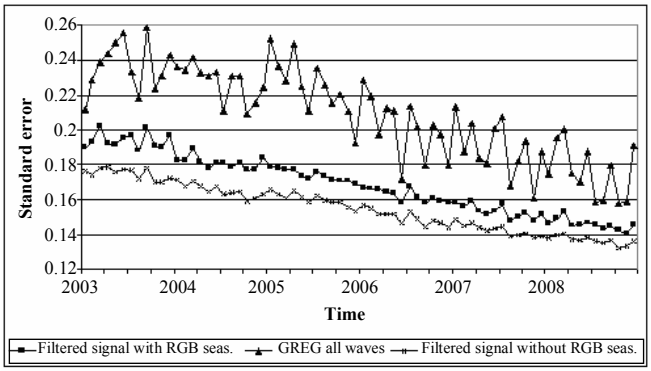

Figure 5.12 Standard errors of the GREG estimates based on all waves and filtered estimates for two different time series models for the monthly unemployment rate

Discussion

In this paper a multivariate structural time series model is applied to the monthly data of the LFS that accounts for the rotating panel design of this survey. This approach is initially proposed by Pfeffermann (1991) and extended in this paper with a component that models systematic differences in the seasonal effects between the subsequent waves. Compared with the GREG estimator, which is currently applied in the regular LFS, the time series model results in a substantial increase of the accuracy of the estimates of the unemployment rate. Firstly, the model explicitly estimates the RGB in the trend and the seasonal patterns between the first CAPI wave and the four subsequent CATI waves. Secondly, the time series model borrows strength from data observed in preceding periods via the assumed model for the population parameter and the AC between the survey errors of the different panels.

The RGB induced by the rotating panel design is substantial. The bias in the trend results in an underestimation of the unemployment rate in the subsequent waves and its magnitude slightly decreases from -0.8 percent points in the second wave to -1.1 percent points in the fifth wave. The seasonal patterns of the first two waves and the last three waves are also significantly different, since the seasonal pattern in the last three waves is less pronounced.

A parsimonious time series model that accounts for the RGB in the trend but not for the RGB in the seasonal pattern, results in a further reduction of the standard error of the filtered estimates. This, however, results in a biased seasonal pattern in the monthly estimates of the unemployment rates. Since the standard errors of the filtered estimates obtained under this parsimonious model do not reflect this bias, a time series model that accounts for both the RGB in the trend and the seasonal pattern is preferred.

The time series model is identified by adopting a restriction for the RGB parameters which assumes that the first wave is observed without bias. This implies that the estimates based on the first wave are used to benchmark the subsequent waves. If this restriction is used, then an all out effort in each part of the statistical process is required to reduce possible bias in the first wave, *e.g.*, by using the most appropriate data collection mode, reducing nonresponse, optimizing the weighting scheme, *etc*. Based on external information about the bias in the different waves, the restrictions for the RGB components might be adjusted.

The time series approach explored in this paper is appropriate to produce model-based estimates for monthly unemployment figures. Statistics Netherlands, however, is generally rather reserved in the application of model-based estimation procedures for the production of official statistics. Model misspecification might result in severely biased estimates. This bias is not reflected in the standard errors of the Kalman filter estimates. Extensive model selection and evaluation is therefore required for each separate target variable. This hampers a straightforward application of such estimation techniques, since there is generally limited time available for the analysis phase of the regular production process of official releases.

There is, on the other hand, a case for having official series that are based on model-based procedures with appropriate methodology and quality descriptions for situations where direct estimators do not result in sufficiently reliable estimates. The RGB observed under the rotating panel design of the LFS clearly illustrates the existence of non-sampling errors such as measurement errors and panel attrition. Therefore the traditional concepts that observations obtained from sampling units are true fixed values observed without error and that the respondents can be considered as a representative probability sample from the target population, generally assumed in design-based sampling theory, are not tenable under such designs. The application of direct estimators in the case of measurement errors and selective panel attrition will result in severely biased estimates. In the regular estimation procedure a ratio correction is applied to the GREG estimates, which is based on the implicit model assumption that the bias is constant over a period of three years. The time series model applied in this paper can be used to produce estimates that are corrected for the bias introduced by these non-sampling errors in a more advanced way.

This estimation procedure is also applicable in situations where small sample sizes result in unacceptable large standard errors. Small sample sizes arise if official statistics are required for small domains or for short data collection periods like the monthly unemployment figures in the LFS. Most surveys conducted by national statistical institutes operate continuously in time and are based on cross-sectional or rotating panel designs. Consequently, estimation procedures based on time series models that use sample information observed in preceding periods are particularly interesting.

The time series model yields estimates for the trend and seasonal components of the population parameter. Seasonally adjusted parameter estimates and their estimation errors are therefore obtained as a by-product of this estimation procedure. Another major advantage is that this approach accounts for the AC in the survey errors due to the rotating panel design. Pfeffermann *et al*. (1998) show that ignoring these AC, for example with the Henderson filters in X-12-ARIMA (Findley, Monsell, Bell, Otto and Chen 1998), results in spurious trend estimates.

The model can be improved in several ways. Information about registered unemployment and related variables, available in the register of the Office for Employment and Income, can be used as auxiliary variables in the models. If longer series become available, an additional cyclic component might be required to capture economic fluctuations. Another possible improvement is detection and modelling of outliers. Furthermore the model needs to be extended to estimate monthly unemployment rates for different domains using sample information collected in the past as well as cross-sectional data from other small areas, using the approach proposed by Pfeffermann and Burck (1990) and Pfeffermann and Tiller (2006).

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors would like to thank Professor

D. Pfeffermann and Professor S.J. Koopman for their valuable advice during this project as well as the Associate Editor and the referees for giving constructive comments on earlier drafts of this paper.

References

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

Bell, W.R., and Hillmer, S.C. (1990). The time series approach to estimation of periodic surveys. *Survey Methodology*, 16, 195-215.

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

Binder, D.A., and Dick, J.P. (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

Binder, D.A., and Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 239-253.

Box, G.E.P., and Jenkins, G.W.M. (1970). *Time series analysis - forecasting and control*. San Francisco: Holden-Day.

Cantwell, P.J. (1990). Variance formulae for composite estimators in rotating designs. *Survey Methodology*, 16, 153-163.

Doornik, J.A. (1998). *Object-oriented matrix programming using Ox 2.0*. London: Timberlake Consultants Press.

Durbin, J., and Koopman, S.J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.

Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55, 182-199.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. and Chen, B.C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics,* 16, 127-176 (with Discussion).

Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian labour force survey. *Survey Methodology*, 27, 45-51.

Gambino, J., Kennedy, B. and Singh, M.P. (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27, 65-74.

Gurney, M., and Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section on Social Statistics*, American Statistical Association, 242-257.

Hansen, M.H., Hurwitz, W.N. and Meadow, W.G. (1953). *Sample survey methods and theory*, 2. New York: John Wiley & Sons, Inc.

Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

Harvey, A.C., and Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society*, Series A, 163, 303-339.

Holbrook, A.L., Green, M.C. and Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires. *Public Opinion Quarterly*, 67, 79-125.

Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Section on Social Statistics*, American Statistical Association, 300-303.

Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.

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

Kumar, S., and Lee, H. (1983). Evaluation of composite estimation for the Canadian labour force survey. *Survey Methodology*, 9, 178-201.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

Nieuwenbroek, N., and Boonstra, H.J. (2002). Bascula 4.0 reference manual, BPA nr: 279-02-TMO, Statistics Netherlands, Heerlen.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.

Pfeffermann, D., and Rubin-Bleuer, S. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 149-163.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

Pfeffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16, 339-348.

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.

Rao, J.N.K., and Graham, J.E. (1964). Rotating designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

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

Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. Review paper, NCRM/008, National Centre for Research Methods, City University London.

Särndal, C.-E., and Lundström, S. (2005). Estimation in surveys with nonresponse. New York: John Wiley & Sons, Inc.

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

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

Scott, A.J., and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

Scott, A.J., Smith, T.M.F. and Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.

Singh, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 120-129.

Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian labour force survey with a rotating panel design. *Survey Methodology*, 27, 33-44.

Tam, S.M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review*, 55, 63-73.

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch labour force survey. *Journal of the Royal Statistical Society*, Series A, 171, 581-613.

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. Research paper, Statistics Netherlands, Heerlen (http://www.cbs.nl/en-GB/menu/methoden/research/discussionpapers/archief/2009/default.htm?Languageswitch=on).

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 191-201
Statistics Canada, Catalogue No. 12-001-X

Estimates for small area compositions subjected to informative missing data

Li-Chun Zhang [1]

Abstract

Estimation of small area (or domain) compositions may suffer from informative missing data, if the probability of missing varies across the categories of interest as well as the small areas. We develop a double mixed modeling approach that combines a random effects mixed model for the underlying complete data with a random effects mixed model of the differential missing-data mechanism. The effect of sampling design can be incorporated through a quasi-likelihood sampling model. The associated conditional mean squared error of prediction is approximated in terms of a three-part decomposition, corresponding to a *naive* prediction variance, a positive correction that accounts for the hypothetical parameter estimation uncertainty based on the latent complete data, and another positive correction for the extra variation due to the missing data. We illustrate our approach with an application to the estimation of Municipality household compositions based on the Norwegian register household data, which suffer from informative under-registration of the dwelling identity number.

Key Words: Conditional MSE of prediction; EMPQL algorithm; Generalized SPREE; Not missing-at-random; Two-way contingency table.

1. Introduction

Small area (or domain) population counts cross-classified by various social-economic characteristics are increasingly demanded for fund allocation, regional planning and social-economic research. Purcell and Kish (1980) outlined the so-called "Structure preserving estimation" (SPREE), which operates by modifying the small area estimates in a way so that they vary from one area to another in accordance with the variation that exists in another known auxiliary table of the same dimension. Typically the auxiliary table is obtained from a previous census, or some administrative register containing similar information. Zhang and Chambers (2004) developed a generalized SPREE (GSPREE) approach. Both fixed effects and random effects mixed models were introduced, and the restricted log-linear model underlying SPREE was shown to be a special case. This provides means for reducing the potential bias of the traditional SPREE estimates. We refer to Ghosh, Natarajan, Stroud and Carlin (1998) and Longford (1999) for alternative hierarchical and empirical Bayes approaches to this type of data.

In this paper we extend the GSPREE approach to situations subjected to missing data. This can be useful in sample surveys where nonresponse is unavoidable. We concentrate on *small area compositions* that can be arranged in a two-way table, where one of the two dimensions refers to the small areas and the other refers to the categories of interest. The cell counts summarize to a fixed area total that may or may not be known. For instance, each person between 16 and 74 years of age can be classified according to the labour force status "employed", "unemployed" and "not in the labour force". The sum of the three counts inside a small area is the total number of persons between 16 and 74 years of age within this area.

In the context of small area composition we say that the missing-data mechanism is *informative* provided it varies across the categories of interest. As such it is also *not* missing-at-random (Rubin 1976). In addition, the overall rate of missing differs across the areas. Differential missingness as such leads to distortion of the underlying complete data, and bias if the estimation is carried out as if the observed data were complete. We propose a double mixed modeling approach that combines the random effects mixed model for the underlying complete data with a random effects mixed model of the missing-data mechanism. The double-smoothing approach is outlined in Section 2.

It should be noted that national statistical offices that conduct large scale surveys will have accounted for missing data by weighting adjustments or imputation. This, however, will have been done at levels that are significantly higher than the small areas, and will be for variables that do not necessarily correspond to those of interest for the small areas. When available, the adjusted totals can be incorporated into the GSPREE as marginal totals for iterative proportional fitting (IPF). But modeling of the differential probabilities of missing across the small areas will generally remain a matter of interest.

It should also be noticed that informative missing data as such makes it less straightforward to assess the potential bias of any estimation approach. SPREE may be biased on two accounts: (i) the underlying restricted log-linear assumptions are likely to be unrealistic, (ii) direct IPF may fail to account for the differential probabilities of missing

1. Li-Chun Zhang, Statistics Norway, Kongensgate 6, PB 8131 Dep. N-0033 Oslo, Norway. E-mail: lcz@ssb.no.

adequately. The proposed double mixed modeling approach deals with problem (i) by GSPREE modeling of the underlying complete data, and it deals with problem (ii) by introducing a more flexible missing-data model, as we shall discuss in Section 2.2. Nevertheless, bias is likely to persist to a certain extent. Since the estimation of model parameters and random effects is more complicated under the double mixed modeling approach, alternative estimation methods that are able to preserve the computational simplicity of SPREE, while making more adequate adjustment for informative missing data, are worth investigating in future.

When it comes to the assessment of estimation uncertainty, Booth and Hobert (1998) argued for the conditional mean squared error of prediction (CMSEP) given the observed data. We extend their approach and derive approximate CMSEP in the current multivariate incomplete-data situation. This results in a three-part decomposition of the CMSEP, corresponding to a *naive* prediction variance, a positive correction that accounts for the hypothetical estimation uncertainty of the parameters based on the latent complete data, and another positive correction for the extra variation due to the missing data. The details are given in Section 3.

Estimation procedures for the parameters, the CMSEP and the small area compositions are described in Section 4. In Section 5 we apply our approach to derive estimates of the Municipality household compositions based on the Norwegian household register, which suffers from informative under-registration of the dwelling identity number (DIN). A summary is given in Section 6.

2. Double mixed modeling

2.1 Random effects mixed model in the complete-data case

2.1.1 Models for finite population

The small area counts can be arranged in a two-way contingency table, denoted by $\mathbf{X} = \{X_{ak}\}$, where $a = 1, \ldots, A$ indexes the small areas and $k = 1, \ldots, K$ the categories of interest. The interest of estimation is the within-area proportions given by

$$\theta_{ak}^X = X_{ak} / X_{a.} = X_{ak} \Big/ \sum_{j=1}^{K} X_{aj}$$

referred to as compositions since $\sum_k \theta_{ak}^X = 1$. Typically under the GSPREE approach we assume that the marginal totals $\{X_{a.}\}$ and $\{X_{.k}\}$, also known as the allocation structure, are either known or can be reliably estimated, in which case estimating $\{\theta_{ak}^X\}$ is equivalent to estimating $\{X_{ak}\}$. For simplicity we then make no distinction between counts and compositions in the exposition. Otherwise, without the allocation structure, one can still use our approach to estimate $\{\theta_{ak}^X\}$ but not $\{X_{ak}\}$.

Assume that we have available an auxiliary table of the same dimension, denoted by $\mathbf{X}^0 = \{X_{ak}^0\}$, and the corresponding within-area proportions $\{\theta_{ak}^0\}$. To model $\theta_a^X = (\theta_{a1}^X, \ldots, \theta_{aK}^X)^T$ we use the *multinomial standardized-log (mslog)* link function, given by

$$\mu_{ak}^X = \log \theta_{ak}^X - K^{-1} \sum_{j=1}^{K} \log \theta_{aj}^X \tag{1}$$

and similarly for μ_{ak}^0 and θ_{ak}^0. Zhang and Chambers (2004) introduced the following generalized linear structural mixed model (GLSMM)

$$\mu_{ak}^X = \lambda_k + \beta \mu_{ak}^0 + v_{ak}$$

where

$$\sum_{k=1}^{K} \lambda_k = 0 \quad \text{and} \quad \sum_{k=1}^{K} v_{ak} = 0 \tag{2}$$

and $\mathbf{v}_{a(1)} = (v_{a2}, \ldots, v_{aK})^T$ assumes a multivariate normal distribution with covariance matrix $G = G(\delta)$, where δ contains the variance parameters. Notice that there is no area-specific term in (2) because $\sum_k \mu_{ak} = \sum_k \mu_{ak}^0 = 0$. The term "structural" refers to the fact that this is a model of the finite-population parameters $\{\theta_{ak}^X\}$ directly, although the emphasis is not common in the small area estimation literature. For instance, the well-known Fay-Herriot model (Fay and Herriot 1979) is "structural" in the same sense.

There is an important interpretation of the model (2) in terms of the log-linear interactions of $\{\theta_{ak}\}$ due to the choice of the link function (1), *i.e.*,

$$\mu_{ak}^X = \alpha_k + \alpha_{ak}^X \tag{3}$$

where by the standard theory of log-linear models (*e.g.*, Agresti 2002), we have

$$\log X_{ak} = \log X_{a.} + \log \theta_{ak}^X = \alpha_0^X + \alpha_a^X + \alpha_k^X + \alpha_{ak}^X$$

for $\alpha_0^X = (AK)^{-1} \sum_{a,k} \log X_{ak}$, and $\alpha_a^X = K^{-1} \sum_k \log X_{ak} - \alpha_0^X$, and $\alpha_k^X = A^{-1} \sum_a \log X_{ak} - \alpha_0^X$, and $\alpha_{ak}^X = \log X_{ak} - \alpha_a^X - \alpha_k^X - \alpha_0^X$, such that $\sum_a \alpha_a^X = \sum_k \alpha_k^X = \sum_a \alpha_{ak}^X = \sum_k \alpha_{ak}^X = 0$. We refer to (3) as the log-linear identity, and we refer to the log-linear parameters α_{ak}^X as the (first-order) interactions of the compositions θ_{ak}^X as well as the counts X_{ak}. Similar identity holds for μ_{ak}^0. Zhang and Chambers (2004) showed that the GLSMM is equivalent to the following *proportional interactions mixed model* (*PIMM*)

$$\alpha_{ak}^X = \beta \alpha_{ak}^0 + v_{ak} + O_p(A^{-1/2}). \tag{4}$$

The parameters λ_k's in (2) do not entail any model restriction beyond the PIMM, and they do not affect the

interactions. The parameter β is called the proportionality coefficient. Clearly, SPREE based directly on the association structure $\{X^0_{ak}\}$ amounts to setting $\beta \equiv 1$ and $v_{ak} \equiv 0$. We therefore refer to the model (2) as a GSPREE model, which contain both fixed and random effects extensions of the SPREE model.

2.1.2 Model for sample

To complete the model specification we assume sample classifications $\mathbf{x} = \{x_{ak}\}$. Let

$$\mathbf{t}_a = (t_{a1}, \ldots, t_{aK})^T = (t_1(\mathbf{x}_a), \ldots, t_K(\mathbf{x}_a))^T$$

be such that $E(t_{ak} | \mathbf{v}) = E(t_{ak} | \mathbf{X}) = \theta^X_{ak}$, where $\mathbf{v} = \{v_{ak}\}$. The expectation is typically with respect to the sampling design. However, it can also be taken under a suitable model of the sampling distribution, such as a multinomial model for $\mathbf{x}_a$ provided simple random sampling within each area. We therefore make no distinction in the notation.

We assume that $\mathbf{t}_a$ is independent of $\mathbf{t}_{a'}$ for $a \neq a'$, and put

$$V(t_{ak}) = \nu_1 \omega_k(\mathbf{X}_a) \quad \text{and} \quad \text{Cov}(t_{ak}, t_{aj}) = \nu_1 \omega_{kj}(\mathbf{X}_a) \quad (5)$$

where $\omega_k(\cdot)$ and $\omega_{kj}(\cdot)$ are specified variance and covariance functions, and ν_1 is the dispersion parameter that may or may not be known. This is essentially the quasi-likelihood set-up for dependent data (McCullagh and Nelder 1989). The dependence on $\mathbf{X}_a$ allows us to incorporate the sampling design effect, in which case the expectations in (5) may be evaluated with respect to the sampling distribution. This is an important reason why we do not directly assume that the distribution of $\mathbf{t}_a$ belongs to the exponential family, as *e.g.*, in the generalized linear mixed models (Breslow and Clayton 1993).

2.1.3 Parameter estimation

Zhang and Chambers (2004) outline an iterative weighted least square (IWLS) algorithm for the GLSMM (2), which is a variation of the PQL approach (Schall 1991; Breslow and Clayton 1993). Let $\mu_a = (\mu^X_{a1}, \ldots, \mu^X_{aK})^T$. The GLSMM (2) can formally be given by

$$\mu_a = g(\theta_a) = H_a \zeta + B \mathbf{v}_{a(1)}$$

where $g(\theta_a)$ is the mslog link function, and $\zeta = (\lambda_2, \ldots, \lambda_K, \beta)^T$, and $\mathbf{v}_{a(1)} = (v_{a2}, \ldots, v_{aK})^T$. The $K \times K$ design matrix H_a and $K \times (K-1)$ design matrix B are, respectively,

$$H_a = [B_{K \times K-1} \quad \mu^0_a] \quad \text{and} \quad B = \begin{pmatrix} -\mathbf{1}^T_{K-1} \\ I_{K-1 \times K-1} \end{pmatrix}$$

where $\mathbf{1}$ is a vector of 1 and I is an identity matrix. Define the working variables

$$\mathbf{z}_a \overset{\text{def.}}{=} \mu_a + \mathbf{e}_a = H_a \zeta + B \mathbf{v}_a + \mathbf{e}_a \text{ and } \mathbf{e}_a = Q(\mathbf{t}_a - \theta^X_a) \quad (6)$$

where $Q = \partial \mu^X_a / \partial \theta^X_a$ is the Jacobian matrix of partial derivatives. Denote by R_a the conditional covariance matrix of $\mathbf{t}_a$ given θ^X_a defined by (5). Under the PQL approach we assume that $\mathbf{e}_a$ has an approximate multivariate normal distribution with covariance matrix QR_aQ^T, and apply standard methods for linear mixed models (LMM) to the linearized data (6). Variants of the PQL approach differ in the estimation of the variance parameters δ. The details are omitted here.

2.1.4 On model hierarchy

The GLSMM (2) is specified at the finite population level. More generally, we may consider the finite population $\{X_{ak}\}$ to be randomly generated from an infinite super-population. Let θ_{ak} be the within-area probability that a unit of the super-population belongs to the cell (a, k), where $\sum_k \theta_{ak} = 1$. Conditional on $X_{a.} = \sum_k X_{ak}$, the within-area counts $(X_{a1}, \ldots, X_{aK})^T$ follow the multinomial distribution with parameters $(\theta_{a1}, \ldots, \theta_{aK})^T$. A *multinomial standardized-log mixed model* (*MSLMM*) of $\{\theta_{ak}\}$ is given by

$$\mu_{ak} = \lambda_k + \beta \mu^0_{ak} + v_{ak} \quad (7)$$

where

$$\sum_{k=1}^K \lambda_k = 0 \quad \text{and} \quad \sum_{k=1}^K v_{ak} = 0$$

where μ_{ak} is given by θ_a through the mslog link function.

Unlike the GLSMM (2), the equation (7) defines a regression model. There are then three different hierarchy one may choose from in the sample survey situation:

1. Assume the GLSMM (2) for the finite population and the quasi-likelihood model (5) for the sample, yielding the GSPREE approach of Zhang and Chambers (2004).

2. Assume the MSLMM (7) for the super-population and model sample data $\mathbf{t}_a$ based on θ_a directly, yielding a purely model-based two-level approach.

3. Assume the MSLMM (7) for the super-population, and assume that the finite population totals $\mathbf{X}_a$ follow the multinomial distribution given θ_a, and assume the quasi-likelihood model (5) given $\mathbf{X}_a$, yielding a general three-level model.

Provided the finite population is large, it makes little difference in practice to adopt the GSPREE approach, in

which case one does not have to deal explicitly with one extra level of hierarchy. But the distinction between (2) and (7) becomes necessary if the areas are so small that the stochastic variation in $\mathbf{X}_a$ is not negligible compared to the sampling variation in $\mathbf{x}_a$ (or $\mathbf{t}_a$). In our application later, we have register data that would have given us the interested population counts $\{X_{ak}\}$ had they not suffered from missing data. And the small area level of aggregation is so detailed that the stochastic variation in $\mathbf{X}_a$ can not be ignored. We therefore adapt the GSPREE approach by (a) adopting the MSLMM (7) instead of the GLSMM (2), and (b) modeling $\mathbf{X}_a$ as a 'sample', albeit a very large one, from the super-population directly.

2.2 A random effects mixed model of missing data

Missing data add another level of stochastic variation on top of the underlying complete data. In the exposition below, we consider the sample counts $\{x_{ak}\}$ as the complete data, which is the most common situation in practice. Our application later in Section 5 can be viewed as a special case where $\mathbf{X} = \mathbf{x}$.

Denote by $\mathbf{y}_a = (y_{a1}, \ldots, y_{aK})^T$ the observed cell counts, for $a = 1, \ldots, A$. Suppose that, conditional on x_{ak} and a random effect b_a,

$$E(y_{ak} \mid x_{ak}, b_a) = x_{ak} p_{ak}$$

and (8)

$$V(y_{ak} \mid x_{ak}, b_a) = \nu_2 c_{ak} p_{ak}(1 - p_{ak})$$

where c_{ak} is a known constant, and ν_2 is the dispersion parameter. We assume that y_{ak} is independent of y_{aj} for $k \neq j$, *i.e.*, missing data are independent from one cell to another. Let the units in the complete sample cell (a, k) be indexed by $i = 1, \ldots, n_{ak}$. Let $r_{i,ak} = 1$ if the i^{th} unit is observed, and $r_{i,ak} = 0$ if it is missing. The parameter p_{ak} is the assumed probability of $r_{i,ak} = 1$ inside cell (a, k). To see this, let $x_{i,ak}$ be the contribution of the i^{th} unit to x_{ak}, *i.e.*, $x_{ak} = \sum_{i=1}^{n_{ak}} x_{i,ak}$, such that $y_{ak} = \sum_{i=1}^{n_{ak}} r_{i,ak} x_{i,ak}$ and

$$E(y_{ak} \mid x_{1,ak}, \ldots, x_{n_{ak},ak}, b_a) = \sum_{i=1}^{n_{ak}} x_{i,ak} E(r_{i,ak} \mid b_a) = \sum_{i=1}^{n_{ak}} x_{i,ak} P(r_{i,ak} = 1 \mid b_a) = x_{ak} p_{ak}.$$

Notice that p_{ak} does not depend on the value of $x_{i,ak}$, but only the position of the unit in the two-way table. We assume that p_{ak} depends on b_a through the logistic link function given by

$$\eta_{ak} = \log(p_{ak}/(1 - p_{ak})) = \xi_k + b_a$$

where (9)

$$b_a \sim N(0, \sigma^2).$$

The fixed effects ξ_k's allow the probability of missing to depend on the categories of interest, the area-level random effect b_a allows it to vary across the areas in addition.

Obviously, under the assumptions (8) and (9), the missing data cause bias in the estimates of the λ_k's, if the observed table $\mathbf{y}$ is treated as if it were complete. Moreover, it distorts the estimation of the first-order interactions $\{\alpha_{ak}^X\}$. We have,

$$\log p_{ak} = (\xi_k + b_a) - \gamma_{ak} \quad \text{where } \gamma_{ak} = \log(1 + \exp(\xi_k + b_a)).$$

The first-order interactions of $\{p_{ak}\}$ are then given by $\alpha_{ak}^p = -\tilde{\gamma}_{ak} = -(\gamma_{ak} - \overline{\gamma}_{a.} - \overline{\gamma}_{.k} + \overline{\gamma}_{..})$, for the row and column means $\overline{\gamma}_{a.}$ and $\overline{\gamma}_{.k}$ and the overall mean $\overline{\gamma}_{..}$. These are non-zero unless $\xi_k = \xi$. By (8) the interactions of the expected observed table are given by

$$\alpha_{ak}^{E(\mathbf{y}|\mathbf{x},\mathbf{b})} = \alpha_{ak}^x + \alpha_{ak}^p = \alpha_{ak}^x - \tilde{\gamma}_{ak} \neq \alpha_{ak}^x$$

such that the estimates of $\{\alpha_{ak}^X\}$ will be biased if $\mathbf{y}$ is treated as $\mathbf{x}$.

It is worth noting that, as far as the estimation of the interactions is concerned, it is in principle possible to treat the observed table $\mathbf{y}$ as if it were the complete table $\mathbf{x}$ under a particular missing-data model given by

$$\log p_{ak} = \xi'_k + b'_a. \tag{10}$$

This is because the first-order interactions of $\{p_{ak}\}$ are all zero under (10), in which case we have $\alpha_{ak}^{E(\mathbf{y}|\mathbf{x})} = \alpha_{ak}^{\mathbf{x}}$. Disregarding the range restrictions, the assumption (10) defines an informative missing-data mechanism where the probability of missing varies across the categories of interest, while the area effect modifies all the within-area probabilities of missing by a factor $\exp(b'_a)$, such that $p_{ak}/\sum_{j=1}^K p_{aj} = \exp(\xi'_k)/\sum_j \exp(\xi'_j)$ remains constant. The model (9), however, is more flexible since it allows the random effects to affect the interactions. Both (9) and (10) will be examined in Section 5.

Finally, we notice that allowing for component-wise random effects in the model (9) may cause identification problems. For instance, assume simple random sampling from the finite population, in which case the interactions of the expected complete table are given by $\alpha_{ak}^{E(\mathbf{x}|\mathbf{X})} = \alpha_{ak}^X$. With component-wise b_{ak} in the model (9) we have $\log p_{ak} = \xi_k + b_{ak} + \gamma_{ak}$, where $\gamma_{ak} = \log(1 + \exp(\xi_k + b_{ak}))$. It follows from (4) and (8) that the interactions of the expected table $E(\mathbf{y} \mid \mathbf{x}, \mathbf{b})$ is given by $\beta\alpha_{ak}^0 + v_{ak} + b_{ak} - \tilde{\gamma}_{ak}$. But there is no information in the observed data to distinguish between the two random effects v_{ak} and b_{ak}.

3. Conditional mean squared errors of prediction

We adopt the approach of Booth and Hobert (1998) and use the CMSEP as a measure of the uncertainty in prediction. Like them we consider the CMSEP on the linear-predictor scale. In vector form the μ_{ak}'s in (1) belong to the following class of linear functions

$$\mu_a = H_a\zeta + B_a\mathbf{v}_a \tag{11}$$

where μ_a is the area-specific vector of linear predictors, and ζ is the vector of fixed effects, and $\mathbf{v}_a$ is the vector of area-specific random effects, and H_a and B_a are the corresponding design matrices. All the quantities have been specified in (6) for the GLSMM (2), where we actually have $B_a = B$. But we shall adopt the slightly more general formulation (11) in the following. Let $\hat{\zeta}$ and $\hat{\mathbf{v}}_a$ be, respectively, the estimates of ζ and $\mathbf{v}_a$ based on observations subjected to missing data, denoted by $\mathbf{y}_a$ for $a = 1, \ldots, A$. The CMSEP of $\hat{\mu}_a = H_a\hat{\zeta} + B_a\hat{\mathbf{v}}_a$ is defined as

$$\text{CMSEP}_a = E\{(\hat{\mu}_a - \mu_a)(\hat{\mu}_a - \mu_a)^T \mid \mathbf{y}_a\}.$$

We introduce first a decomposition through the hypothetical best predictor (BP) based on $\mathbf{x}_a$, given by $\dot{\mu}_a = E(\mu_a \mid \mathbf{x}_a, \zeta, \delta) = H_a\zeta + B_aE(\mathbf{v}_a \mid \mathbf{x}_a, \zeta, \delta)$, when the parameters are known. We have

$$\begin{aligned}\text{CMSEP}_a &= E\{E((\dot{\mu}_a - \mu_a)(\dot{\mu}_a - \mu_a)^T \mid \mathbf{x}_a) \mid \mathbf{y}_a\} \\ &\quad + E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{y}_a\} \\ &= E\{B_a\text{Cov}(\mathbf{v}_a, \mathbf{v}_a \mid \mathbf{x}_a)B_a^T \mid \mathbf{y}_a\} \\ &\quad + E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{y}_a\}\end{aligned}$$

because $\dot{\mu}_a - \mu_a$ and $\hat{\mu}_a - \dot{\mu}_a$ are conditionally independent of each other given $\mathbf{x}_a$: $\dot{\mu} - \mu$ depends on the random effects $\mathbf{v}_a$, whereas $\hat{\mu}_a - \dot{\mu}_a$ depends on random variations in the other areas. Next, for the second term on the right-hand side, we introduce a decomposition through the hypothetical estimated best predictor (EBP) based on the complete data $\mathbf{x}$, denoted by $\tilde{\mu}_a = H_a\tilde{\zeta} + B_a\tilde{v}_a$, where $(\tilde{\zeta}, \tilde{\delta})$ are the parameter estimates based on $\mathbf{x}$, and $\tilde{v}_a = E(\mathbf{v}_a \mid \mathbf{x}_a, \tilde{\zeta}, \tilde{\delta})$. We have

$$\begin{aligned}E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{y}_a\} &\approx E\{(\hat{\mu}_a - \dot{\mu}_a)(\hat{\mu}_a - \dot{\mu}_a)^T\} \\ &= E\{E((\tilde{\mu}_a - \dot{\mu}_a)(\tilde{\mu}_a - \dot{\mu}_a)^T \mid \mathbf{x})\} \\ &\quad + E\{E((\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T \mid \mathbf{x})\} \\ &= E\{(\tilde{\mu}_a - \dot{\mu}_a)(\tilde{\mu}_a - \dot{\mu}_a)^T\} \\ &\quad + E\{E((\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T \mid \mathbf{x})\}.\end{aligned}$$

The first approximation is correct to the order of $O_p(A^{-1})$, and can be justified as the number of areas tends to infinity. Intuitively, this makes sense if the information from any single area is asymptotically negligible compared to the information from all the other areas together. Next, the decomposition follows because $\tilde{\mu}_a - \dot{\mu}_a$ and $\hat{\mu}_a - \tilde{\mu}_a$ are independent of each other given $\mathbf{x}$: the former is a constant given $\mathbf{x}$.

In this way, we obtain an approximate CMSEP with a three-part decomposition

$$\text{CMSEP}_a \approx h_{1a}(\mathbf{x}_a; \zeta, \delta) + h_{2a}(\zeta, \delta) + h_{3a}(\mathbf{x}; \tilde{\zeta}, \tilde{\delta}, \psi)$$

where ψ contains the parameters of the conditional distribution of $\mathbf{y}_a$ given $\mathbf{x}_a$, and

$$h_{1a}(\mathbf{x}_a; \zeta, \delta) \overset{\text{def.}}{=} B_a\text{Cov}(\mathbf{v}_a, \mathbf{v}_a \mid \mathbf{x}_a)B_a^T \tag{12}$$

$$h_{2a}(\zeta, \delta) \overset{\text{def.}}{=} E\{(\tilde{\mu}_a - \dot{\mu}_a)(\tilde{\mu}_a - \dot{\mu}_a)^T\} \tag{13}$$

$$h_{3a}(\mathbf{x}; \tilde{\zeta}, \tilde{\delta}, \psi) \overset{\text{def.}}{=} E\{(\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T \mid \mathbf{x}\}. \tag{14}$$

The three *h*-terms correspond, respectively, to a conditional prediction variance due to the random effects, a positive correction that accounts for the uncertainty in the estimation of the parameters based on the latent complete data, *i.e.*, the sampling variation, and another positive correction for the extra variation due to the randomness in the missing data. Alternative approximations are possible. For instance, one might use $E\{B_a\text{Cov}(\mathbf{v}_a, \mathbf{v}_a \mid \mathbf{x}_a)B_a^T \mid \mathbf{y}_a\}$ instead of h_{1a}, or replace h_{3a} with the unconditional $E\{(\hat{\mu}_a - \tilde{\mu}_a)(\hat{\mu}_a - \tilde{\mu}_a)^T\}$. The expressions (12) - (14) are chosen because they produce a clean separation between the sampling variation in the complete data and the extra variation owing to the missingness given the complete data. The difference from the CMSEP in the complete-data case (Booth and Hobert 1998) comes down to the third term h_{3a}.

4. Estimation

4.1 Parameter estimation

The structure of the data suggests an iterative procedure similar to the EM algorithm (Dempster, Laird and Rubin 1977). Given the current values of the parameters and the random effects, we calculate at the E-step the conditional expected complete two-way table $E(\mathbf{x} \mid \mathbf{y}, \mathbf{m})$. At the M-step we estimate the two random effects mixed models separately by some maximum penalized quasi-likelihood (MPQL) procedures. Iterations between the two yield an EMPQL algorithm.

For the E-step, let $I_{i,ak}=1$ if the sample unit *i* belongs to the $(a,k)^{\text{th}}$ cell, and $I_{i,ak}=0$ otherwise. The value is observed provided $r_{i,ak}=1$, but is unknown if $r_{i,ak}=0$. Let θ_{ak} be the generic compositions, depending of the adopted model. Suppose that

$$P[I_{i,ak}=1 \mid i\in s]=d_{ak}\theta_{ak}$$

where *s* denotes the complete sample, and d_{ak} is some known constant which accounts for the eventual sampling design effect. For example, simple random sampling implies that $d_{ak}=1$ for all (a,k). An example of $d_{ak}\neq 1$ is when the sampling units are households, which are selected by a probability proportional to the household size. Let $m_{ak}=x_{ak}-y_{ak}=\sum_{i;r_{i,ak}=0}I_{i,ak}x_{i,ak}$. We have $E(x_{ak}\mid \mathbf{y}_a, m_{a.})=y_{ak}+E(m_{ak}\mid m_{a.})$, where

$$\begin{aligned}E(m_{ak}\mid m_{a.}) &= \sum_{i;r_{i,ak}=0}E(I_{i,ak}\mid r_{i,ak}=0)x_{i,ak}\\ &= m_{a.}P[I_{i,ak}=1\mid r_{i,ak}=0]\\ &= m_{a.}(1-p_{ak})d_{ak}\theta_{ak}\Big/\Big\{\sum_j(1-p_{aj})d_{aj}\theta_{aj}\Big\}. \quad (15)\end{aligned}$$

Having thus 'completed' the sample data, we move to the MPQL-step, where we apply the IWLS algorithm outlined in Section 2.1.3, respectively, to the complete-data model and the missing-data model conditional on the complete data.

4.2 Estimation of CMSEP

Evaluating the CMSEP at the estimated parameter values yields a plug-in estimate of the CMSEP. Of the three *h*-terms, h_{1a} is of the order $O_p(1)$, whereas both h_{2a} and h_{3a} are of the order $O_p(A^{-1})$, when the number of areas tends to infinity while the within-area sample sizes remain bounded. The results of Booth and Hobert (1998) and Prasad and Rao (1990), obtained in the univariate complete-data case, suggest that the bias in the plug-in estimate $\hat{h}_{1a}$ is of the same order as $\hat{h}_{2a}$ and $\hat{h}_{3a}$. These authors developed second-order correction through the Taylor expansion. We do not pursue such second-order asymptotics in this paper. Approximate expressions of the *h*-terms that accompany the EMPQL algorithm are given below.

Take first h_{1a} by (12). Based on the linearized data (6), the covariance matrix $\mathrm{Cov}(\mathbf{v}_a,\mathbf{v}_a\mid \mathbf{z}_a)$ does not depend on either $\mathbf{z}_a$ or $\mathbf{x}_a$. This is convenient because we then have

$$\begin{aligned}h_{1a}(\mathbf{x}_a;\zeta,\delta) &\approx B_a\mathrm{Cov}(\mathbf{v}_a,\mathbf{v}_a\mid \mathbf{z}_a)B_a^T\\ &= B_a(G-GB_a^TV_a^{-1}B_aG)B_a^T \quad (16)\end{aligned}$$

where $V_a=B_aGB_a^T+QR_aQ^T$ is the marginal covariance matrix of $\mathbf{z}_a$.

Next, take h_{2a} by (13). Let $\phi=(\zeta^T,\delta^T)^T$. Expanding $\tilde{\phi}$ around ϕ yields $\tilde{\mu}_a-\dot{\mu}_a\approx\dot{\mu}_a'(\tilde{\phi}-\phi)$, where $\dot{\mu}_a'=\partial\dot{\mu}_a/\partial\phi$, such that

$$h_{2a}\approx\dot{\mu}_a'\,\mathrm{Cov}(\tilde{\phi},\tilde{\phi})\dot{\mu}_a'^T. \quad (17)$$

Based on (6) we derive $\dot{\mu}_a=H_a\zeta+D_a\dot{\mathbf{u}}_a$, where $D_a=B_aGB_a^TV_a^{-1}$ and $\dot{\mathbf{u}}_a=\mathbf{z}_a-H_a\zeta$. Denote by *I* the identity matrix. The partial derivatives in $\dot{\mu}'_a$ are given by

$$\partial\dot{\mu}_a/\partial\zeta=(I-D_a)H_a$$

and

$$\partial\dot{\mu}_a/\partial\delta_j=(\partial D_a/\partial\delta_j)\dot{\mathbf{u}}_a=(I-D_a)B_a(\partial G/\partial\delta_j)B_a^TV_a^{-1}\dot{\mathbf{u}}_a$$

where δ_j is the j^{th} variance parameter in the covariance matrix $G(\delta)$ of $\mathbf{v}_a$. To obtain $\mathrm{Cov}(\tilde{\phi},\tilde{\phi})$, suppose that the PQL approach is based on the following quasi log-likelihood

$$\ell=\sum_a\ell_a$$

and

$$\ell_a=-\frac{1}{2}\log|V_a|-\frac{1}{2}(\mathbf{z}_a-H_a\zeta)^TV_a^{-1}(\mathbf{z}_a-H_a\zeta).$$

The so-called sandwich formula yields then

$$\mathrm{Cov}(\tilde{\phi},\tilde{\phi})=\left(-\frac{\partial^2\ell}{\partial\phi^2}\right)^{-1}\left\{\sum_{a=1}^{A}\left(\frac{\partial\ell_a}{\partial\phi}\right)\left(\frac{\partial\ell_a}{\partial\phi}\right)^T\right\}\left(-\frac{\partial^2\ell}{\partial\phi^2}\right)^{-1}.$$

Finally, take h_{3a} by (14). Similarly as above we have $\tilde{\mu}_a=(I-\tilde{D}_a)H_a\tilde{\zeta}+\tilde{D}_a\tilde{\mathbf{z}}_a$ evaluated at $\phi=\tilde{\phi}$, and $\hat{\mu}_a=(I-\hat{D}_a)H_a\hat{\zeta}+\hat{D}_a\hat{\mathbf{z}}_a$, where $\hat{\mathbf{z}}_a$ is derived from $\hat{\mathbf{t}}_a=\mathbf{t}(\hat{\mathbf{x}}_a)$ for $\hat{\mathbf{x}}_a=E(\mathbf{x}_a\mid\mathbf{y}_a,m_{a.};\hat{\phi},\hat{\psi})$. Expanding $\hat{\phi}$ around $\tilde{\phi}$ and retain only the leading term, we obtain

$$\hat{\mu}_a-\tilde{\mu}_a\approx\breve{\mu}_a-\tilde{\mu}_a=\tilde{D}_a(\breve{\mathbf{z}}_a-\tilde{\mathbf{z}}_a)$$

where $\breve{\mu}_a=(I-\tilde{D}_a)H_a\tilde{\zeta}+\tilde{D}_a\breve{\mathbf{z}}_a$, and $\breve{\mathbf{z}}_a$ is derived from $\breve{\mathbf{t}}_a=\mathbf{t}(\breve{\mathbf{x}}_a)$ for $\breve{\mathbf{x}}_a=E(\mathbf{x}_a\mid\mathbf{y}_a,m_{a.};\tilde{\phi},\hat{\psi})$. That is, we ignore the terms involving $\hat{\phi}-\tilde{\phi}$. The remaining variation in $\breve{\mathbf{z}}_a$ is due to the estimation of the missing-data model alone. Expanding $\hat{\psi}$ around ψ, we obtain, by the chain rule,

$$h_{3a}\approx C_a\mathrm{Cov}(\hat{\psi},\hat{\psi}\mid\mathbf{x})C_a^T \quad (18)$$

and

$$C_a=\left\{D_a\left(\frac{\partial\mathbf{z}_a}{\partial\mathbf{t}_a}\right)\left(\frac{\partial\mathbf{t}_a}{\partial\mathbf{x}_a}\right)\left(\frac{\partial\mathbf{x}_a}{\partial\mathbf{p}_a}\right)\left(\frac{\partial\mathbf{p}_a}{\partial\eta_a}\right)\left(\frac{\partial\eta_a}{\partial\psi}\right)\right\}_{\phi=\tilde{\phi},\psi}$$

where we assume that $E(\hat{\psi}\mid\mathbf{x})=\psi$ and $E[\breve{\mathbf{z}}_a\mid\mathbf{x}]=\tilde{\mathbf{z}}_a$. Whereas the sandwich formula yields $\mathrm{Cov}(\hat{\psi},\hat{\psi}\mid\mathbf{x})$ under the conditional model of **y** given **x**, similarly to $\mathrm{Cov}(\tilde{\phi},\tilde{\phi})$ above.

4.3 Estimation of small area compositions

Suppose first that the GLSMM, defined by (2) and in combination with (5), has been estimated, upon which we obtain $\hat{\mu}_a^X$, and $\hat{\theta}_{ak}^X = \exp(\hat{\mu}_{ak}^X)/\sum_j \exp(\hat{\mu}_{aj}^X)$.

When the marginal totals $X_{a.}$ and $X_{.k}$ are known, it makes sense to apply the IPF, starting with the estimated table $\{\hat{\theta}_{ak}^X\}$. The difference from SPREE, which starts with the auxiliary table $\mathbf{X}^0$, is that the interactions have been re-estimated. On convergence we obtain the estimated small area counts, denoted by $\hat{\mathbf{X}} = \{\hat{X}_{ak}\}$, and the corresponding compositions, denoted by $\theta_{ak}^{\hat{X}} = \hat{X}_{ak}/\sum_j \hat{X}_{aj}$, which are different from the direct model estimates $\hat{\theta}_{ak}^X$ that have provided the starting values for the IPF.

Often in practice, while the area totals $\{X_{a.}\}$ may be known, the marginal totals $\{X_{.k}\}$ need to be estimated based on the survey data available, separately using a method that is appropriate for the aggregated level. The IPF is still worth considering as long as these estimated marginal totals are judged to be more reliable and/or less biased than the aggregated small area estimates $\sum_a X_{a.}\hat{\theta}_{ak}^X$. The reason is that the estimated interactions $\hat{\alpha}_{ak}^X$ are preserved in the IPF, *i.e.*, $\alpha_{ak}^{\hat{X}} = \hat{\alpha}_{ak}^X$. By the log-linear identity (3), the difference between the direct model estimate $\hat{\theta}_{ak}^X$ and final estimate $\theta_{ak}^{\hat{X}}$ is due to the difference in the estimates of the main effects $\{\alpha_k^X\}$. Thus, less biased estimates of $\{X_{.k}\}$ are expected to yield less biased estimates of $\{\alpha_k^X\}$ and, thereby, less biased estimates of $\{\theta_{ak}^X\}$.

Suppose next that the MSLMM (7) combined with (5) have been estimated. We may express the interest of estimation, *i.e.*, $\{\mu_{ak}^X\}$, in terms of $\mathbf{z}_a$ defined as

$$\begin{aligned}\mathbf{z}_a &= H_a\zeta + B_a\mathbf{v}_a + \mathbf{e}_a = H_a\zeta + B_a\mathbf{v}_a + \mathbf{e}_a^X + \mathbf{e}_a^{x|X} \\ &= \mu_a^X + \mathbf{e}_a^{x|X} = H_a\zeta + \mathbf{v}_a^X + \mathbf{e}_a^{x|X}\end{aligned}$$

where $\mathbf{e}_a^X = Q(\theta_a^X - \theta_a)$ and $\mathbf{e}_a^{x|X} = Q(\mathbf{t}_a - \theta_a^X)$. In accordance we have $R_a = R_a^X + R_a^{x|X}$, where $R_a^X = \mathrm{Cov}(\theta_a^X, \theta_a^X \mid \theta_a)$ and $R_a^{x|X} = \mathrm{Cov}(\mathbf{t}_a, \mathbf{t}_a \mid \theta_a^X)$. It follows that

$$\hat{\mu}_a^X = H_a\hat{\zeta} + (B_a\hat{G}B_a^T + \hat{Q}\hat{R}_a^X\hat{Q}^T)\hat{V}_a^{-1}(\hat{\mathbf{z}}_a - H_a\hat{\zeta}). \quad (19)$$

The rest follows as above where μ_a^X is estimated directly under the GLSMM.

5. Example: Register-based small area household compositions

5.1 Register household data

Register-based household data have undergone considerable development in Norway. One of the goals is to produce detailed household statistics that traditionally are only available from the census. For this purpose the registration of a unique dwelling identity number (DIN) was initiated in the last census in 2001. The work is not yet completed, and the DIN is still missing for about 6% of the people residing in the country. The rate of missing is differential as it varies over the household type as well as across the Municipalities, the latter of which is a reflection of the overall effort of the local administration regarding the registration of the DINs.

A household register can be complied in a year after the census based on a number of data sources. The most important ones include the central population register (CPR), the DIN-register and the census household file (CH01). Even without the DIN a register household can be compiled based on the other information available. But the result suffers from informative under-registration of the DIN. For instance, a typical source of bias is cohabitants living without children, because such a couple appear as two single-person households in the CPR, unless they have already been identified as a household in the CH01. Nevertheless, historic as well as cross-country comparisons suggest that the national totals are acceptable. A more urgent problem lies on lower levels of aggregation. For example, changes from the census in 2001 are unlikely large in certain Municipalities, including the capital city Oslo where the increase in the proportion of single-person households is almost three times as high as it is in the rest of the country - see top-left plot in Figure 1. And a large part of the problem in Oslo can be explained by a combination of high proportion of cohabitants living without children and low DIN-registration rate (indeed, the lowest in the country).

5.2 Set-up of data

We shall illustrate our approach using these register household data. The target population contains all persons living at multiple-dwelling addresses at the beginning of year 2005, who do not belong to households of married people or registered partners; the latter household types are excluded because the DIN is not critical for compiling the households of these people. There is no distinction between the finite population and the sample in this case, *i.e.*, $\mathbf{X} = \mathbf{x}$. The households that have registered DINs are treated as the 'observed' sample $\mathbf{y}$, whereas the households that do not have registered DINs are viewed as the missing. In this way the population consists of 713,387 persons, of which 558,136 persons have registered DINs. The overall rate of missing is about 22%.

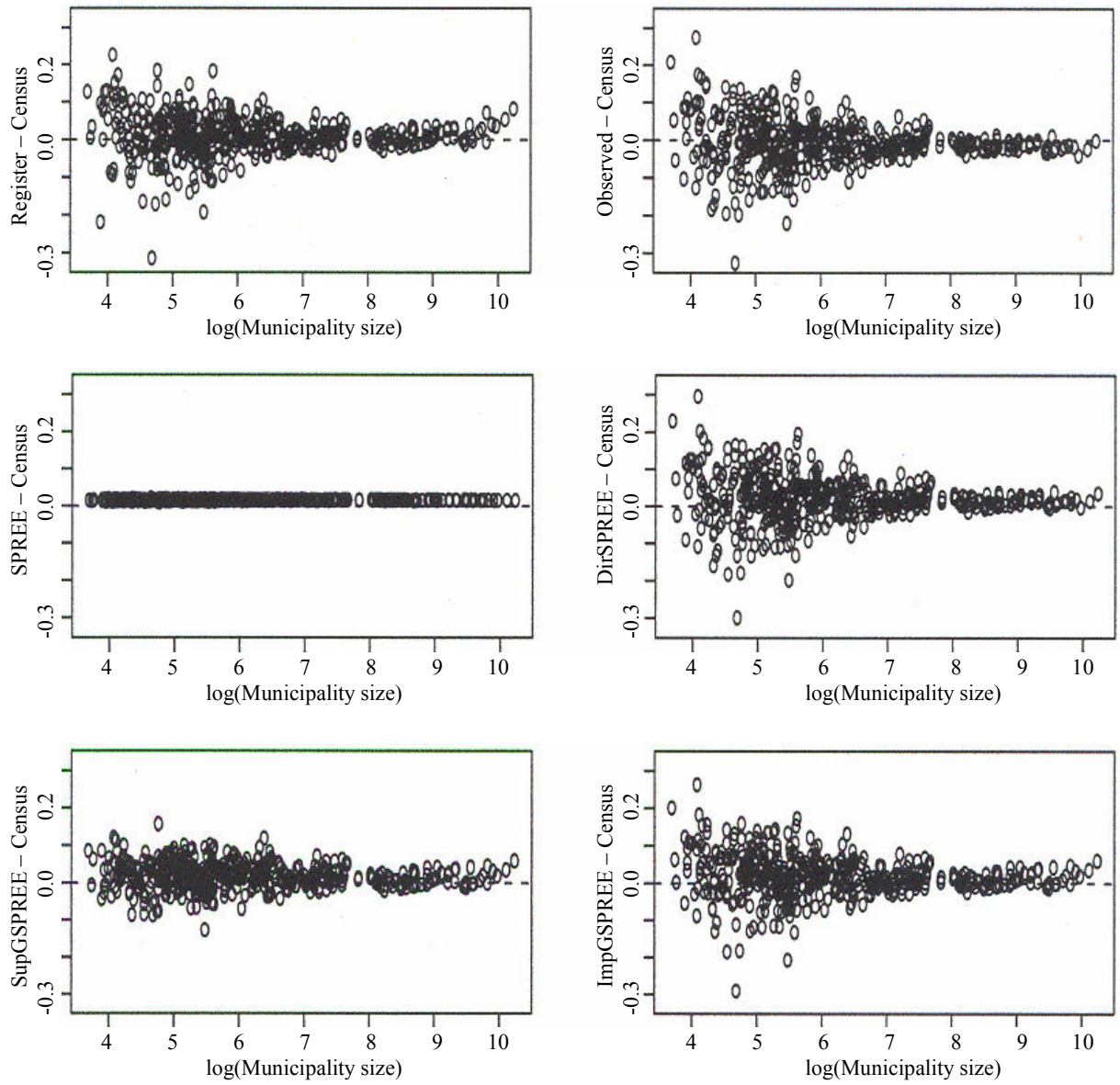

Figure 1 **Difference between estimates of proportion of Single-person households and census counts in 2001 against log Municipality size: Register households (top-left), Households with registered DINs (top-right), SPREE based on census (middle-left), DirSPREE based on households with registered DINs (middle-right), SupGSPREE of super-population proportions (bottom-left), and ImpGRSREE of imputed finite-population proportions (bottom-right). The dashed line marks no difference**

Let the Municipalities be the small areas of this study, where $A = 433$. The households are classified into 4 categories: $k = 1$ for "Single-person", $k = 2$ for "Single-parent", $k = 3$ for "Cohabitants", and $k = 4$ for "Other", *i.e.*, $K = 4$. Let i index the households, and let x_i be the number of persons living in the household. Let $X_{ak} = x_{ak}$ be the number of persons in the $(a,k)^{\text{th}}$ cell in the population, and let y_{ak} be the corresponding 'observed' cell count. Let N_{ak} be the number of households in the $(a,k)^{\text{th}}$ cell, and let n_{ak} be the corresponding number of 'observed' households. Notice that only the total number of persons is known in each area, but not the total number of households. However, provided cell-specific probability of DIN-registrations, an estimator of N_{ak} based on $\hat{X}_{ak}$ is given by $\hat{N}_{ak} = n_{ak}\hat{X}_{ak}/y_{ak}$. We shall therefore concentrate on the estimation of X_{ak} here.

Let $\{X^0_{ak}\}$ be the corresponding cell counts from the last census in 2001. Let $X'_{ak} = y_{ak} + m'_{ak}$ be the register counts in 2005, where m'_{ak} is the number of persons without the DIN. A register household can be considered as a form of imputed household that may suffer from informative missing of DINs. The register area total is correct, *i.e.*,

$X'_{a.} = X_{a.}$, and the national totals $\{X'_{.k}\}$ are considered acceptable. The question is whether estimates of $\{X_{ak}\}$ can be derived, based on the 'observed' **y** and the allocation structure $\{X_{a.}\}$ and $\{X'_{.k}\}$, that better accounts for the differential missing DINs.

5.3 Set-up of model

Scatter plots of the register first-order interactions $\{\alpha_{ak}^{X'}\}$ against the census interactions $\{\alpha_{ak}^{0}\}$ provide motivation for the PIMM (4). To chose between the GLSMM (2) and the MSLMM (7), we look at the difference between the register proportion $\theta_{ak}^{X'}$ and the corresponding census proportion θ_{ak}^{0}, *i.e.*, $\theta_{ak}^{X'} - \theta_{ak}^{0}$, plotted against $\log X_{a.}$: the case of $k = 1$ is shown in the top-left plot of Figure 1. Clearly, the variance of the difference increases as $X_{a.}$ decreases, and is not constant of $X_{a.}$. Notice that we are dealing with estimation at a very low level of aggregation here, where *e.g.*, the median value of all $\{X'_{ak}\}$ is only 70. We therefore adopt the model (7) for θ_{ak}, the quasi-likelihood (5) for $X_{ak} = x_{ak}$, and the quasi-likelihood (8) and the model (9) for y_{ak}.

For the quasi-likelihood (5) we assume $\nu_1 = 1$. Let $t_{ak} = X_{ak}/X_{a.}$. We have

$$V(t_{ak}) = N_{a.}^{-1}\theta_{ak}(1-\theta_{ak})\bar{X}_a^{(2)}/\bar{X}_{a.}^2$$

and

$$\mathrm{Cov}(t_{ak}, t_{aj}) = -N_{a.}^{-1}\theta_{ak}\theta_{aj}\bar{X}_a^{(2)}/\bar{X}_{a.}^2.$$

where $\bar{X}_a^{(2)} = \sum_{i=1}^{N_{a.}} x_i^2/N_{a.}$ and $\bar{X}_{a.} = X_{a.}/N_{a.}$. Since $x_i \geq 1$, we have $\bar{X}_a^{(2)} \geq \bar{X}_{a.}^2$, and over-dispersion compared to the Multinomial-$(N_{a.}, \theta_a)$ distribution. We calculate the factor $\bar{X}_a'^{(2)}/\bar{X}_{a.}'^2$ based on the register data, which is then used as $\bar{X}_a^{(2)}/\bar{X}_{a.}^2$ in the estimation below. Moreover, for the quasi-likelihood (8) we assume $\nu_2 = 1$, and

$$V(y_{ak}|n_{ak}) = V\left(\sum_{i=1}^{n_{ak}} r_{i,ak}x_{i,ak}\right) = \left(\sum_i x_{i,ak}^2\right)V(r_{i,ak}) \Rightarrow c_{ak} = \left(\sum_i x_{i,ak}^2\right).$$

5.4 Estimation results

Six different estimators of the proportion of Single-person households (*i.e.*, for $k = 1$) are illustrated in Figure 1.

To start with, we have the direct register proportions $\theta_{a1}^{X'}$ in the top-left plot, and the 'observed' proportions θ_{a1}^{y} in the top-right plot. On average the proportion is increased based on the entire register compared to the census in 2001, whereas it is slightly decreased according to the 'observed' part only. This demonstrates that the missing DINs are informative, as explained before. Inclusion of the register households without the DINs raises the proportion of Single-person households. But the result is implausible in some of the largest Municipalities. Of course, large bias also exists among the smaller Municipalities, but these are not easily detectable in a plot like this one.

Next, in the middle-left plot of Figure 1, estimates are obtained by SPREE using the census counts $\{X_{ak}^0\}$ as the starting values. For the simple two-way table here, this yields an almost constant adjustment of the census proportions, with negligible change in the between-area variation. In the middle-right plot, estimates are obtained by SPREE using the 'observed' table $\{y_{ak}\}$ as the starting values. Notice that, to start with the observed sample counts would be too unstable to be useful in usual survey sampling situations, but it is a viable option here because of the large amount of 'observed' data. To distinguish from the standard SPREE we shall refer to it as the *direct* SPREE (DirSPREE). As noted earlier, DirSPREE is unbiased under the assumption (10) of informative missingness. Indeed, it is seen to lead to useful adjustments for the largest Municipalities.

In the bottom-row plots of Figure 1, estimates are obtained using the double-mixed modeling approach. The estimates of the bottom-left plot are obtained by the IPF starting with the estimated super-population compositions $\{\hat{\theta}_{ak}\}$, denoted by *SupGSPREE*. The extreme post-censal development in the largest Municipalities are reduced. But the changes from the census-proportions are clearly over-shrunk towards to the population average for the smaller areas. The variation is *e.g.*, much less than that of $\theta'_{ak} - \theta^0_{ak}$ in the top-left plot. The estimates of the bottom-right plot are derived from the imputed finite-population counts, denoted by *ImpGSPREE*, which are calculated at the E-step of the EMPQL algorithm. The estimates for the largest Municipalities are similar to those of SupGSPREE, and the variation in the changes from the census-proportions is similar to that of DirSPREE.

5.5 Estimation of CMSEP

Approximate CMSEP of the ImpGSPREE compositions can be derived similarly as in Section 3. Denote by $\hat{X}_{ak}$ the ImpGSPREE count, and by $\dot{X}_{ak}$ the BP based on known conditional distribution of $\mathbf{X}_a$ given $(\mathbf{y}_a, m_{a.})$. We have

$$\mathrm{CMSEP}(\hat{\mathbf{X}}_a) \approx E\{(\dot{\mathbf{X}}_a - \mathbf{X}_a)(\dot{\mathbf{X}}_a - \mathbf{X}_a)^T \mid \mathbf{y}_a, m_{a.}\} + E\{(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)^T\}.$$

Moreover, let $\tilde{\phi}$ be the hypothetical estimate of ϕ based on the complete data $\mathbf{x} = \mathbf{X}$, and let $\hat{\psi}$ be the estimate of ψ based on the observed data. Let Q_1 and Q_2 be, respectively, the Jacobian matrix of partial derivatives $\partial\dot{\mathbf{X}}_a/\partial\phi$ and $\partial\dot{\mathbf{X}}_a/\partial\psi$. We have

$$\begin{aligned} E\{(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \dot{\mathbf{X}}_a)^T\} &\approx E\{(\tilde{\mathbf{X}}_a - \dot{\mathbf{X}}_a)(\tilde{\mathbf{X}}_a - \dot{\mathbf{X}}_a)^T\} \\ &\quad + E\{(\hat{\mathbf{X}}_a - \tilde{\mathbf{X}}_a)(\hat{\mathbf{X}}_a - \tilde{\mathbf{X}}_a)^T \mid \mathbf{X}\} \\ &\approx Q_1\mathrm{Cov}(\tilde{\phi}, \tilde{\phi})Q_1^T + Q_2\mathrm{Cov}(\hat{\psi}, \hat{\psi} \mid \mathbf{X})Q_2^T. \end{aligned}$$

Together, these lead to a three-part decomposition of the CMSEP similar to (12) - (14). In the estimation of the CMSEP below we ignore the effect of IPF. This is justified in our case because the IPF essentially amounts to a constant multiplicative adjustment very close to unity, as can be seen in the middle-left plot in Figure 1.

The CMSEP of a DirSPREE count is calculated as a 'sampling' variance that is induced by missing-at-random within each cell of the two-way table, plus a squared bias term which is estimated by the squared difference between the ImpGSPREE count and the corresponding DirSPREE count, provided the assumption (9) is a more appropriate model for the missing data than the assumption (10).

The estimated root CMSEPs (rcmsep) are given in Figure 2. On average both are decreasing as the Municipality size increases. However, for some of the largest Municipalities, the CMSEP of the DirSPREE proportion is abnormally large for Single-person and Cohabitants households due to the bias term. On the whole the CMSEP of the ImpGSPREE composition is clearly smaller than that of the DirSPREE. The h_{1a} -term, corresponding to the prediction variance of $\mathbf{X}_a$, is by far the dominating contribution to the CMSEP (over 99% in many areas). This is understandable since there are over 550 thousand people in the 'observed' sample, such that the uncertainty in parameter estimation is comparatively negligible. But the quoted percentage will be lower in a sample survey situation, as the estimation uncertainty summarized in terms h_{2a} and h_{3a} increases.

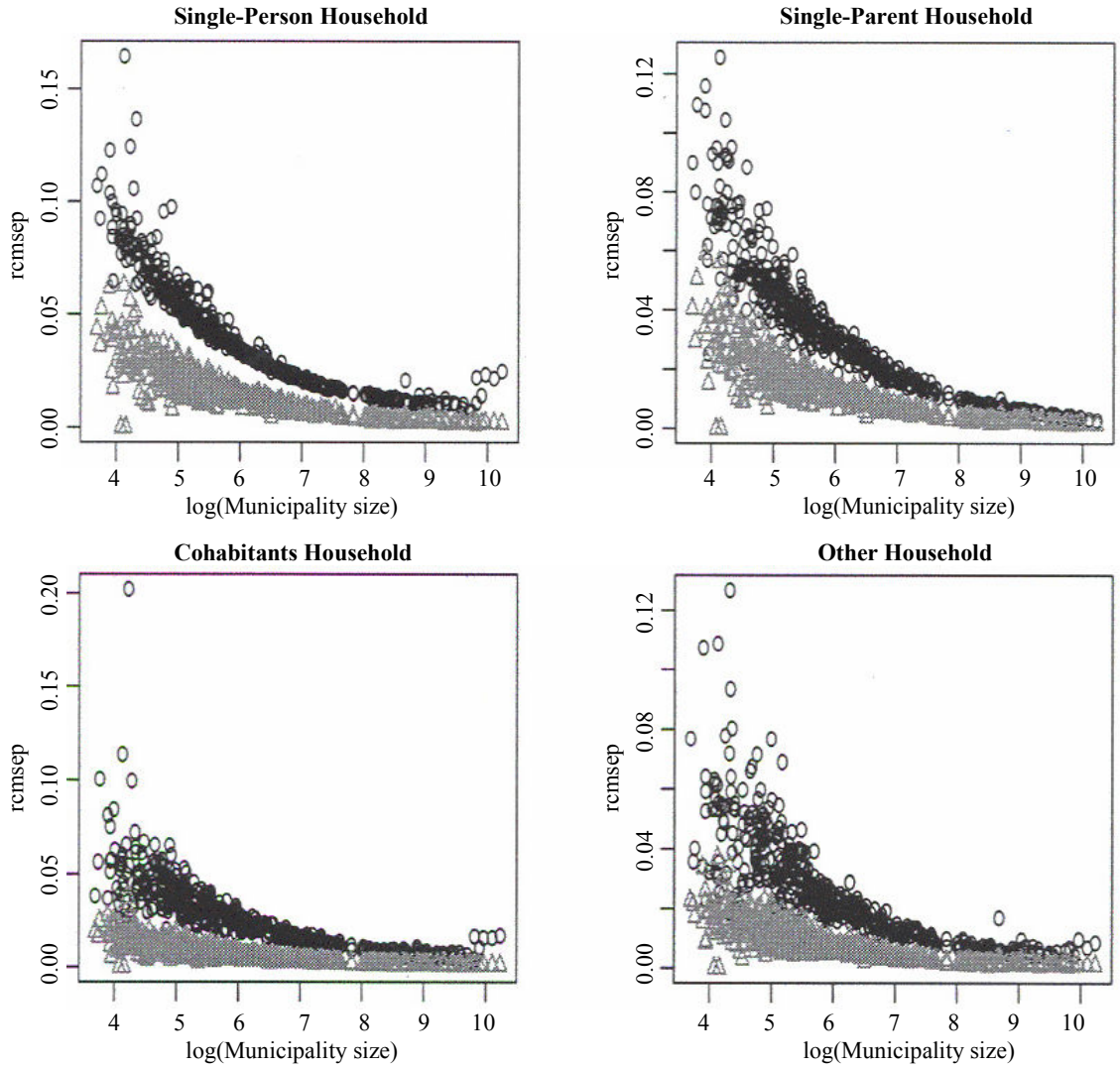

Figure 2 Estimated root conditional mean squared error of prediction (rcmsep) of DirSPREE (circle) and ImpGSPREE (triangle) of Municipality household proportions

6. Summary

In the above we outlined a double-mixed modeling approach that extends the GSPREE methodology to estimation of small area compositions subjected to differential missing data. An approximate CMSEP was derived which contains a three-part decomposition, corresponding to the prediction variance of the unknown random effect, the sampling variance in the absence of missing data, and the extra variance due to the missing data, respectively. The approach was applied to the Norwegian register household data, which yielded useful adjustments for informative missing of dwelling identity numbers.

Acknowledgements

I am thankful to the referees and the Associate Editor for comments and suggestions that have helped to improve the presentation.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Booth, J.G., and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262-272.

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

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*, 93-273-282.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Longford, N. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society*, Series A, 162, 227-245.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. London: Champman and Hall.

Prasad, N.G., and Rao, J.N.K. (1990). The estimation of mean square errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.

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

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727.

Zhang, L.-C., and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society*, Series B, 66, 479-496.

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 203-214
Statistics Canada, Catalogue No. 12-001-X

Small area population prediction via hierarchical models

Debora F. Souza, Fernando A.S. Moura and Helio S. Migon [1]

Abstract

This paper proposes an approach for small area prediction based on data obtained from periodic surveys and censuses. We apply our approach to obtain population predictions for the municipalities not sampled in the Brazilian annual Household Survey (PNAD), as well as to increase the precision of the design-based estimates obtained for the sampled municipalities. In addition to the data provided by the PNAD, we use census demographic data from 1991 and 2000, as well as a complete population count conducted in 1996. Hierarchically non-structured and spatially structured growth models that gain strength from all the sampled municipalities are proposed and compared.

Key Words: Markov Chain Monte Carlo (MCMC); Population projection; Spatial models.

1. Introduction

Like many other countries, the demand for detailed and updated small area statistics has been steadily growing in Brazil. This increasing demand is motivated by the need to have a more precise picture of subregions and has been driven by issues of distribution, equity and disparity. For instance, there may exist subregions or subgroups that are not keeping up with the overall average in certain respects. Therefore, there is a need to identify such regions and to have statistical information at that geographical level before taking any possible remedial action. Besides these national requirements, local authorities are faced with the need of having reliable estimates, such as demographic characteristics, for analysis, planning and administration purposes.

In Brazil, one important example of the demand for reliable estimates is related to how constitutionally mandated federal revenue sharing is apportioned annually to the various municipalities (Brazil is a federated republic made up of states and the Federal District. The states are divided into municipalities, which share characteristics of cities and counties - they can contain more than one urban area, but they have a single mayor and municipal council). The predicted number of inhabitants in a municipality is used by the federal government as a criterion to distribute funding. Hence, there is a need to obtain reliable municipal population forecasts in order to fairly apply this criterion, regulated by federal law.

An important source of demographic data is the annual Household Survey (PNAD). However, this survey is not designed to produce estimates at the municipal level. In other words, apart from a few municipalities, the municipal sample sizes are not large enough to yield acceptable standard errors when the direct survey estimates are used. Furthermore, a considerable number of municipalities are not sampled at all.

The current approach to obtain municipal population estimates is based on making prediction for a larger area at first, and then using some auxiliary information to allocate the total predicted population to the municipalities. In turn, prediction for a larger area is done by assuming that birth, mortality and migration rates are the same for all municipalities. The major drawback of this approach is that it relies on the assumed model evolution. It does not take into account all uncertainties and does not provide, in general, error measures of the estimates.

The small area estimation problem has received attention in the statistical literature due to the growing demand for detailed statistical information from the public and private sectors. An excellent and updated account of methods and applications of small area estimation can be found in Rao (2003). The main source of small area data is provided by periodic surveys whose sample sizes are not large enough to provide reliable estimates for the areas. One way of tackling this problem is to gain strength from all areas and through other sources of related data. As stated in Pfeffermann (2002), the sources of data suitable for this task can be classified into two categories: data obtained from other similar areas with respect to the characteristic of interest and past data obtained for the characteristic of interest and auxiliary information. In our demographic context, the main source of related data is provided by the 1991 and 2000 censuses and a complete count of the population carried out in 1996.

The aim of this work is to obtain estimates of the municipal populations based on survey data provided by the PNAD and census data. A non-structured hierarchical model is proposed and its fitness and predictive power are

1. Debora F. Souza, Department of Methods and Quality, IBGE, Rio de Janeiro, 20031-170. E-mail: debora.souza@ibge.gov.br; Fernando A.S. Moura, Universidade do Brasil-UFRJ, Rio de Janeiro, Brazil. E-mail: fmoura@im.ufrj.br; Helio S. Migon, Universidade do Brasil-UFRJ, Rio de Janeiro, Brazil. E-mail: migon@im.ufrj.br.

evaluated. We also consider a spatially structured hierarchical model, in the spirit of Moura and Migon (2002), since the population per area and its growth pattern might be related to the development of its neighboring areas. For the sake of simplicity, from now on we respectively call the non-structured hierarchical and spatially structured hierarchical models as the Hierarchical model and Spatial model.

In Section 2 the main data sources used in this work are described. In Section 3, the proposed models and a model selection criteria are presented. Applications with real and a simulated data are presented in Section 4. Finally, Section 5 contains a brief summary with an outline for future research.

2. Data set

The input data for the models introduced in Section 3 are taken from the annual Household Surveys (PNADs) from 1992 to 1999, the 1991 and 2000 census data and a complete enumeration of the population carried out in 1996. In order to evaluate the proposed approach, the municipalities of São Paulo State are considered as the areas of interest.

In this section we present a brief description of these data sources, reporting their main advantages and limitations. The population direct estimates of sampled municipalities were obtained from the PNAD. As explained in Section 3, these estimates are regarded as the input data for making inference about our target parameters. The two censuses and the 1996 population count are also utilized in our application.

The Brazilian Demographic Census is the main source of information about the population. It is carried out every ten years, usually in the beginning of the decade. Although the objective is to count all the population, some enumeration errors are found. The magnitude of the errors is evaluated through a post enumeration survey carried out soon after the completion of the census.

The annual Household Survey (PNAD) is designed to produce basic information about the socioeconomic situation of the country. The investigation unit is the household, for which yearly information about the number of dwellers, their gender, education level, employment, *etc*. is collected. The survey is not carried out in a census year, and was also not conducted in 1994 for administrative reasons. The sample is selected by a three-stage cluster sampling design. The primary and secondary units are respectively the municipality and enumeration areas (with 250 households on average). The municipalities are stratified according to their population sizes as obtained from the last census. In the first stage, all municipalities belonging to the metropolitan regions and the state capitals (which in Brazil are normally the largest cities in the respective states) are sampled. The municipalities whose populations are greater than some cutoff value are also included in the sample with probability one. The ones left are stratified and two of them are sampled from each stratum with probability proportional to their population sizes.

The enumeration areas are sampled with probability proportional to the number of households residing in the area in the last census. Finally, in the last stage the households are sampled systematically with equal probability from a list, which is updated at the beginning of the survey. The municipalities and enumeration districts are kept the same in all the surveys carried out in the same decade, while households are sampled every year.

Since each area is sampled with probability proportional to its respective number of households, it could be argued that the sampling mechanism is informative with respect to the population of the area. However, since the response variable actually used in this work is the area density, it is reasonable to assume that the sample selection mechanism is not relevant. Thus, this issue is not exploited in this work. A good reference about how to make small area inference under informative sampling is Pfeffermann and Sverchkov (2007). We also recommend Pfeffermann, Moura and Silva (2006) for readers interested in how to employ a Bayesian approach to hierarchically modeling under informative sampling.

3. Model specification

3.1 Exponential growth model

Let y_t be sample values of a distribution belonging to an exponential family with expected value given by $\pi_t = E(y_t \mid \theta_t)$ where θ_t is a vector of unknown parameters.

An important and wide class of exponential growth models parameterized by $(\alpha, \beta, \gamma, \phi)$ is defined as:

$$\pi_t = [\alpha + \beta \exp(\gamma t)]^{1/\phi}. \qquad (1)$$

Some special well-known cases in the literature are:

(1) Logistic: with $\phi = -1$, $\pi_t^{-1} = \alpha + \beta \exp(\gamma t)$;

(2) Gompertz: with $\phi = 0$, defining (1) as $\log(\pi_t) = \alpha + \beta \exp(\gamma t)$;

(3) Modified exponential: with $\phi = 1$, $\pi_t = \alpha + \beta \exp(\gamma t)$.

The main advantage of using model (1) is the possibility of keeping the observations y_t in the original scale, changing only the trajectory of π_t, making interpretation easy. Furthermore, the time intervals do not need to be of the same length, allowing the data to come from different reference sources (see Section 4 for further details).

When $\psi = \exp(\gamma) < 1$, the process is non-explosive, implying that π_t converges to $\alpha^{1/\phi}$ when $t \to \infty$, with the

convention that for $\phi = 0$, this quantity is equal to $\log(\alpha)$. When $\psi > 1$, the curves are concave for $\phi \geq 0$ and $\beta > 0$, leading to an explosive process. This class of models is called the generalized exponential growth model. Migon and Gamerman (1993) show how the exponential growth model can be viewed as a particular case of a general dynamic model.

3.2 Hierarchical growth models

In this paper our main parameters of interest π_{it} are nonlinear exponential growth functions with some parameters that are hierarchically or spatially structured. Spatially structured models provide alternative ways for connecting similar neighboring areas. We further assume that the sampling variance σ_{it}^2 follows a model that depends on the sample size in the respective municipality. In this work, hierarchical and spatial models are fitted and compared.

We assume that the population sizes are available for all the m municipalities of São Paulo State for the census years of 1991 and 2000, as well as the complete population count in 1996. From now on, we simply refer to them as the census data. In order to improve the hypothesis of exchangeability of the parameters describing the mean of the process, our response variables are set as the sampled municipal density estimates instead of the municipal population estimates. See also the end of Section 2 for further reasons for using the densities.

For each period, estimates of these quantities are available only for $k < m$ first-stage units municipalities of the PNAD sample. In order to estimate the municipal density, we simply divide the total population estimate by the respective municipal area.

Let y_{it} be the population density obtained from the census data or estimated by the PNAD at time t, $t = 1, \ldots n$ for the i^{th} municipality, $i = 1, \ldots, m$. Our aim is to make inferences about the true population density π_{it} for the population of all municipalities, including those that are not sampled. In the next section, true municipal population densities π_{it} are modeled via a stochastic nonlinear hierarchical growth function. We assume that the random quantities y_{it} are normally distributed with mean π_{it} and variance σ_{it}^2.

We use a Bayesian approach in this work. Therefore, predictions are described by probability distributions, giving the opportunity for users to analyze the uncertainties involved in the decision process. This fact is one of the advantages, among many others, of using this kind of approach.

Only in the census years are the y_{it} obtained for all the municipalities of São Paulo State. Although the census attempts to obtain complete enumeration of the whole population, coverage errors can occur. The following model is assumed therefore for the census data and the data obtained from the PNAD, with exception that the variances σ_{it}^2 are set to be smaller for the census data (see Section 3.4 and also the final remarks in Section 5):

$$\begin{aligned} y_{it} &= \pi_{it} + \varepsilon_{it},\ \varepsilon_{it} \sim N(0,\ \sigma_{it}^2) \\ \pi_{it} &= \{\alpha_i + \beta \exp(\gamma_i t)\}^{1/\phi} \\ \alpha_i &= \alpha + \xi_{\alpha_i},\ \xi_{\alpha_i} \sim N(0,\ \sigma_{\xi_\alpha}^2) \\ \gamma_i &= \gamma + \xi_{\gamma_i},\ \xi_{\gamma_i} \sim N(0,\ \sigma_{\xi_\gamma}^2) \end{aligned} \tag{2}$$

where the prior distributions of α, β and γ are given by: $\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$, $\beta \sim N(\mu_\beta, \sigma_\beta^2)$, $\gamma \sim N(\mu_\gamma, \sigma_\gamma^2)$. It should be noted that information from all areas is obtained through the hierarchical structure of the parameters α_i, and γ_i. Another way of borrowing information between municipalities is to assume that α_i are spatially structured (see Section 3.3). Supposing that the mean π_{it} is non-explosive, the parameter $\alpha^{1/\phi}$ can be regarded as the value at which the mean municipal population stabilizes. The parameters β and γ affect the evolution of the density over time. The prior distributions of α, β and γ can be chosen by taking advantage of some prior demographic knowledge of the expected population evolution. In our application, we set $\phi = 1$, implying that for $t = 0$ the true value density in each municipality is given by $\alpha_i + \beta$. The hierarchical structure imposed on the parameters α_i, implies that the expected value of the true density for any municipality at $t = 0$ is $\alpha + \beta$. To assume that the growth parameters, γ_i, have a hierarchical structure means that the densities have different growth rates but share the same mean. A small simulation study (see Section 4.1) guides us to keep the β parameter fixed for all areas, without any loss of generality, since the levels are still different for different municipalities. In all models considered in our application, we assume that $\tau_\alpha^2 = \sigma_\alpha^{-2} \sim G(a_\alpha, b_\alpha)$, $\tau_\gamma^2 = \sigma_\gamma^{-2} \sim G(a_\gamma, b_\gamma)$. In order to assign vague priors, in Section 4.2 we set small values for the parameters related to these precision prior distributions.

The assumption that the mean function π_{it} is given by an exponential growth curve allows adjusting for increasing or decreasing population density. The sources of data used have different reference data and are not equally spaced in time. In this case, the use of an exponential growth curve yields an extra advantage, since we can simply make a scale of time in order to conform with the different data sources, as explained in the application section 4.

3.3 Spatial model

In the Hierarchical model presented in the previous section, the information from all areas is combined in order to predict the population of a particular area. However, it is reasonable to assume that two or more neighboring municipalities have more similar demographic densities

than two other arbitrarily chosen ones. The regional structure is represented in the joint prior distribution of the random spatial effects. We consider that two areas are neighbors if they share a border.

In our proposed model, the demographic density in an area i at time t, π_{it}, is affected by its neighboring areas by adding random spatial effects δ_{α_i} to the parameters α_i, that is, $\alpha_i = \alpha + \delta_{\alpha_i}$, where α is a term representing the intercept. Therefore, α_i vary only with the spatial effect, representing a local effect, while the growth parameters γ_i's are regarded as similar among all areas (overall effect).

The relationship between neighboring areas is defined in the prior distributions of δ_{α_i}. The prior joint distribution of $\delta_\alpha = (\delta_{\alpha_1}, ..., \delta_{\alpha_m})'$ given the hyperparameter σ_α^2, is defined as in Mollié (1996):

$$p(\delta_\alpha \mid \sigma_\alpha^2) \propto \frac{1}{\sigma_\alpha^{m/2}} \exp\left\{-\frac{1}{2\sigma_\alpha^2}\sum_{i=1}^{m}\sum_{k<i} w_{ik}(\delta_{\alpha_i} - \delta_{\alpha_k})^2\right\} \quad (3)$$

where w_{ik} are the weights associated with the regional structure. The weights were chosen such that $w_{ik} = 1$, if i and k are contiguous, and $w_{ik} = 0$, otherwise. The distribution of $\delta_\alpha \mid \sigma_\alpha^2$ is evidently improper, since we can add any constant to all of the δ_{α_i} and $p(\delta_\alpha \mid \sigma_\alpha^2)$ is not affected. Thus, we must impose a constraint to ensure that the model is identifiable. We set $\sum_{i=1}^{m}\delta_{\alpha_i} = 0$ and assign a uniform prior distribution on the whole real line to the intercept α. It is not difficult to see that this procedure leads to a proper $(m-1)$ dimensional likelihood density, see Besag and Kooperang (1995) for further details.

The prior conditional distribution of δ_{α_i}, given the effects δ_{α_k} of the remaining areas and the hyperparameter σ_α^2, is normal with mean and variance given by:

$$E[\delta_{\alpha_i} \mid \delta_{\alpha_k},\ k \in \partial i,\ \sigma_\alpha^2] = \overline{\delta}_{\alpha_i}$$

$$\mathrm{Var}[\delta_{\alpha_i} \mid \delta_{\alpha_k},\ k \in \partial i,\ \sigma_\alpha^2] = \frac{\sigma_\alpha^2}{w_{i+}}$$

where $\overline{\delta}_{\alpha_i}$ denotes the arithmetic mean of the δ_{α_j} for $k \in \partial i$ (the contiguous areas of i), and $w_{i+} = \sum_{k=1}^{m} w_{ik}$ is the number of neighboring municipalities of i.

Figure 1 shows the demographic densities of São Paulo municipalities in 1991. These municipalities tend to be concentrated geographically according to density classes. This suggests that the spatial model can be usefully applied.

3.4 Modeling the sampling variances

Since we use data from two different sources, it makes sense to assume that the sampling variances vary over time. Furthermore, we can also consider that the variances change with the areas.

For the years in which the data are provided by the PNAD, we assume the following model for the sampling variances:

$$\log(\sigma_{it}^2) = \eta_0 + \eta_1 .(1/n_i) \quad (4)$$

with n_i representing the number of enumeration areas sampled in the i^{th} area. This model captures the expectation that the variance gets smaller as the sample size increases.

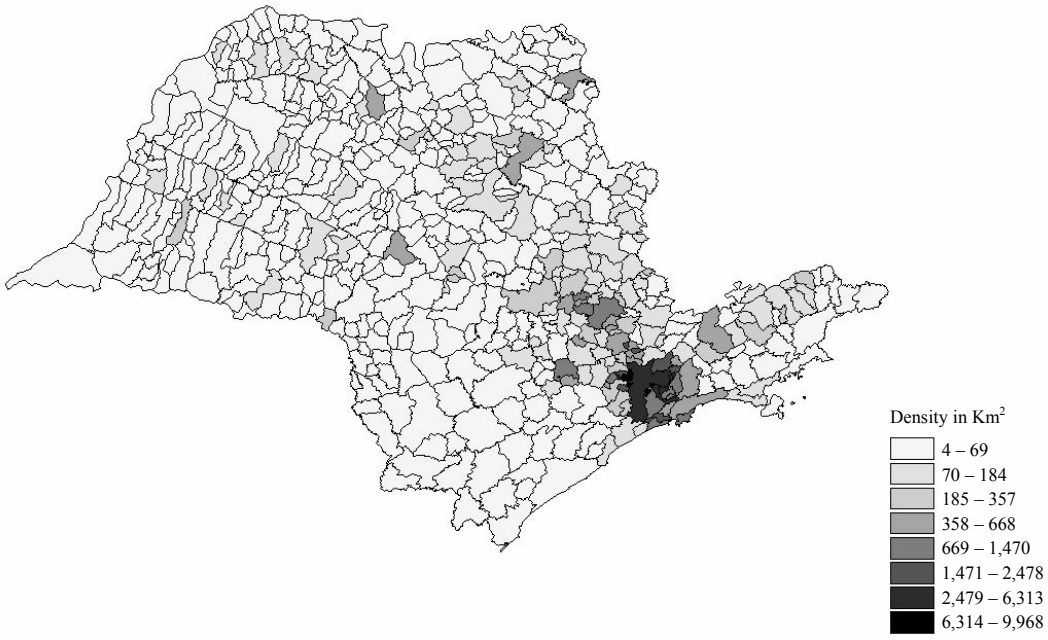

Figure 1 Population densities of São Paulo municipalities in 1991

For the years that the censuses were carried out, we assumed that σ_{it}^2 is known and $\log(\sigma_{it}^2) = \log(v_{it})$ where v_{it} is calculated in such a way that the census coverage error is 5% for all areas. This hypothesis implies that the true population in each area for census years lies in the interval given by the observed population in the census plus or minus 5% of this value. Therefore, for the census years we set the standard deviation as: $\sigma_{it} = 0.05 * (y_{it}/2)$. Assuming known variance in the census years is a way of giving more weight to census data, since one would expect a complete census to provide more reliable information than survey data. Independent normal distributions are assumed for the parameters η_0 and η_1: $\eta_k \sim N(\mu_{\eta_k}, \phi_{\eta_k})$; $k = 0,1$. In order to assign vague priors to the η's, we set both prior means as zero and large values for the ϕ_η's. See Section 4.2 for details.

3.5 Summary of the models

The prior distributions of the common parameters of the Spatial and Hierarchical models are the same as already described for the former. The distributions of the random spatial effects are specified in Section 3.3. The variance σ_{it}^2 in the Spatial model was stated as in the Hierarchical model. A summary of the models in Section 4 is presented in Table 1. For the sake of simplicity, the application was carried out by fixing $\phi = 1$ in both models.

3.6 Computational issues

The posterior distributions of the parameters for the models proposed cannot be obtained in closed forms. Therefore, it is necessary to use numerical approximation methods. One alternative, often used and easy to implement, is to generate samples of these distributions based on the Markov Chain Monte Carlo (MCMC) algorithm. Since the full conditional distributions of all the model parameters have closed form, except for the vector $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_k)$, we employed the Gibbs sampler algorithm with one acceptance/rejection algorithm step for sampling from the vector $\boldsymbol{\gamma}$. Let π_{it} be the population density in the i^{th} area at time t. The following steps summarize how to sample from the posterior distribution of π_{it}:

1. Generate $\alpha_i^{(l)}, \beta^{(l)}, \gamma_i^{(l)}, \alpha^{(l)}, \gamma^{(l)}, \tau_\alpha^{2(l)}, \tau_\gamma^{2(l)}, \eta_0^{(l)}$ and $\eta_1^{(l)}$ for $l = 1, ..., M$, where M is the number of MCMC samples generated from the full conditional distributions of all model parameters including the random effects;
2. Calculate $\pi_{it}^{(l)} = \alpha_i^{(l)} + \beta^{(l)}\exp(\gamma_i^{(l)}t)$;

Three informal checks for convergence, based on graphical techniques, were applied for assessing the convergence when fitting our proposed models. They consist of observing the histogram, the trace and the autocorrelation function for each of the sampled values calculated. The histogram analysis allows us to identify possible departures from convergence, such as the presence of multiple modes. The trace of the multiple chains simulated in parallel, each one with different starting points and overdispersed with respect to the target distribution, provides a rough indication of stationary behavior when the sequences of values tend to oscillate in the same region. The plot of the autocorrelation function allows identifying whether the sampling can be regarded as independent.

In addition to these informal checks, other more formal criteria were applied. The criteria introduced by Brooks and Gelman (1998) and implemented in WinBugs 1.4 (Spiegelhalter, Thomas, Best and Lunn 2004) permit diagnosing whether dispersion within chains is larger than dispersion between chains. Consider I parallel chain and a parameter of interest λ. Let λ_i^j be the j^{th} value of the i^{th} chain, for $i = 1, ..., K$ and $j = 1, ..., J$. Then the variances between chains $\hat{B}$ and within chains $\hat{W}$ are given by

$$\hat{B} = J(K-1)^{-1}\sum_{i=1}^{K}(\bar{\lambda}_i - \bar{\lambda})^2$$

and

$$\hat{W} = \{K(J-1)\}^{-1}\sum_{i=1}^{K}\sum_{j=1}^{J}(\lambda_i^j - \bar{\lambda}_i)^2$$

where $\bar{\lambda}_i$ and $\bar{\lambda}$ respectively are the average of observations of chain i, $i = 1, ..., K$ and the global average. Under convergence, all these KJ values are drawn from the posterior of λ and the variance of λ can be consistently estimated by $\hat{B}, \hat{W}$ and the weighted average $\hat{\sigma}_\lambda^2 = (1 - 1/J)\,\hat{W} + (1/J)\,\hat{B}$.

Table 1
Summary of the models employed

model	parameters	variance	prior distribution
Hierarchical	$\alpha_i = \alpha + \xi_{\alpha_i}$ β $\gamma_i = \gamma + \xi_{\gamma_i}$	$\log(\sigma_{it}^2) = \eta_0 + \eta_1(1/n_i)$, for survey data σ_{it}^2 is assumed to be known for census data	$\eta_0 \sim N(\mu_{\eta_0}, \phi_{\eta_0})$ $\eta_1 \sim N(\mu_{\eta_1}, \phi_{\eta_1})$
Spatial	$\alpha_i = \alpha + \delta_{\alpha_i}$ β $\gamma_i = \gamma + \xi_{\gamma_i}$	$\log(\sigma_{it}^2) = \eta_0 + \eta_1(1/n_i)$ in the survey σ_{it}^2 is assumed to be known for census data	$\delta_{\alpha_i} \mid \delta_{\alpha,-i}, \tau_\alpha^2 \sim N(\bar{\delta}_{\alpha_i}, \tau_\alpha^2/w_{i+})$ $\sum_{i=1}^{m}\delta_{\alpha_i} = 0$ $\eta_0 \sim N(\mu_{\eta_0}, \phi_{\eta_0})$ $\eta_1 \sim N(\mu_{\eta_1}, \phi_{\eta_1})$

If the chains have not yet converged, then initial values will still be influencing the trajectories and $\hat{\sigma}^2_\lambda$ will overestimate σ^2_λ until stationarity be reached. On the other hand, before convergence, $\hat{W}$ will tend to underestimate σ^2_λ. Following these reasoning, Brooks and Gelman (1998) proposed an iterated graphical approach, which is implemented in WinBugs 1.4. It allows to check if: (i) the weighted posterior variance estimated $\hat{\sigma}^2_\lambda$ and the within-chain variance $\hat{W}$ stabilize as a function of J, and (ii) the variance reduction factor, $\hat{R} = \hat{\sigma}^2_\lambda / \hat{W}$, approaches 1.

4. Application

In this section we present two applications of our approach, the first one with a simulated data set and the second one with the real data set that motivated this work. The simulation study aims to check if the parameters of interest are being properly estimated, as well as to perform some sensitivity analysis with respect to the form of the prior distributions used for fitting the model.

4.1 Application to simulated data

We carried out a small simulation study fitting the Hierarchical and Spatial models presented in Section 3. The true model hyperparameters related to the growth curve were fixed as $\alpha = 40$, $\beta = 25$, $\gamma = 0.05$. Thus, we are considering a situation where the population size approximately doubles in 25 years. The parameters related to the sampling variance model were fixed as $\eta_0 = 6.5$, $\eta_1 = 0.5$. Finally, the precision parameters were respectively set as $\tau^2_\alpha = 0.0001$ and $\tau^2_\gamma = 400$. The precision τ^2_α and τ^2_γ were fixed to be in agreement with the scales of the quantities they respectively measure. The intercept presents more relative variation between areas than the growth parameter, which is expected in practical situations.

Since it is well recognized that the form of the priors has more impact on the component of variance parameters than the fixed parameters, we fitted the simulated data using two different vague priors for the parameters related to the variances: uniform for the standard deviation, which is one of the priors recommended by Gelman (2006) for linear hierarchical models, and gamma for the precision, commonly used as the default in some computational packages. In the first case, we assigned $\sigma_\alpha \sim U(0, 1{,}000)$ and $\sigma_\gamma \sim U(0, 100)$, where $\sigma_\alpha = 1/\tau_\alpha$ and $\sigma_\gamma = 1/\tau_\gamma$. In the second case, we considered $\tau^2_\alpha \sim G(0.001, 0.001)$ and $\tau^2_\gamma \sim G(0.001, 0.001)$. For the other parameters, we set $\alpha \sim U(-\infty, +\infty)$, for the Spatial Model (see Section 3.3 for further details) and $\alpha \sim N(0, 10^6)$ for the Hierarchical model. For the others parameters we set $\beta \sim N(0, 10^6)$, $\gamma \sim N(0, 10^2)$, $\eta_0 \sim N(0, 10^4)$ and $\eta_1 \sim N(0, 10^4)$ for both models. The effect of the number of small areas is also investigated. We simulated separate data from the Hierarchical and Spatial models with $m = 60$ and $m = 100$ areas in each case. For each combination of the number of areas and the model employed we generated 200 data sets. Therefore, a total of 800 sets of artificial data was simulated. The distribution of the sample sizes within the areas is the same for the simulated data sets with 60 and 100 areas. Table 2 presents the relative frequencies of the small areas sample sizes for the both simulated data sets. These sample sizes are very similar to the sample sizes in the real data that underlines this simulation study. The number of neighbors employed in the spatial model varies from 1 to 12 and each area has on average 5 neighbors. We considered a total period of $n = 9$ years.

Table 2
Relative frequencies of the small area samples sizes for both simulated data sets

Sample size	Relative frequency
2	0.05
5	0.20
8	0.25
10	0.25
12	0.20
15	0.05

In order to get rid of chain correlation, we generated 20,000 samples after discarding the first 10,000. There is no evidence for non-convergence of the Hierarchical and the Spatial model parameters. A careful analysis of some outputs obtained from the MCMC samples for some simulation sets suggests that convergence was achieved for all model parameters. We assessed the statistical properties of the population density (π_{it}) estimates by investigating the average of the absolute relative error of the estimates (ARE) and the mean square error (MSE), respectively given by:

$$\mathrm{ARE}_{i,t} = \frac{1}{200}\sum_{l=1}^{200} \frac{|\hat{\pi}^{(l)}_{i,t} - \pi^{(l)}_{i,t}|}{\pi^{(l)}_{i,t}}$$

and

$$\mathrm{MSE}_{i,t} = \frac{1}{200}\sum_{l=1}^{200} (\hat{\pi}^{(l)}_{i,t} - \pi^{(l)}_{i,t})^2,$$

$i = 1, \ldots, m$, $t = 1, \ldots, n$. There is no much variation, as far as the ARE values are concerned. For the two models fitted and both small area sample sizes tried, the ARE values are around 1.5%.

Table 3 shows a summary of the MSE values obtained from the simulations carried out under the Spatial and Hierarchical models with 60 and 100 areas and respectively assigning gamma and uniform priors to the precision and to the standard deviation of the parameters related to the variance. It can be seen from Table 3 that the MSEs are not affected by the use of different vague priors. It is noteworthy that increasing the number of areas from 60 to 100 results in a small decrease of 6% in the median of the MSE for the Spatial model. However, for the case of the Hierarchical model, the decrease is about 13%.

Table 3
Summary of mean square error distribution for the spatial and hierarchical models

Model	Num. of areas	Gamma prior			Uniform prior		
		1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
Spatial	60	0.398	1.741	3.574	0.394	1.737	3.595
	100	0.525	1.637	3.538	0.524	1.641	3.517
Hierarchical	60	0.542	2.218	6.262	0.646	2.223	6.278
	100	0.594	1.959	5.593	0.596	1.960	5.619

We also investigated the percentage coverage of nominal 95% credible intervals. The results are presented in Table 4. As far as this simulation study is concerned, the intervals for the parameters of interest have in general the correct coverage percentages for both models investigated and these results do not depend on wether we have 60 or 100 areas. However, with a small number of areas we could face convergence problems unless we tighten the priors for the hyperparameters. The simulation study reveals that the population prediction is not affected by the forms of the vague priors assigned to the variance of the intercept term.

Table 4
The coverage rates of nominal 95% credible intervals for the population densities

Model	Num. of Areas	Gamma prior coverage(%)	Uniform prior coverage(%)
Spatial	60	96	96
	100	96	96
Hierarchical	60	94	94
	100	95	95

We analyzed the model fit when data generated from a model were fitted by the correct and the wrong models. Figure 2 presents the mean square error for the following situations: (a) data generated from the Spatial model and fitted by the Spatial and Hierarchical models and (b) data generated from the Hierarchical model and fitted by the Spatial and Hierarchical models. Since the form of the priors assigned to the parameters related to the variance does not affect the inference, we set uniform priors for both models. The ARE measures are shown in Figure 3.

It can be seen from Figure 2 that when the data are generated from the simpler model (Hierarchical) the more complex estimation procedures (Spatial) do not suffer any appreciable worsening of efficiency. On the other hand when the data are generated from the more complex model (Spatial) the simpler estimator (Hierarchical) has some inferior properties. However, this result does not hold for the ARE measurements. Figure 3 shows that fitting the model not used for generating the data results in appreciable increase in the relative bias. As it might be expected, model fitting and diagnostics are crucial in order to get suitable prediction of the small area population.

4.2 Application to real data

The PNAD data sets from 1992 to 1999, (excluding 1994 and 1996) and the population census data of 1991, 1996 and 2001 were used in our application. Our areas of interest are all the municipalities in São Paulo State, a total of 572 areas, of which 111 areas were sampled by the PNAD survey. Figure 4 shows the areas sampled by the PNAD, classified by the sampling definition: areas belong to metropolitan regions and self-representing areas (sampled with probability equal to 1) and non-self-representing areas. It should be noted that the census and PNAD have different periods of reference. We set $t = 0$ for the 1991 census. Thus, the values of t for the data provided by the PNAD are equal to the number of years between the reference period of the 1991 census and the respective PNAD. For instance, a survey datum provided by the PNAD 18 months after the 1991 census corresponds to $t = 1.5$.

Figure 5 shows the estimated coefficient of variation of the direct estimator by areas'sample sizes. These estimates are based on PNAD data. It can be seen that these coefficients of variation vary considerably with the areas and tend to decrease as the sample size increases. The high values of these coefficients show the difficulty in using only the direct estimator to provide municipal estimates. Furthermore, we cannot make any prediction for nonsampled areas by using only the direct estimators.

4.3 Specification of the prior distributions

The mean of the normal prior distributions of the parameters α, β and γ, related to the population evolution, were assigned by first expanding the function $\alpha + \beta\exp(\gamma t)$ around zero in a Taylor series up to the second order and then equating the resulting expression to the values of the mean density in the 1991 and 2000 censuses and the 1996 population count. In the absence of prior information, we considered a reasonably large value (10^6) for the prior variances of α, β and γ. Thus, we set $\alpha \sim U(-\infty, +\infty)$ (see Section 3.3 for further details), for the Spatial Model and $\alpha \sim N(370, 10^6)$, for the Hierarchical model and $\beta \sim N(726, 10^6)$, $\gamma \sim N(0.04, 10^6)$ for both models. The reason for this adjustment is to obtain a reasonable value of the prior means, but one that is essentially vague. Regarding the precisions and η_0, η_1, we assigned relatively vague priors: $\tau_\alpha^2 \sim \text{Ga}(0.001, 0.001)$, $\tau_\gamma^2 \sim \text{Ga}(0.001, 0.001)$, $\eta_0 \sim N(0, 10^6)$ and $\eta_1 \sim N(0, 10^6)$.

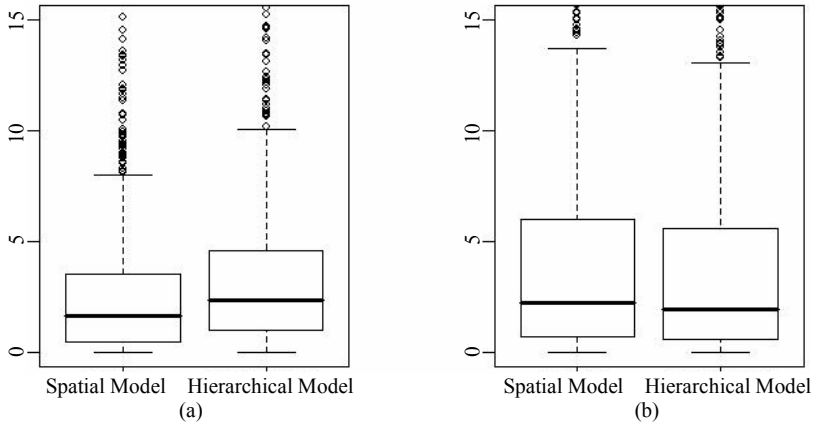

Figure 2 **Box plots of mean square error (MSE) for the cases: (a) data generated from the Spatial model and respectively fitted by the Spatial and Hierarchical models and (b) data generated from the Hierarchical model and respectively fitted by the Spatial and Hierarchical models**

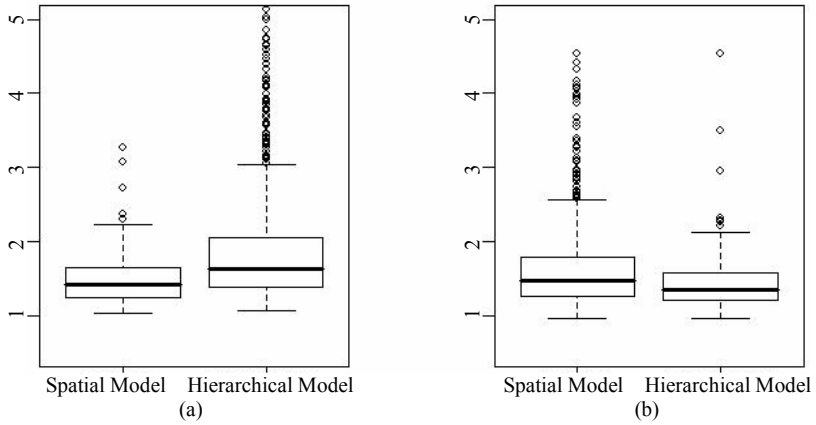

Figure 3 **Box plots of absolute relative error (ARE) for the cases: (a) data generated from the Spatial model and respectively fitted by the Spatial and Hierarchical models and (b) data generated from the Hierarchical model and respectively fitted by the Spatial and Hierarchical models**

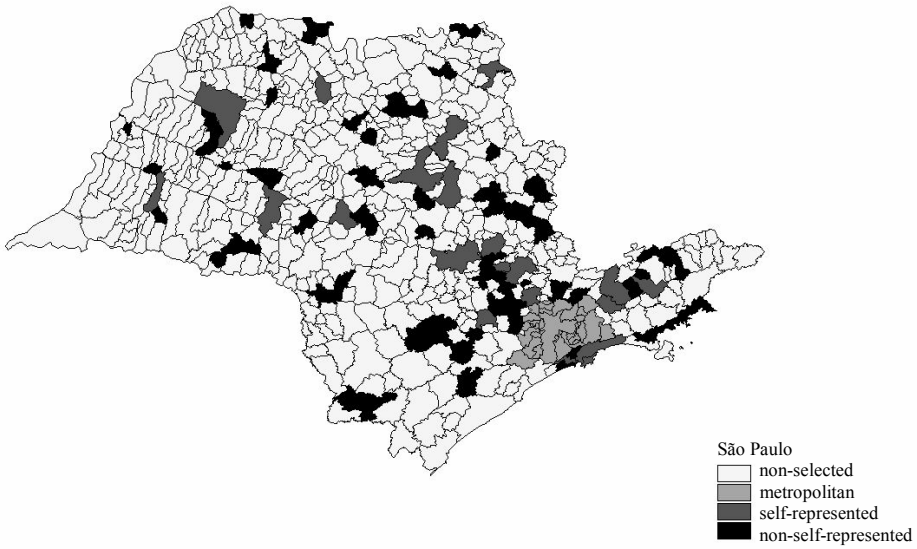

Figure 4 São Paulo municipalities sampled by the PNAD classified by the sampling definition

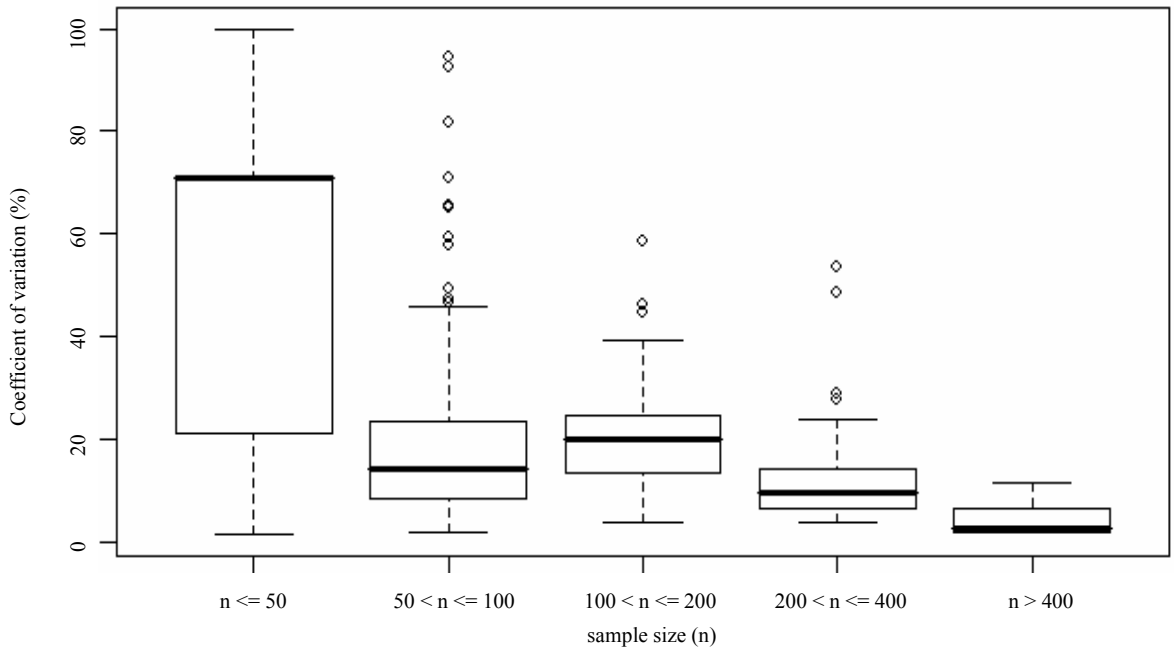

Figure 5 Boxplot of the coefficients of variation of the direct population estimates

4.4 Some results

We generated 20,000 samples after discarding the first 5,000. There is no evidence for non-convergence of the Hierarchical and the Spatial model parameters. A careful analysis of the MCMC outputs suggests that convergence was achieved for all model parameters. We summarize the results obtained by fitting the Hierarchical model (3) to the data provided by the PNAD survey. The posterior means of the model parameters were used as the point estimates. Table 5 presents these estimates together with the respective square root of the posterior variance. It can be seen from Table 5 that the estimate of η_1 is significantly positive, which agrees with what is expected by equation 4: the greater the sample size, the smaller σ^2_{it}.

Table 5
Summary of the model (2) parameter posterior distributions

parameter	posterior mean	posterior std
α	892.500	202.000
β	105.700	1.278
γ	0.072	0.008
η_0	10.620	0.133
η_1	3.185	0.484
τ^2_α	2.174E-7	2.961E-8
τ^2_γ	139.000	19.560

Figure 6 shows that the posterior means of the parameters α and γ that index the hierarchical model seem to be spatially distributed. The parameters of neighboring areas seem more alike than those of distant areas, which suggests applying the Spatial model.

4.5 Model selection

The Expected Prediction Deviance (EPD) (Gelfand and Ghosh 1998) measure was applied to help choose the most suitable model. The EPD measure is the sum of two terms. The first term, denoted by G, can be interpreted as a goodness-of-fit measure and the second term, denoted by P, as a penalty term for underfitted as well as overfitted models. The respective expressions for G and P are given by: $G=\sum_{i=1}^{m}\sum_{t=1}^{n}(y_{it}-E(y_{it}^{\text{rep}}|M))^2$ and $P=\sum_{i=1}^{m}\sum_{t=1}^{n}V(y_{it}^{\text{rep}}|M)$, where the expectations and the variances are with respect to the posterior predictive distribution associated with a future observation (y_{it}^{rep}) of y_{it} generated under the assumed model (M). According to this criterion, the smaller its value, the better the model. As can be seen in Table 6, the EPD criterion slightly favors the Spatial model.

4.6 Analysis of the results

The most disaggregated level for which the PNAD provides precise estimates is the metropolitan region, which is a set of contiguous municipalities. In order to validate the results obtained with the spatial model, population estimates for the greater São Paulo metropolitan region were

compared to the official statistics projections. The posterior distribution of $\mu_t = \sum_{i=1}^{r} \pi_{it} * A_i$ is easily obtained by adding $\mu_t^{(l)} = \sum_{i=1}^{r} \pi_{it}^{(l)} * A_i$ to the MCMC algorithm, where μ_t represents the total population of the metropolitan region at time t and r is the number of municipalities belonging to that metropolitan region.

Table 6
Measures for selecting models for demographic density

Model	*G*	*P*	EPD
Hierarchical	1.37E+09	6.14E+09	7.51E+09
Spatial	1.05E+09	6.19E+09	7.24E+09

Figure 7 compares the population estimates (μ_t) of the São Paulo metropolitan region obtained by the Spatial model and the official statistics. The solid lines represent the limits of the 95% credible intervals of μ_t, while the dotted line shows the respective point estimates. The symbol (+) represents the observed official statistics. It is noteworthy that some official statistics projections are outside of the credibles inferior limit (including the 1991 Census). This indicates that further investigations should be made in order to find out the reasons for these discrepancies. However, when we compare them at municipality level, the overall conclusion is that the model predictions and official statistics reasonable agree. The 95% credible intervals contains 92.4% of the official statistics projections. The average of the absolute relative error (ARE) between the estimated population density and the official statistics projection are 3%. These ARE measures are on average nearly the same for selected and non-selected municipalities.

Figure 8 compares the point estimates of the population sizes (μ_{it}) with the official projection statistics and the official census population sizes for a sampled municipality. The official projection methodology assumes that a set of small areas and a larger area, which contains them, have the same population growth rate pattern. The population of the larger area is projected by a component method and then proportionally allocated to the small areas. The component method uses data from the most recent census as well as the number of births and deaths and net migrations obtained from administrative records. The component method projects the population for a time t by adding the population in a previous time with the number of births and net migrations and subtracting the number of deaths in the same time interval.

The solid lines represent the 95% credible intervals for μ_{it} obtained by the Spatial model, while the dotted line shows the respective posterior means. The symbol (+) represents the official population projection for the inter-census period and the observed population in the census years. It is noteworthy that the point estimates are relatively close to the official projection statistics and the population obtained in the census year. This indicates that the use of the proposed model yields reliable estimates at municipality levels, with the extra advantage of providing a measure of the respective error.

We also analyze the estimates obtained for some municipalities not sampled in the PNAD. Figure 9 shows the model predictions, the 95% credible intervals, the official projection statistics and the observed population values in the censuses for a non-sampled municipality (+). It can be seen that the predictions obtained by the Spatial model reasonably agree with the official figures.

5. Final remarks

The model used in this article identifies the population growth trend of the municipalities. Reasonable estimates of the municipal populations are obtained for years with survey data, as well as for the years where census data are available. The point estimates have good precision and reasonably agree with estimates obtained for larger areas using other technique. The past information can be updated as soon as estimates become available from a new census or survey. Furthermore, the proposed approach provides the probability distribution of the quantity of interest, aiding the decision-making process.

Further work should be done in order to allow for autocorrelation of the parameters of interest over time. Extra information about the sampling variance estimates of the direct estimators could also be regarded as additional data. The assumption that the census coverage error is distributed symmetrically around zero could be relaxed by assigning a non-symmetric distribution to it. However a good knowledge of the shape of the distribution is required, which might be difficult in practice.

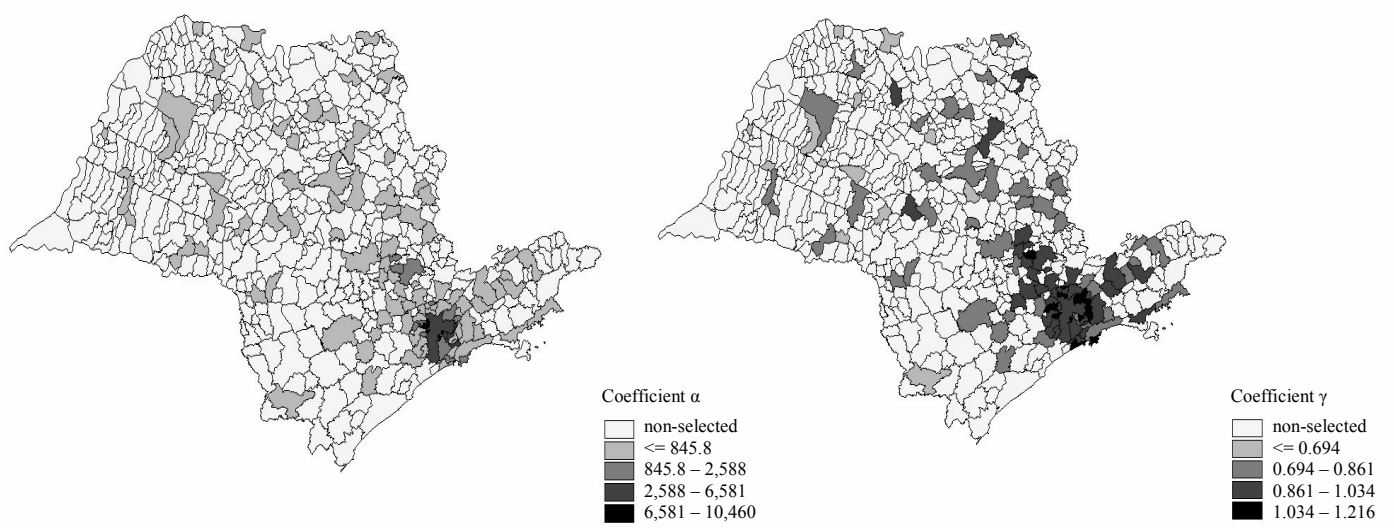

Figure 6 Posterior means of the parameters α and γ obtained by the hierarchical model

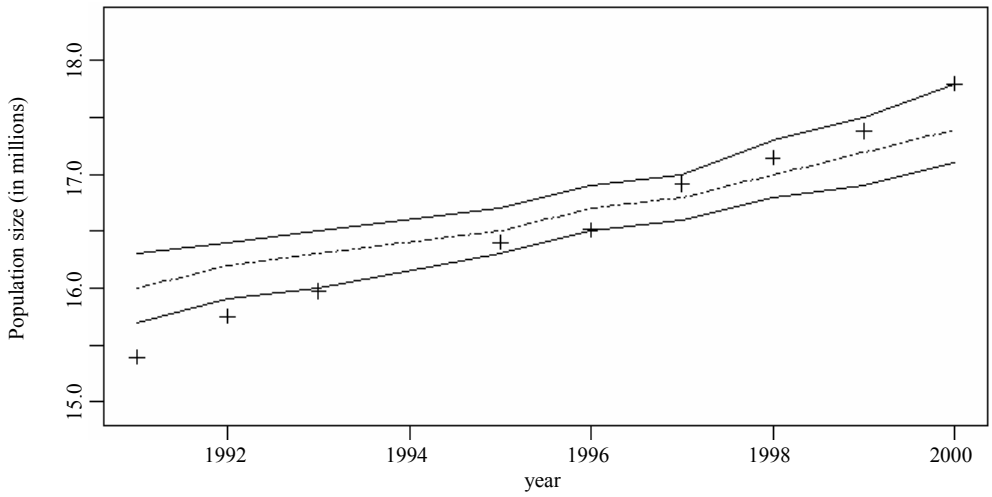

Figure 7 Comparison between the population sizes predicted by the spatial model and the official statistics (+) for the metropolitan region

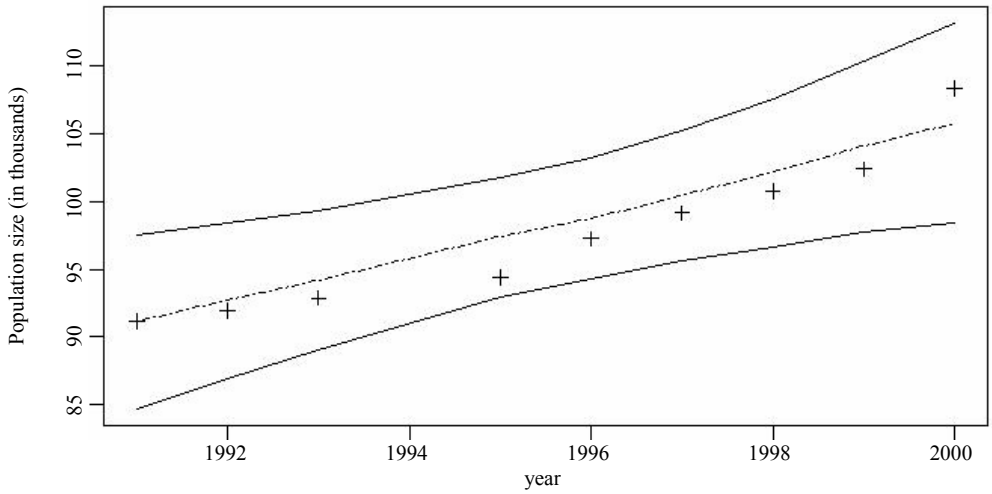

Figure 8 Comparison between the population sizes predicted by the spatial model and the official statistics (+) for a sampled municipality

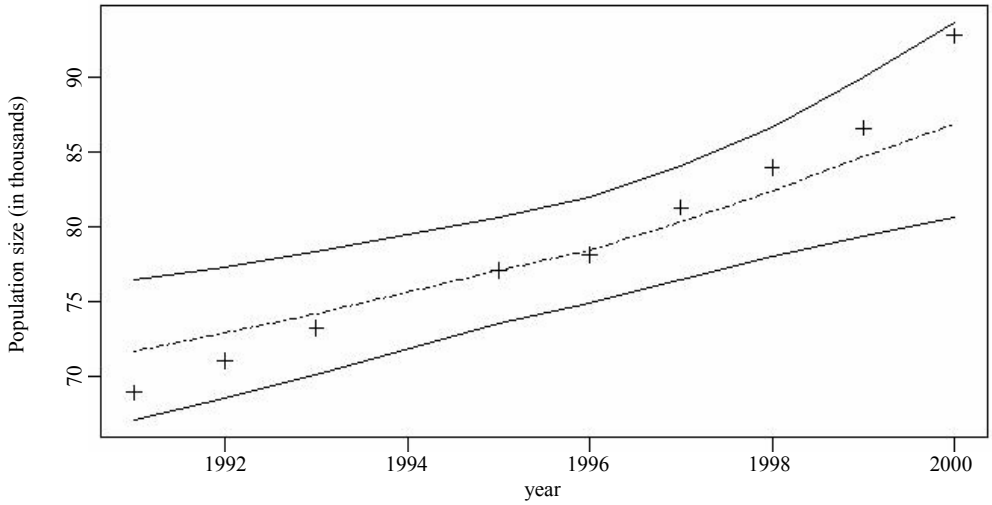

Figure 9 Population sizes predicted by the spatial model and the official statistics (+) for a non-sampled municipality

Acknowledgements

The authors thank an associate editor and two reviewers for their constructive comments and suggestions. The work of Fernando Moura and Helio Migon was funded in part by a research grant from the Brazilian National Council for the Development of Science and Technology (CNPq).

References

Besag, J., and Kooperang, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.

Brooks, S.P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 4, 434-455.

Gelfand, A.E., and Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85, 1, 1-11.

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

Migon, H., and Gamerman, D. (1993). Generalized exponential growth models: A bayesian approach. *Journal of Forecasting*, 12, 573-584.

Mollié, A. (1996). Bayesian mapping of disease. In: Gilks, W.R.; Richardson, S.; Spieghalter, D.J. Markov Chain Monte Carlo in Practice. New York: Chapman & Hall, 359-379.

Moura, F.A.S., and Migon, H.S. (2002). Bayesian spatial models for small area estimation of proportions. *Statistical Modeling*: *An International Journal*, 2, 3, 183-201.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review*, 70, 1, 125-143.

Pfeffermann, D., Moura, F.A.S. and Silva, P.L.N. (2006). Multi-level modeling under informative sampling. *Biometrika*, 93, 4, 943-959.

Pfeffermann, D., and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of American Statistical Association*, 102, 1427-1439.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology.

Souza, D.F. (2004). Estimação de População em Nível Municipal via Modelos Hierárquicos e Espaciais. Unpublished master's dissertation. Universidade Federal do Rio de Janeiro.

Spiegelharter, D.J., Thomas, A., Best, N. and Lunn, D. (2004). WinBUGS User Manual Version 1.4. MRC Biostatistics Unit, Cambridge.

Survey Methodology, december 2009
Vol. 35, No. 2, pp. 215-225
Statistics Canada, Catalogue No. 12-001-X

Variance estimation in the presence of nonrespondents and certainty strata

Jun Shao and Katherine J. Thompson [1]

Abstract

Business surveys often use a one-stage stratified simple random sampling without replacement design with some certainty strata. Although weight adjustment is typically applied for unit nonresponse, the variability due to nonresponse may be omitted in practice when estimating variances. This is problematic especially when there are certainty strata. We derive some variance estimators that are consistent when the number of sampled units in each weighting cell is large, using the jackknife, linearization, and modified jackknife methods. The derived variance estimators are first applied to empirical data from the Annual Capital Expenditures Survey conducted by the U.S. Census Bureau and are then examined in a simulation study.

Key Words: Covariate dependent nonresponse; Jackknife; Linearization; Ratio adjustment; Uniform nonresponse.

1. Introduction

Many business surveys use a one-stage stratified simple random sample without replacement design. Because of the skewness of the sampled populations, these designs generally include both certainty and non-certainty strata. With such designs, the sampling rates in the non-certainty strata are generally negligible (*e.g.*, less than 20 percent in all strata). However, if the ultimate sampling unit is large business entity such as a company, the size of the universe is much smaller and often sampling fractions should not be ignored in computation of variance estimates.

Most surveys have nonresponse. We consider surveys using weighting adjustment for nonresponse. For certainty strata, there is no sampling error and, hence, standard variance formulas do not include any component for certainty strata. When nonresponse is present, however, there is an estimation error even in a certainty stratum, which is often an appreciable component of the total estimation error.

The purpose of this paper is to develop some methods for variance estimation that take into account the weighting adjustment for nonresponse and the existence of certainty strata. After introducing notation and assumptions in Section 2, we show that the jackknife and linearization variance estimators ignoring nonresponse in certainty strata, which are often currently used in many surveys, underestimate the true variance of the weight adjusted estimated population total. By directly deriving an approximate variance formula, we obtain two consistent variance estimators. These variance estimators are also consistent if there are non-certainty strata with large sampling fractions. A modified jackknife variance estimator taking into account the variability due to nonresponse in certainty strata is also derived.

In Section 3, we compare variance estimators using five years' of data from the Annual Capital Expenditures Survey (ACES) conducted by the U.S. Census Bureau. Simulation results are presented in Section 4 using a population generated from 2003 ACES data. Our simulation results show that the variance estimators ignoring certainty strata have large negative biases; the derived consistent variance estimators perform well when stratum sample sizes are all large and perform inconsistently otherwise; and the jackknife variance estimator ignoring all sampling fractions overestimates. Some concluding remarks are given in Section 5.

2. Main results

Consider a stratified sample without replacement from a finite population containing H strata. Let n_h and N_h be the sample and population size of stratum h, respectively, y_{hj} be a variable of interest that may have nonresponse, and x_{hj} be a covariate that takes positive values and does not have nonresponse, where j is the index of population unit and h is the index for stratum. Using the sample-response path considered by Fay (1991) and Shao and Steel (1999), we view the finite population as a census with y, x values and nonrespondents, *i.e.*, each unit j in stratum h of the finite population is associated with an indicator I_{hj} (= 1 if y_{hj} is a respondent and = 0 if y_{hj} is a nonrespondent). Our sample is taken from this finite population, and if unit j in stratum h is in the sample, y_{hj} is a respondent if $I_{hj} = 1$ and a nonrespondent if $I_{hj} = 0$.

Let E_s and V_s be the expectation and variance, respectively, with respect to sampling and E_m, V_m, and P_m be the expectation, variance, and probability, respectively,

1. Jun Shao, University of Wisconsin-Madison and U.S. Census Bureau; Katherine J. Thompson, U.S. Census Bureau. E-mail: shao@stat.wisc.edu.

with respect to the model m specified in one of the following assumptions.

Assumption M. Values of (y_{hj}, x_{hj}, I_{hj}) in the finite population are independently generated from a superpopulation model m. The finite population is divided into P sub-populations such that, within sub-population p, the response probability $P_m(I_{hj} = 1 \mid y_{hj}, x_{hj}) = P_m(I_{hj} = 1 \mid x_{hj}) > 0$, $E_m(y_{hj} \mid x_{hj}) = \beta_p x_{hj}$, and $V_m(y_{hj} \mid x_{hj}) = \sigma_p^2 x_{hj}$, where β_p and σ_p are unknown parameters depending on p.

Assumption P. The finite population is divided into P sub-populations such that, under a superpopulation model, $P_m(I_{hj} = 1 \mid y_{hj}, x_{hj}) = \pi_p > 0$ is constant within sub-population p.

The sub-population in Assumption M or Assumption P is called nonresponse adjustment weighting cell (or weighting cell for short), since we handle nonrespondents by weight adjustment within each weighting cell. (If imputation is applied within each sub-population, then sub-populations are called imputation cells.) In applications, weighting cells may be strata, or unions of strata (strata are collapsed when they have insufficient respondents), or may cut across strata. Assumption M involves a prediction model between y_{hj} and x_{hj} and a covariate-dependent response mechanism within each weighting cell. The response mechanism under Assumption P is the within-weighting-cell uniform response mechanism and is often referred to as the quasi-random response model. Assumption P is stronger than Assumption M in terms of the response mechanism. However, Assumption M requires an explicit model between y_{hj} and x_{hj} within each weighting cell. In this paper we assume either Assumption M or Assumption P. Estimators that can be justified under Assumption P are referred to as the "quasi-randomization" estimators (Oh and Scheuren 1983).

When we study asymptotic consistency of estimators, we consider the limiting process of $k_p \to \infty$ for all p with fixed H and P, where k_p is the sample size in weighting cell p. If weighting cells are the same as strata or unions of strata, then $k_p \to \infty$ is the same as $n_h \to \infty$ for all h.

After the ratio-adjustment for nonresponse, we consider the following estimator of the total of y-values in the finite population:

$$\hat{Y} = \sum_p \sum_h \sum_{j \in s_h} \left(\frac{\hat{X}_p}{\hat{X}_{pr}} w_{hj} \right) \delta_{phj} I_{hj} y_{hj} = \sum_p \frac{\hat{X}_p}{\hat{X}_{pr}} \hat{Y}_{pr}, \quad (1)$$

where p is the index for weighting cell, s_h is the sample in stratum h, δ_{phj} is the indicator for the weighting cell p, and w_{hj} is the survey weight constructed for the stratified sampling,

$$\hat{X}_p = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} x_{hj}, \quad \hat{X}_{pr} = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} I_{hj} x_{hj},$$

and

$$\hat{Y}_{pr} = \sum_h \sum_{j \in s_h} w_{hj} \delta_{phj} I_{hj} y_{hj}.$$

In the special case where weighting cells are the same as strata,

$$\hat{Y} = \sum_h \frac{\hat{X}_h}{\hat{X}_{hr}} \hat{Y}_{hr}, \quad (2)$$

where

$$\hat{X}_h = \sum_{j \in s_h} w_{hj} x_{hj}, \quad \hat{X}_{hr} = \sum_{j \in s_h} w_{hj} x_{hj} I_{hj},$$

and

$$\hat{Y}_{hr} = \sum_{j \in s_h} w_{hj} y_{hj} I_{hj}.$$

When the covariate $x_{hj} \equiv 1$, $\hat{Y}$ is referred to as the count estimator. The count estimator controls respondent estimates to frame population totals. When the weighting cells are the same as strata, the count estimator uses the unweighted cell response rates, as recommended in Vartivarian and Little (2002).

Under Assumption M or P,

$$E_m E_s(\hat{Y} - Y) = E_s E_m(\hat{Y} - Y) = 0,$$

where Y is the finite population total of y values, and the total variance

$$V_{m,s}(\hat{Y} - Y) = E_m[V_s(\hat{Y})] + V_m[E_s(\hat{Y}) - Y].$$

Let $V_1 = E_m[V_s(\hat{Y})]$ and $V_2 = V_m[E_s(\hat{Y}) - Y]$. To estimate V_1, it suffices to estimate the sampling variance $V_s(\hat{Y})$. Since $\hat{Y}$ defined by (1) is a sum of ratios and each of $\hat{X}_p$, $\hat{X}_{pr}$, and $\hat{Y}_{pr}$ is a weighted total of variables and indicators, we can apply the stratified jackknife variance estimator

$$v_{J1} = \sum_h \left(1 - \frac{n_h}{N_h}\right) \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left(\hat{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \hat{Y}_{(hk)} \right)^2 \quad (3)$$

(see Wolter 1985 or Shao and Tu 1995), where $\hat{Y}_{(hj)}$ is the jackknife analog of $\hat{Y}$ when unit j in stratum h is deleted. Note that sampling fractions are incorporated in this formula. When $k_p \to \infty$ for all weighting cells, the standard result for the complete data case (see, *e.g.*, Krewski and Rao 1981) implies that the jackknife estimator v_{J1} is consistent for the sampling variance $V_s(\hat{Y})$, under Assumption M or P. Since V_1 is the expectation of $V_s(\hat{Y})$, v_{J1} is also consistent for V_1 under some minor conditions.

Since the function in (1) is the sum of ratios and data in different weighting cells are independent, a linearization estimator of $V_s(\hat{Y})$ can be derived using Taylor's expansion.

When weighting cells are the same as strata, for example, $\hat{Y}$ is given by (2) and is a separate ratio estimator whose linearization variance estimator can be obtained using standard techniques. An alternative way to derive a linearization variance estimator is to linearize the jackknife estimator v_{J1} (Thompson and Yung 2006). The resulting estimator is

$$v_{L1} = \sum_h \frac{n_h}{n_h - 1} \sum_{j \in s_h} \left\{ \sum_p \left[\frac{\hat{X}_p}{\hat{X}_{pr}} (\overline{e}_{ph} - w_{hj} e_{phj} I_{hj} \delta_{phj}) + \frac{\hat{Y}_{pr}}{\hat{X}_{pr}} (\overline{x}_{ph} - w_{hj} x_{hj} \delta_{phj}) \right] \right\}^2, \quad (4)$$

where $e_{phj} = y_{hj} - (\hat{Y}_{pr}/\hat{X}_{pr}) x_{hj}$, $\overline{e}_{ph} = n_h^{-1} \sum_{j \in s_h} w_{hj} e_{phj} I_{hj} \delta_{phj}$, and $\overline{x}_{ph} = n_h^{-1} \sum_{j \in s_h} w_{hj} x_{hj} \delta_{phj}$. The estimator in (4) is exactly the same as the standard linearization variance estimator for the separate ratio estimator in (2) when weighting cells are the same as strata. Like v_{J1}, v_{L1} is consistent for V_1 when $k_p \to \infty$ under Assumption M or P, which follows from the standard result for the complete data case (Krewski and Rao 1981).

Since ratio is a smooth function, under Assumption M or P,

$$E_s(\hat{Y}) = \sum_p E_s \left(\frac{\hat{X}_p \hat{Y}_{pr}}{\hat{X}_{pr}} \right) \approx \sum_p \frac{E_s(\hat{X}_p) E_s(\hat{Y}_{pr})}{E_s(\hat{X}_{pr})} = \sum_p \frac{X_p Y_{pr}}{X_{pr}},$$

where

$$X_p = \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} x_{hj},$$

$$X_{pr} = \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} x_{hj},$$

$$Y_{pr} = \sum_h \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} y_{hj},$$

and $\mathcal{P}_h$ is the finite population in stratum h. Let Y_p be the same as X_p with x_{hj} replaced by y_{hj}. Then

$$V_2 = V_m[E_s(\hat{Y}) - Y] \approx \sum_p V_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right).$$

Note that V_2 is small if the nonresponse rate is low ($V_2 = 0$ if there is no nonresponse) or if the model under Assumption M is highly predictive. If the overall sampling fraction, $\sum_h n_h / \sum_h N_h$, converges to 0, then V_2/V_1 converges to 0 and, hence v_{L1} and v_{J1} are consistent estimators of the total variance $V_{m,s}(\hat{Y}) = V_1 + V_2 \approx V_1$. Note that V_1 does not contain the variation from certainty strata due to nonresponse. Because the y-values from certainty strata are influential in the total Y in many surveys, and because in applications it is difficult to tell how small $\sum_h n_h / \sum_h N_h$ has to be for the convergence $V_2/V_1 \to 0$ to take place, it is necessary to estimate V_2.

Under Assumption M, let $\tilde{E}_m, \tilde{V}_m$, and $\tilde{C}_m$ be the conditional expectation, variance, and covariance, respectively, given all x-values and response indicators. Since

$$\tilde{E}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) = 0,$$

we obtain

$$\begin{aligned} V_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) &= E_m \left[\tilde{V}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] + V_m \left[\tilde{E}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] \\ &= E_m \left[\tilde{V}_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] \\ &= E_m \left[\frac{X_p^2}{X_{pr}^2} \tilde{V}_m(Y_{pr}) - 2 \frac{X_p}{X_{pr}} \tilde{C}_m(Y_{pr}, Y_p) + \tilde{V}_m(Y_p) \right] \\ &= E_m \left[\frac{X_p^2}{X_{pr}^2} \sigma_p^2 X_{pr} - 2 \frac{X_p}{X_{pr}} \sigma_p^2 X_{pr} + \sigma_p^2 X_p \right] \\ &= \sigma_p^2 E_m \left(\frac{X_p^2}{X_{pr}} - X_p \right). \end{aligned}$$

Under Assumption P, let V_m^I be the variance with respect to I_{hj}'s. Since

$$E_m^I \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \approx 0,$$

we obtain

$$\begin{aligned} V_m \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) &\approx E_m \left[V_m^I \left(\frac{X_p Y_{pr}}{X_{pr}} - Y_p \right) \right] \\ &\approx E_m \left[\frac{1 - \pi_p}{\pi_p} \sum_h \sum_{j \in P_h} \delta_{hj} \left(y_{hj} - \frac{Y_p}{X_p} x_{hj} \right)^2 \right] \\ &\approx E_m \left[\left(\frac{X_p^2}{X_{pr}} - X_p \right) S_p^2 \right], \end{aligned}$$

where

$$S_p^2 = \frac{1}{X_{pr}} \sum_h \sum_{j \in \mathcal{P}_h} \delta_{hj} \left(y_{hj} - \frac{Y_p}{X_p} x_{hj} \right)^2.$$

Since X_p and X_{pr} can be estimated by $\hat{X}_p$ and $\hat{X}_{pr}$, respectively, to estimate V_2 we only need to find an estimator of σ_p^2 or S_p^2. Under Assumption M, a regression estimator of β_p is $\hat{Y}_{pr}/\hat{X}_{pr}$ and a consistent estimator of σ_p^2 based on regression residuals is

$$\hat{\sigma}_p^2 = \frac{1}{\hat{X}_{pr}} \sum_h \sum_{j \in s_h} \delta_{phj} I_{hj} w_{hj} \left(y_{hj} - \frac{\hat{Y}_{pr}}{\hat{X}_{pr}} x_{hj} \right)^2.$$

From the theory of sampling, $\hat{\sigma}_p^2$ is also a consistent estimator of S_p^2 under Assumption P. Hence, under Assumption M or P, a consistent estimator of V_2 is

$$v_{L2} = \sum_p \hat{\sigma}_p^2 \left(\frac{\hat{X}_p^2}{\hat{X}_{pr}} - \hat{X}_p \right). \tag{5}$$

The subscript L indicates that this estimator is based on linearization.

In some applications $\sum_h n_h / \sum_h N_h$ is negligible and nonresponse in noncertainty strata has negligible contribution to the variance component V_2, *i.e.*,

$$V_2 \approx V_m \left[\sum_p \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} - Y_{cp} \right) \right], \tag{6}$$

where the subscript c stands for certainty strata,

$$X_{cp} = \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{P}_h} \delta_{phj} x_{hj}, \quad X_{cpr} = \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} x_{hj},$$
$$Y_{cp} = \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{P}_h} \delta_{phj} y_{hj}, \quad Y_{cpr} = \sum_{h \in \mathcal{C}} \sum_{j \in \mathcal{P}_h} \delta_{phj} I_{hj} y_{hj},$$

and $\mathcal{C}$ is the collection of indices of certainty strata. A consistent jackknife estimator of V_2 can be obtained as follows. Note that X_{cp}, X_{cpr}, and Y_{cpr} are estimators, since $\mathcal{P}_h = s_h$ for $h \in \mathcal{C}$, but Y_{cp} is not an estimator because of nonresponse. Thus, we cannot apply the jackknife to the function $X_{cp} Y_{cpr} / X_{cpr} - Y_{cp}$. From the previous derivation we note that, under Assumption M,

$$V_2 \approx E_m \tilde{V}_m \left[\sum_p \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} - Y_{cp} \right) \right]$$
$$= E_m \left[\sum_p \left(1 - \frac{X_{cpr}}{X_{cp}} \right) \tilde{V}_m \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} \right) \right].$$

Similarly, under Assumption P, the result holds with $\tilde{V}_m$ replaced by V_m^I. Hence, we can apply the jackknife to the estimator $X_{cp} Y_{cpr} / X_{cpr}$. Let

$$\breve{Y} = \sum_p \sqrt{1 - \frac{X_{cpr}}{X_{cp}}} \left(\frac{X_{cp} Y_{cpr}}{X_{cpr}} \right)$$

and $\breve{Y}_{(hj)}$ be the jackknife analog of $\breve{Y}$ after unit j in $h \in \mathcal{C}$ is deleted, when we treat $X_{cp} Y_{cpr} / X_{cpr}$ as estimators. Then a jackknife estimator of V_2 is

$$v_{J2} = \sum_{h \in C} \frac{N_h - 1}{N_h} \sum_{j \in \mathcal{P}_h} \left(\breve{Y}_{(hj)} - \frac{1}{N_h} \sum_{k \in P_h} \breve{Y}_{(hk)} \right)^2$$

($n_h = N_h$ and $s_h = \mathcal{P}_h$ when $h \in \mathcal{C}$). The factor $\sqrt{1 - X_{cpr} / X_{cp}}$ in the formula for $\tilde{Y}$ makes the appropriate adjustment for nonresponse. Under Assumption P, $X_{cpr} / X_{cp} \approx \pi_p$ is the response rate, which can be view as a "sampling" fraction for certainty strata.

The resulting jackknife estimator of the total variance $V_1 + V_2$ is then $v_{J1} + v_{J2}$. Since $n_h = N_h$ (*i.e.*, $1 - n_h / N_h = 0$) if stratum h is a certainty stratum, it is easy to see that $v_{J1} + v_{J2}$ is equal to

$$v_J = \sum_h \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left(\tilde{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \tilde{Y}_{(hk)} \right)^2, \tag{7}$$

where

$$\tilde{Y}_{(hj)} = \begin{cases} \breve{Y}_{(hj)} & \text{if stratum } h \text{ is a certainty stratum} \\ \hat{Y}_{(hj)} \sqrt{1 - \dfrac{n_h}{N_h}} & \text{if stratum } h \text{ is not a certainty stratum.} \end{cases}$$

Compared with the jackknife variance estimator v_{J1} in (3), v_J in (7) addresses the variability due to nonresponse in certainty strata, whereas v_{J1} does not. Under (6) and Assumption M or P, v_J is consistent.

Finally, the jackknife estimator that ignores all sampling fractions is:

$$\tilde{v}_J = \sum_h \frac{n_h - 1}{n_h} \sum_{j \in s_h} \left(\hat{Y}_{(hj)} - \frac{1}{n_h} \sum_{k \in s_h} \hat{Y}_{(hk)} \right)^2. \tag{8}$$

This estimator seems to be conservative, although it is not theoretically justified.

In summary, we have the following estimators of the total variance $V_{m,s}(\hat{Y})$:

1. The jackknife estimator v_{J1} defined in (3), which underestimates when V_2 / V_1 is not negligible.
2. The linearization estimator v_{L1} defined in (4), which is asymptotically equivalent to v_{J1}.
3. $v_L = v_{L1} + v_{L2}$ with v_{L2} is defined in (5), which is consistent.
4. $v_{JL} = v_{J1} + v_{L2}$, which is asymptotically equivalent to v_L.
5. The jackknife variance estimator v_J defined in (7), which is consistent when (6) holds.
6. The jackknife estimator $\tilde{v}_J$.

Under stratified simple random sampling and Assumption P, v_L is approximately the same as the variance estimator obtained by treating the set of respondents as an additional phase of the stratified simple random sample (*i.e.*, a two-phase sample design) and applying standard variance formula (when $x_{hj} \equiv 1$) or the variance formula for calibration estimators (Kott 1994, Särndal, Swensson and Wretman 1992, and Hidiroglou and Särndal 1998). This variance estimator, however, is not consistent when Assumption P does not hold.

3. Empirical comparisons

In this section, we apply the variance estimators described in Section 2 to five years of empirical data from the employer component of the ACES introduced in Section 1. Section 3.1 provides background on the ACES analysis variables, sample design, and estimation procedures. Section 3.2 presents the empirical comparisons.

3.1 Background of ACES

The ACES collects data about the nature and level of capital expenditures in non-farm businesses operating within the United States. Respondents report capital expenditures, broken down by type (expenditures on Structures and expenditures on Equipment) for the calendar year in all subsidiaries and divisions for all operations within the United States.

The ACES universe contains two sub-populations: employer companies (ACE-1) and non-employer (ACE-2) companies. (A nonemployer company is one that has no paid employees, has annual business receipts of $1,000 or more ($1 or more in the construction industries), and is subject to federal income taxes. Most nonemployers are self-employed individuals operating very small un-incorporated businesses, which may or may not be the owner's principal source of income). Different forms are mailed to sample units depending on whether they are ACE-1 companies or ACE-2 companies. New ACE-1 and ACE-2 samples are selected each year, both with stratified simple random sample without replacement designs. The ACE-1 sample comprises approximately seventy-five percent of the ACES sample (roughly 46,000 companies selected per year for ACE-1, and 15,000 selected per year for ACE-2). In the ACE-1 design, units are stratified into size-class strata within each industry on the sampling frame. There are five separate ACE-1 strata in each industry, consisting of one certainty stratum (referred to as stratum 10) and four non-certainty strata defined by company size within industry (denoted by 2A through 2D, ranked from largest to smallest within industry), with approximately 500 non-certainty strata in each year's design. Sampling fractions in the large-size class-within-industry strata (2A) can be fairly high: in most years, approximately 55% of the sample in 2A strata are sampled at rates between 0.5 and 1. Sampling fractions in the other three size class within-industry strata are usually less than 0.20. Design weights range from 1 to 1,000, depending on industry and size-class strata. The ACE-2 component is much less highly stratified, with between a total of six to eight size-class strata used each year, and sampling fractions less than 0.01 in all strata. Our empirical analysis is restricted to the ACE-1 component of the survey, which meets all of the conditions described in the previous section.

The ACES publishes total and year-to-year change estimates. Estimates are published for the entire survey, and by industry code as indicated by the respondent units (not necessarily the industry code on the sampling frame). If there is no nonresponse, variances are estimated using the delete-a-group jackknife variance estimator (Kott 2001). To account for unit nonresponse, the ACE-1 component uses the ratio-adjustment procedure presented in Section 2 with administrative payroll data as the auxiliary variable x. Weighting cells are the design strata, provided that there is at least one respondent in the cell. Cell collapsing is extremely rare and is hereafter ignored in this paper. More details concerning the ACES survey design, methodology, and data limitations are available on-line at http://www.census.gov/csd/ace.

Although the ACE-1 survey design is fairly typical for a business survey, the collected data are not. Smaller companies often report legitimate values of zero for capital expenditures, and consequently the majority of the estimates are often obtained from the certainty and large non-certainty (2A) companies. As the capital expenditures are further cross-classified, the incidence of reported zeros (especially among smaller companies) increases.

3.2 Comparisons

To assess the effect of the unit non-response weight adjustment procedure on the ACE-1 standard errors, we computed variance estimates from unit nonresponse adjusted ACE-1 data using the ratio estimator with payroll as the auxiliary variable, in four industries, each with high sampling rates in the large company non-certainty strata (2A). The selected industries represent a cross-section of the sectors represented in the ACES. These industries and their North American Industrial Classification System (NAICS) codes are: Oil and Gas Extraction (211100), Nonmetallic Mineral Mining and Quarrying (212300), Other Miscellaneous Manufacturing (339900), and Architectural, Engineering, and Related Services (541300). In subsequent tables and discussions, industries are referred to by their NAICS code.

Table 1 presents variance estimate comparisons using five years' of ACE-1 survey data for three characteristics: the total capital expenditures (Total), capital expenditures on structures (Structures), and capital expenditures on equipment (Equipment). For comparison, the variance estimates are presented as a ratio to v_{J1} in Table 1. The estimated totals are also included. (Note that these totals are not the same as the published estimates, since they are computed using the industry classification on the frame, not the industry classification provided by the respondent).

As expected, the jackknife estimator v_{J1} and the linearization jackknife estimator v_{L1} are very close for all

variables. The consistent variance estimators (v_L and v_{JL}) are all noticeably larger than their corresponding jackknife counterparts (v_{L1} and v_{J1}). In general, most capital expenditures are reported by certainty or large non-certainty companies, so effect on variance estimation of including non-respondent component in the variance estimator is noticeable. The jackknife estimator v_J, which adjusts for the effect of certainty strata, is generally between v_{J1} and v_{JL}. In some cases, v_J is equal to or very close to v_{J1}, indicating that the variability due to nonresponse mainly comes from non-certainty strata with large sampling fractions. The jackknife estimate $\tilde{v}_J$, which ignores sampling fractions, is much larger than any other estimates.

4. Simulation results

In this section, we present a simulation study using data modeled from the ACE-1 industries presented in the previous section. Section 4.1 describes the simulation settings. Section 4.2 presents and summarizes the results.

4.1 Simulation settings

We modeled our population using respondent data from the 2003 data collection of the three key items collected by the survey (Total, Structures, and Equipment). Frame data for the auxiliary variable (payroll) were available for all units. The complete population data were generated using the SIMDAT algorithm (Thompson 2000) with modeling cells equal to sampling strata and population size equal to the original frame size in each cell. Table 2 provides sampling fractions and correlation coefficients with the payroll for the modeled data in each stratum.

In the simulation, stratified simple random samples were selected from the generated population. We examine the statistical properties of the six variance estimators described in Section 2 over repeated samples under the following two different response mechanisms applied to the sample data:

1. The covariate-dependent response mechanism obtained by randomly applying response propensities modeled from the survey data with payroll as the covariate, which yields very high probabilities of responding to the large units and very small probabilities to the small (non-certainty) units;
2. The within-stratum uniform response mechanism obtained by using the observed survey response rate as the within-stratum response probability.

On the average, response probabilities in the individual stratum within industry were 0.85, 0.76. 0.77, 0.76, and 0.68 for strata 10, 2A, 2B, 2C, and 2D, respectively.

We selected 5,000 samples from the population, computed $\hat{Y}$ in (1) from each sample with nonresponse and weight adjustment, and computed the empirical mean and variance of the 5,000 $\hat{Y}$ values. This was done for each industry and each item, with two adjustment methods: the ratio and count estimators. When $\hat{Y}$ is the ratio estimator using the payroll as the auxiliary variable, the absolute value of the empirical relative bias is under 1.4% and is smaller than 1% in most cases. For the count estimator under the within-stratum uniform response mechanism, its absolute value of the empirical relative bias is under 0.5%. The count estimator is not approximately unbiased in theory under the covariate-dependent response mechanism. In the simulation, however, its absolute value of the empirical relative bias is under 1% in most cases and has a maximum value of 2.7%. The empirical variance of the 5,000 $\hat{Y}$ values was used as the "true value" of the variance of $\hat{Y}$.

4.2 Results

In 2,000 of the 5,000 samples, we computed the six different variance estimates for all three items, four industries, and two weight adjustment methods. We examined the statistical properties of each of variance estimator over repeated samples using the relative bias (RB) defined as

$$\frac{\text{the average of } 2,000 \text{ variance estimates}}{\text{the true variance}} - 1,$$

the stability (ST) defined as

$$\frac{\sqrt{\text{the empirical mean squared error of variance estimate}}}{\text{the true variance}},$$

and the error rate (ER) defined as the empirical proportion of the approximate 90% confidence intervals ($\hat{Y} \pm 1.645\sqrt{\text{variance estimate}}$) from 2,000 samples that do not contain the true population total.

Tables 3 and 4 respectively report the simulation results under the two response mechanisms. The results from these tables can be summarized as follows.

1. Two variance estimators ignoring V_2, v_{J1} and v_{L1}, have large negative relative biases in general. The error rates of the related confidence intervals are also large.
2. Two consistent variance estimators, v_L and v_{JL}, have very similar performances and are generally much better than v_{J1} and v_{L1} in terms of the relative bias and the error rate of the related confidence intervals.
3. The jackknife variance estimator v_J performs well in industries 339900 and 541300, but may have large positive relative biases in industries 211000 and 212300. We think that this is a "small sample" effect, since v_J is justified by asymptotic consistency and the sizes of the certainty strata in industries 211000 and 212300 are 26 and 30, respectively (Table 2). The sizes of the certainty

strata for the other two industries are 158 and 160, respectively. In fact, the performance of v_L and v_{JL} is generally better in industries 339900 and 541300.

4. In some cases v_J has more than 10% negative relative biases, which is caused by the fact that some non-certainty strata have large sampling fractions, *i.e.*, the approximation (6) does not hold enough.
5. The jackknife variance estimator $\tilde{v}_J$ ignoring all sampling fractions has very large positive relative biases and is too conservative.

5. Concluding remarks

When nonresponse is present in certainty strata (or strata with large sampling fractions), the jackknife and the linearization variance estimators that ignore certainty strata (or strata with large sampling fractions) are not acceptable because of their large negative biases. We derive two asymptotically unbiased and consistent variance estimators by adding an extra term that accounts the variability from nonresponse in certainty strata (or strata with large sampling fractions). We also derive a modified jackknife estimator that is consistent when the certainty strata are the only strata that contribute to the variance due to nonresponse (*i.e.*, Assumption (6) holds).

Our simulation results show that the three derived variance estimators perform well when stratum sample sizes are all large and perform inconsistently otherwise, and that the jackknife variance estimator that ignores all sampling fractions is very conservative.

Compared with the linearization method, the jackknife requires more computational resources but it has other advantages such as being easy to program, using a single recipe for different problems, and not requiring complicated or separate derivations for different estimators. Our linearization variance estimator given in (4) is in fact obtained by linearizing the jackknife estimator in (3).

Table 1
Variance estimates for $\hat{Y}$ with ratio adjustment in ACE-1 survey

Industry	Item	Year	$\hat{Y}$	v_{J1}	$\frac{v_{L1}}{v_{J1}}$	$\frac{v_L}{v_{J1}}$	$\frac{v_{JL}}{v_{J1}}$	$\frac{v_J}{v_{J1}}$	$\frac{\tilde{v}_J}{v_{J1}}$
211000	Total	2002	1.63E+7	4.63E+11	0.97	1.14	1.17	1.00	17.3
		2003	2.28E+7	6.87E+12	0.95	1.21	1.26	1.00	2.81
		2004	2.30E+7	2.45E+12	0.98	1.23	1.25	1.00	4.77
		2005	3.08E+7	4.29E+12	0.98	1.22	1.24	1.19	4.77
		2006	4.18E+7	6.29E+12	0.99	1.17	1.19	1.00	8.78
	Structures	2002	1.31E+7	3.99E+11	0.97	1.14	1.17	1.00	15.3
		2003	1.86E+7	5.78E+12	0.94	1.22	1.27	1.00	2.78
		2004	1.70E+7	8.39E+11	0.99	1.42	1.43	1.00	11.3
		2005	2.64E+7	3.84E+12	0.98	1.22	1.24	1.16	4.64
		2006	3.55E+7	5.41E+12	0.99	1.19	1.21	1.00	8.76
	Equipment	2002	3.20E+6	6.14E+10	0.98	1.15	1.17	1.00	7.26
		2003	4.18E+6	8.39E+11	0.97	1.22	1.24	1.00	1.70
		2004	6.01E+6	1.54E+12	0.97	1.13	1.16	1.00	1.39
		2005	4.33E+6	1.34E+11	0.97	1.22	1.25	1.15	6.17
		2006	6.31E+6	7.14E+11	0.99	1.12	1.13	1.00	2.68
212300	Total	2002	1.56E+6	4.14E+10	0.81	1.06	1.24	1.20	3.19
		2003	1.33E+6	1.21E+10	0.94	1.18	1.24	1.36	5.43
		2004	2.01E+6	2.86E+10	0.96	1.60	1.65	2.20	6.04
		2005	1.96E+6	1.93E+10	0.98	1.12	1.14	2.30	6.04
		2006	2.28E+6	2.19E+10	0.96	1.26	1.30	3.22	11.7
	Structures	2002	2.22E+5	4.36E+8	1.00	1.11	1.11	1.64	8.61
		2003	1.49E+5	2.27E+8	0.96	1.28	1.32	1.48	7.32
		2004	4.14E+5	1.03E+8	0.96	46.6	46.6	75.3	426
		2005	2.23E+5	9.33E+8	0.99	1.12	1.13	1.32	1.88
		2006	2.20E+5	1.88E+9	0.97	1.20	1.22	1.19	2.29
	Equipment	2002	1.33E+6	4.05E+10	0.81	1.06	1.25	1.15	2.86
		2003	1.18E+6	1.13E+10	0.94	1.20	1.26	1.32	5.07
		2004	1.60E+6	2.82E+10	0.96	1.40	1.44	1.53	3.30
		2005	1.73E+6	1.62E+10	0.97	1.16	1.19	2.33	6.69
		2006	2.06E+6	2.14E+10	0.96	1.26	1.30	2.94	10.8

Table 1 (continued)
Variance estimates for $\hat{Y}$ with ratio adjustment in ACE-1 survey

Industry	Item	Year	$\hat{Y}$	v_{J1}	$\frac{v_{L1}}{v_{J1}}$	$\frac{v_{L}}{v_{J1}}$	$\frac{v_{JL}}{v_{J1}}$	$\frac{v_{J}}{v_{J1}}$	$\frac{\tilde{v}_{J}}{v_{J1}}$
339900	Total	2002	1.75E+6	1.94E+10	0.99	1.27	1.29	1.10	3.71
		2003	1.58E+6	2.99E+10	0.98	1.24	1.27	1.10	1.60
		2004	1.70E+6	1.00E+10	0.99	1.40	1.40	1.69	4.61
		2005	1.77E+6	2.55E+10	0.99	1.28	1.29	1.25	3.02
		2006	1.94E+6	5.51E+10	0.99	1.23	1.25	1.12	2.15
	Structures	2002	2.99E+5	1.21E+9	0.99	1.24	1.24	1.09	3.55
		2003	1.93E+5	8.54E+8	0.99	1.27	1.28	1.09	1.75
		2004	2.10E+5	2.00E+8	0.99	1.86	1.87	2.08	5.89
		2005	2.56E+5	5.07E+8	0.99	1.80	1.81	1.97	9.61
		2006	5.97E+5	4.93E+10	0.99	1.19	1.20	1.01	1.16
	Equipment	2002	1.45E+6	1.62E+10	0.99	1.27	1.28	1.07	3.02
		2003	1.39E+6	2.71E+10	0.97	1.24	1.27	1.09	1.58
		2004	1.49E+6	9.14E+9	0.99	1.40	1.41	1.62	4.61
		2005	1.51E+6	2.45E+10	0.99	1.22	1.23	1.15	2.12
		2006	1.34E+6	5.65E+9	0.99	1.42	1.43	1.60	6.20
541300	Total	2002	3.38E+6	2.32E+10	0.99	1.47	1.48	1.67	5.02
		2003	3.09E+6	2.61E+10	0.99	1.26	1.27	1.05	1.62
		2004	3.97E+6	1.12E+11	1.00	1.23	1.23	1.03	1.37
		2005	4.94E+6	2.54E+11	1.00	1.20	1.20	1.04	1.71
		2006	4.96E+6	2.82E+10	1.00	1.40	1.40	1.75	8.36
	Structures	2002	7.41E+5	6.32E+9	1.00	1.70	1.71	1.64	7.47
		2003	4.29E+5	3.32E+9	1.00	1.29	1.29	1.01	1.33
		2004	6.96E+5	4.38E+10	1.00	1.22	1.22	1.00	1.40
		2005	7.12E+5	9.00E+9	1.00	1.25	1.25	1.08	2.08
		2006	8.73E+5	3.44E+9	1.00	1.58	1.59	1.63	9.88
	Equipment	2002	2.96E+6	1.39E+10	0.99	1.37	1.38	1.54	3.95
		2003	2.66E+6	1.94E+10	0.99	1.25	1.26	1.05	1.59
		2004	3.27E+6	5.83E+10	1.00	1.22	1.23	1.04	1.29
		2005	4.23E+6	2.40E+11	1.00	1.19	1.20	1.03	1.59
		2006	4.09E+6	2.35E+10	1.00	1.27	1.28	1.49	5.47

Table 2
Population characteristics for the simulation study

Industry	Stratum	Population size	Sampling fraction	Correlation with Payroll		
				Total	Structures	Equipment
211000	10	26	1.00	0.65	0.53	0.95
	2A	128	0.77	0.68	0.66	0.22
	2B	372	0.11	0.57	0.51	0.51
	2C	1,800	0.02	-0.07	0.00	-0.10
	2D	10,406	0.00	0.28	0.00	0.28
212300	10	30	1.00	0.96	0.95	0.94
	2A	108	0.37	0.85	0.74	0.77
	2B	414	0.07	0.03	0.76	-0.03
	2C	1,310	0.03	0.42	0.13	0.43
	2D	4,762	0.01	0.44	-0.22	0.44
339900	10	158	1.00	0.80	0.40	0.80
	2A	498	0.26	0.40	0.04	0.51
	2B	2,048	0.05	0.20	0.24	0.18
	2C	6,310	0.02	0.19	0.48	0.09
	2D	25,288	0.00	0.37	0.67	0.36
541300	10	160	1.00	0.60	0.56	0.59
	2A	959	0.38	0.20	0.39	0.06
	2B	4,531	0.06	0.28	0.13	0.27
	2C	17,913	0.01	0.08	0.06	0.08
	2D	67,440	0.00	0.13	-0.01	0.15

Table 3
Simulation results (in %) for variance estimation under covariate-dependent response mechanism

Estimate	Item	Industry		v_{J1}	v_{L1}	v_L	v_{JL}	v_J	$\tilde{v}_J$
Ratio	Total	211000	RB	-35.8	-38.1	-10.3	-8.0	39.1	113.9
			ST	49.8	50.4	47.4	48.6	252.9	182.9
			ER	19.6	19.8	12.2	11.8	10.7	1.1
		212300	RB	-20.4	-22.2	-4.48	-2.69	54.8	266.4
			ST	30.3	31.1	26.8	27.3	139.1	268.8
			ER	12.6	12.6	9.9	9.6	6.3	0.1
		339900	RB	-21.2	-22.5	0.26	1.55	-5.34	52.5
			ST	47.3	47.0	55.0	56.0	43.9	67.8
			ER	14.3	14.6	10.4	10.3	10.0	2.6
		541300	RB	-20.7	-21.0	3.83	4.08	-11.6	18.4
			ST	32.7	32.8	34.9	35.0	29.4	32.0
			ER	12.6	12.7	8.6	8.6	10.7	6.2
	Structures	211000	RB	-38.0	-40.1	-11.9	-9.59	33.9	108.1
			ST	51.3	51.9	48.5	49.6	244.4	180.8
			ER	20.9	21.1	12.9	12.6	11.1	1.1
		212300	RB	-23.2	-23.9	-12.4	-11.6	33.2	341.5
			ST	31.5	32.0	27.1	27.0	95.0	344.3
			ER	12.3	12.3	10.4	10.3	6.9	0.1
		339900	RB	-20.0	-20.4	-6.31	-5.88	-10.9	39.8
			ST	42.5	42.7	42.3	42.3	39.9	64.0
			ER	15.9	16.0	12.7	12.6	13.2	5.4
		541300	RB	-20.0	-20.1	0.09	0.33	-15.9	15.7
			ST	42.6	42.5	50.5	50.7	41.1	42.7
			ER	13.1	13.2	9.9	9.9	12.0	6.5
	Equipment	211000	RB	-15.0	-17.3	14.1	16.4	-9.37	27.9
			ST	63.9	62.6	87.7	90.0	64.1	69.6
			ER	16.2	16.7	13.3	13.0	14.7	6.7
		212300	RB	-21.4	-23.3	-4.13	-2.21	39.7	201.1
			ST	31.7	32.5	28.4	29.0	113.7	204.4
			ER	13.3	13.5	10.2	10.1	7.7	0.2
		339900	RB	-21.4	-22.8	1.18	2.57	-7.29	50.8
			ST	51.2	50.9	60.8	61.9	47.9	69.2
			ER	15.5	15.8	11.6	11.4	11.0	2.3
		541300	RB	-19.7	-19.9	6.16	6.43	-11.9	12.8
			ST	33.8	33.9	38.4	38.5	31.0	30.9
			ER	12.5	12.5	8.9	8.9	11.0	7.0
Count	Total	211000	RB	-30.1	-31.9	0.05	1.85	1.05	103.1
			ST	50.4	50.5	55.9	57.3	46.7	113.4
			ER	15.3	15.6	9.0	8.8	8.7	1.0
		212300	RB	-33.2	-34.6	-6.30	-4.96	17.6	204.5
			ST	38.7	39.6	27.7	27.8	42.8	208.6
			ER	14.1	14.7	9.1	8.7	6.9	0.4
		339900	RB	-23.9	-24.6	1.73	2.44	-14.2	46.4
			ST	47.5	47.4	55.2	55.7	43.4	62.4
			ER	13.4	13.5	9.1	9.1	10.7	2.5
		541300	RB	-22.9	-23.2	1.68	1.94	-18.8	15.4
			ST	32.9	33.0	32.0	32.2	30.2	28.9
			ER	11.5	11.6	7.2	7.1	10.6	5.2
	Structures	211000	RB	-30.3	-32.2	-0.15	1.65	-1.27	101.5
			ST	51.3	51.3	57.3	58.7	46.7	112.3
			ER	15.8	16.3	9.6	9.4	9.2	0.8
		212300	RB	-37.4	-38.0	-13.5	-12.9	3.53	250.2
			ST	41.6	42.0	28.9	28.8	32.2	254.8
			ER	15.4	15.6	9.6	9.5	8.1	0.4
		339900	RB	-20.0	-20.3	-4.33	-4.00	-14.5	38.6
			ST	42.3	42.4	42.4	42.4	40.1	62.8
			ER	14.6	14.7	11.9	11.8	13.6	5.0
		541300	RB	-20.9	-21.2	-0.54	-0.32	-18.9	14.5
			ST	41.6	41.6	47.8	48.0	40.6	40.9
			ER	12.5	12.5	9.1	9.1	12.1	6.0
	Equipment	211000	RB	-17.8	-20.0	11.2	13.3	-13.0	26.6
			ST	58.9	58.0	76.4	78.4	57.7	64.1
			ER	15.7	15.8	12.5	12.3	14.5	6.1
		212300	RB	-30.7	-32.2	-4.74	-3.27	12.1	164.3
			ST	37.6	38.6	29.1	29.3	38.7	168.9
			ER	14.1	14.5	9.6	9.5	7.9	0.6
		339900	RB	-24.1	-24.9	2.52	3.27	-15.2	45.0
			ST	51.2	51.1	61.0	61.5	47.7	64.1
			ER	14.8	15.1	9.9	9.8	11.9	2.3
		541300	RB	-21.6	-21.9	4.10	4.39	-18.1	10.1
			ST	33.6	33.7	35.2	35.3	31.5	28.2
			ER	11.1	11.1	7.2	7.1	10.3	5.9

Table 4
Simulation results (in %) for variance estimation under within-stratum uniform response mechanism

Estimate	Item	Industry		v_{J1}	v_{L1}	v_L	v_{JL}	v_J	$\tilde{v}_J$
Ratio	Total	211000	RB	-49.2	-50.4	-17.2	-16.0	89.2	138.4
			ST	55.4	56.1	43.7	43.9	310.5	258.3
			ER	26.9	27.1	13.9	13.8	9.10	1.80
		212300	RB	-5.42	-7.99	16.1	18.7	111.7	337.2
			ST	28.9	28.5	37.4	39.6	179.2	341.7
			ER	13.5	13.8	9.85	9.50	5.85	0.05
		339900	RB	-9.37	-10.5	18.0	19.2	16.5	83.8
			ST	45.8	45.4	59.0	60.1	48.4	95.6
			ER	14.5	14.7	9.55	9.55	8.60	2.65
		541300	RB	-8.83	-9.03	18.2	18.4	6.62	44.5
			ST	26.7	26.8	36.6	36.7	28.2	52.3
			ER	12.6	12.6	8.45	8.45	9.70	5.35
	Structures	211000	RB	-52.6	-53.7	-19.2	-18.0	78.4	128.0
			ST	58.0	58.7	45.3	45.4	290.8	248.5
			ER	28.7	29.0	15.1	14.9	9.80	2.25
		212300	RB	-16.2	-18.1	9.92	11.9	63.5	356.2
			ST	32.0	32.4	37.1	38.5	108.6	361.9
			ER	15.5	16.0	10.9	10.7	6.65	0.35
		339900	RB	-13.2	-13.6	13.9	14.3	1.15	54.9
			ST	47.8	47.8	59.2	59.5	46.3	82.7
			ER	17.2	17.2	12.6	12.5	13.8	6.40
		541300	RB	-8.9	-9.2	19.2	19.5	-2.22	36.0
			ST	39.4	39.3	53.7	54.0	38.6	55.9
			ER	12.9	12.9	8.85	8.85	11.3	6.85
	Equipment	211000	RB	-1.1	-2.75	27.5	29.1	12.8	60.1
			ST	64.4	63.2	88.6	90.3	71.1	89.8
			ER	15.3	15.6	12.0	12.0	12.7	5.10
		212300	RB	-6.3	-8.96	16.8	19.4	90.0	263.1
			ST	30.3	29.8	39.3	41.6	148.6	269.1
			ER	13.9	14.2	10.1	9.60	6.75	0.15
		339900	RB	-8.84	-10.1	19.5	20.7	15.8	84.6
			ST	50.8	50.3	65.7	66.9	52.9	98.8
			ER	15.1	15.3	10.3	10.3	9.50	2.45
		541300	RB	-6.89	-7.1	19.9	20.1	6.76	38.5
			ST	28.6	28.6	40.0	40.1	30.1	48.3
			ER	12.4	12.4	8.60	8.55	10.3	5.90
Count	Total	211000	RB	-27.8	-29.0	14.2	15.4	16.3	149.5
			ST	47.4	47.5	53.1	54.2	44.4	158.1
			ER	16.0	16.2	8.30	8.30	7.45	1.85
		212300	RB	-33.5	-34.9	15.3	16.7	23.9	219.9
			ST	40.0	40.9	38.5	39.5	39.4	228.5
			ER	18.8	19.3	9.80	9.65	8.55	1.90
		339900	RB	-16.5	-17.1	20.2	20.8	4.21	75.7
			ST	45.1	44.0	57.9	58.5	42.4	87.6
			ER	15.6	15.8	9.40	9.35	10.8	3.20
		541300	RB	-9.61	-9.81	18.9	19.1	-0.77	45.0
			ST	26.5	26.5	36.0	36.1	24.7	52.2
			ER	12.4	12.4	8.55	8.55	11.3	4.85
	Structures	211000	RB	-27.5	-28.7	14.5	15.7	14.6	149.0
			ST	48.1	48.1	54.5	55.6	45.1	157.6
			ER	17.1	17.5	9.05	9.00	8.50	2.05
		212300	RB	-39.4	-40.4	11.6	12.6	10.1	238.5
			ST	44.8	45.5	38.8	39.6	32.1	248.4
			ER	20.2	20.7	9.95	9.85	10.3	1.80
		339900	RB	-14.2	-14.6	13.6	14.0	-3.55	53.5
			ST	47.1	47.0	57.3	57.6	45.1	80.5
			ER	17.6	17.7	12.1	12.1	14.7	6.30
		541300	RB	-9.54	-9.76	20.0	20.2	-5.32	36.0
			ST	39.1	39.0	53.3	53.5	38.3	55.8
			ER	12.6	12.6	9.05	9.05	11.9	6.55
	Equipment	211000	RB	-8.12	-9.64	22.7	24.2	1.56	54.3
			ST	58.0	57.1	76.2	77.7	57.5	82.1
			ER	16.5	16.7	12.4	12.4	14.6	6.45
		212300	RB	-28.5	-30.0	17.1	18.6	21.5	189.7
			ST	37.6	38.4	40.9	42.0	38.2	198.9
			ER	18.1	18.5	9.95	9.80	9.25	1.75
		339900	RB	-15.8	-16.4	21.8	22.5	4.69	76.8
			ST	49.5	49.3	64.6	65.2	47.4	91.1
			ER	16.4	16.5	9.45	9.40	11.4	3.20
		541300	RB	-7.53	-7.74	20.2	20.4	0.26	38.8
			ST	28.2	28.2	39.2	39.4	27.2	48.0
			ER	12.7	12.7	8.30	8.25	11.3	5.65

Acknowledgements

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical or methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors thank two referees and an associate editor for their helpful comments and suggestions, and Carol Caldwell, Rita Petroni, and Mark Sands for their useful comments on an earlier version of this paper. Jun Shao's research was partially supported by the NSF Grant SES-0705033.

References

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 429-440.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

Kott, P. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment of unit nonresponse. *Incomplete Data in Sample Surveys*. New York: Academic Press, 20, 143-184.

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

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

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

Thompson, J.R. (2000). *Simulation*: *A Modeler's Approach*. New York: John Wiley & Sons, Inc.

Thompson, K.J., and Yung, W. (2006). To replicate (a weight adjustment procedure) or not to replicate? An analysis of the variance estimation effects of a shortcut procedure using the stratified jackknife. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3772-3779.

Vartivarian, S., and Little, R.J. (2002). On the formation of weighting adjustment cells for unit non-response. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 3553-3558.

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

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 227-234
Statistics Canada, Catalogue No. 12-001-X

Rescaled bootstrap for stratified multistage sampling

John Preston [1]

Abstract

In large scaled sample surveys it is common practice to employ stratified multistage designs where units are selected using simple random sampling without replacement at each stage. Variance estimation for these types of designs can be quite cumbersome to implement, particularly for non-linear estimators. Various bootstrap methods for variance estimation have been proposed, but most of these are restricted to single-stage designs or two-stage cluster designs. An extension of the rescaled bootstrap method (Rao and Wu 1988) to stratified multistage designs is proposed which can easily be extended to any number of stages. The proposed method is suitable for a wide range of reweighting techniques, including the general class of calibration estimators. A Monte Carlo simulation study was conducted to examine the performance of the proposed multistage rescaled bootstrap variance estimator.

Key Words: Bootstrap; Calibration; Multistage sampling; Stratification; Variance estimation.

1. Introduction

Stratified multistage sampling designs are especially suited to large scaled sampled surveys because of the advantage of clustering collection effort. Various methods exist for variance estimation for these complex survey designs. The most commonly used methods are the linearization (or Taylor) method, and resampling methods, such as the jackknife, balance repeated replication and the bootstrap. The linearization method can be quite cumbersome to implement for complex survey designs as it requires the derivation of separate variance formulae for each non-linear estimator. Some approximations are normally required for the variance of non-linear functions, such as ratios and correlation and regression coefficients, and functionals, such as quantiles.

On the other hand, the various resampling methods employ a single variance formulae for all estimators. The replication methods can reflect the effects of a wide range of reweighting techniques, including calibration, and adjustments due to provider non-response and population undercoverage. The jackknife and balance repeated replication methods are only applicable to stratified multistage designs where the clusters are sampled with replacement or the first-stage sampling fractions are negligible. A number of different bootstrap methods for finite population sampling have been proposed in the literature, including the with-replacement bootstrap (McCarthy and Snowden 1985), the rescaled bootstrap (Rao and Wu 1988), the mirror match bootstrap (Sitter 1992a), and the without-replacement bootstrap (Gross 1980; Bickel and Freedman 1984; Sitter 1992b). A summary of these bootstrap methods can be found in Shao and Tu (1995).

Most of these bootstrap methods are restricted to single-stage designs or multistage designs where the first-stage sampling units are selected with replacement or the first-stage sampling fractions are small in most strata. However, in many large scaled sample surveys it is common practice to employ highly stratified multistage designs where units are selected using simple random sampling without replacement at each stage. Some typical examples of these types of surveys are employer-employee surveys, such as the Survey of Employee Earnings and Hours (ABS 2008), and school-student surveys, such as the National Survey on the Use of Tobacco by Australian Secondary School Students (White and Hayman 2006).

McCarthy and Snowden (1985) proposed an extension of their with-replacement bootstrap to two-stage sampling in the special case of equal cluster sizes and equal within cluster sample sizes, while Rao and Wu (1988) and Sitter (1992a) have given extensions of their rescaled bootstrap and mirror match bootstrap methods to two-stage cluster sampling. More recently, Funaoka, Saigo, Sitter and Toida (2006) proposed two Bernoulli-type bootstrap methods, the general Bernoulli bootstrap and the short cut Bernoulli bootstrap, which can easily handle multistage stratified designs where units are selected using simple random sampling without replacement at each stage. The general Bernoulli bootstrap has the advantage that it can handle any combination of sample sizes, but it requires a much larger number of random number generations than the short cut Bernoulli bootstrap.

In this paper, an extension of the rescaled bootstrap procedure to stratified multistage sampling where units are selected using simple random sampling without replacement at each stage is proposed. In Section 2, the notation for stratified multistage sampling is introduced. In Section 3, the extension of the rescaled bootstrap estimator to multistage sampling is described. The main findings of a simulation study are reported in Section 4. Some concluding remarks are provided in Section 5.

1. John Preston, Australian Bureau of Statistics, 639 Wickham Street, Fortitude Valley QLD 4006, Australia. E-mail: john.preston@abs.gov.au.

2. Stratified multistage sampling

For simplicity, the case of stratified three-stage sampling is presented. Consider a finite population U divided into H nonoverlapping strata $U = \{U_1, ..., U_H\}$, where U_h is comprised of N_{1h} primary sampling units (PSU's). At the first-stage, a simple random sample without replacement (SRSWOR) of n_{1h} PSU's are selected with selection probabilities $\pi_{1hi} = n_{1h} / N_{1h}$ within each stratum h. Suppose selected PSU i in stratum h is comprised of N_{2hi} secondary sampling units (SSU's). At the second-stage, a SRSWOR of size n_{2hi} SSU's are selected with selection probabilities $\pi_{2hij} = n_{2hi} / N_{2hi}$ within each selected PSU. Suppose selected SSU j in selected PSU i in stratum h is comprised of N_{3hij} ultimate sampling units (USU's). At the third-stage, a SRSWOR of size n_{3hij} USU's are selected with selection probabilities $\pi_{3hijk} = n_{3hij} / N_{3hij}$ within each selected SSU.

The objective is to estimate the population total $Y = \sum_{h=1}^{H}\sum_{i=1}^{N_{1h}}\sum_{j=1}^{N_{2hi}}\sum_{k=1}^{N_{3hij}} y_{hijk}$, where y_{hijk} is the value for the variable of interest y for USU k in SSU j in PSU i in stratum h. An unbiased estimate of Y is given by:

$$\hat{Y} = \sum_{h=1}^{H}\hat{Y}_h = \sum_{h=1}^{H}\frac{N_{1h}}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{N_{2hi}}{n_{2hi}}\sum_{j=1}^{n_{2hi}}\frac{N_{3hij}}{n_{3hij}}\sum_{k=1}^{n_{3hij}} y_{hijk}$$

where $\hat{Y}_h = (N_{1h} / n_{1h})\sum_{i=1}^{n_{1h}}\hat{Y}_{hi}$, $\hat{Y}_{hi} = (N_{2hi} / n_{2hi})\sum_{j=1}^{n_{2hi}}\hat{Y}_{hij}$ and $\hat{Y}_{hij} = (N_{3hij} / n_{3hij})\sum_{k=1}^{n_{3hij}} y_{hijk}$. This estimator can also be written as $\hat{Y} = \sum_{h=1}^{H}\sum_{i=1}^{n_{1h}}\sum_{j=1}^{n_{2hi}}\sum_{k=1}^{n_{3hij}} w_{hijk} y_{hijk}$, where $w_{hijk} = w_{1hi} w_{2hij} w_{3hijk} = (N_{1h} / n_{1h})(N_{2hi} / n_{2hi})(N_{3hij} / n_{3hij})$ is the sampling weight for USU k in SSU j in PSU i in stratum h.

An unbiased estimate of $\mathrm{Var}(\hat{Y})$ is given by (Särndal, Swensson and Wretman 1992):

$$\begin{aligned}\hat{\mathrm{Var}}(\hat{Y}) = &\sum_{h=1}^{H}\frac{N_{1h}^2}{n_{1h}}(1 - f_{1h})s_{1h}^2 \\ &+ \sum_{h=1}^{H}\frac{N_{1h}}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{N_{2hi}^2}{n_{2hi}}(1 - f_{2hi})s_{2hi}^2 \\ &+ \sum_{h=1}^{H}\frac{N_{1h}}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{N_{2hi}}{n_{2hi}}\sum_{j=1}^{n_{2hi}}\frac{N_{3hij}^2}{n_{3hij}}(1 - f_{3hij})s_{3hij}^2 \quad (2.1)\end{aligned}$$

where $f_{1h} = (n_{1h}/N_{1h})$, $f_{2hi} = (n_{2hi}/N_{2hi})$, $f_{3hij} = (n_{3hij}/N_{3hij})$, $\bar{\hat{Y}}_h = \sum_{i=1}^{n_{1h}}\hat{Y}_{hi}/n_{1h}$, $\bar{\hat{Y}}_{hi} = \sum_{j=1}^{n_{2hi}}\hat{Y}_{hij}/n_{2hi}$, $\bar{y}_{hij} = \sum_{k=1}^{n_{3hij}} y_{hijk}/n_{3hij}$, $s_{1h}^2 = \sum_{i=1}^{n_{1h}}(\hat{Y}_{hi} - \bar{\hat{Y}}_h)^2/(n_{1h} - 1)$, $s_{2hi}^2 = \sum_{j=1}^{n_{2hi}}(\hat{Y}_{hij} - \bar{\hat{Y}}_{hi})^2/(n_{2hi} - 1)$ and $s_{3hij}^2 = \sum_{k=1}^{n_{3hij}}(y_{hijk} - \bar{y}_{hij})^2/(n_{3hij} - 1)$.

3. Rescaled bootstrap for stratified multistage sampling

Rao and Wu (1988) proposed a rescaling of the standard bootstrap method for various sampling designs including stratified sampling. Since the rescaling factors are applied to the survey data values, this method is only applicable to smooth statistics. Rao, Wu and Yue (1992) presented a modification to this rescaled bootstrap method where the rescaling factors are applied to the survey weights, rather than the survey data values. This modified rescaled bootstrap method is equivalent to the original rescaled bootstrap method, but has the added advantage that it is applicable to non-smooth statistics as well as smooth statistics. Kovar, Rao and Wu (1988) showed that when using a bootstrap sample size of $n_h^* = n_h - 1$ the rescaled bootstrap estimator performed well for smooth statistics.

Although bootstrap samples are usually selected with replacement, Chipperfield and Preston (2007) modified the rescaled bootstrap method to the situation where the bootstrap samples are selected without replacement. Under this without replacement rescaled bootstrap method it can be shown that the choice of either $n_h^* = \lfloor n_h / 2 \rfloor$ or $n_h^* = \lceil n_h / 2 \rceil$ is optimal, where the operators $\lfloor x \rfloor$ and $\lceil x \rceil$ round the argument x down and up respectively to the nearest integer. The choice of $n_h^* = \lfloor n_h / 2 \rfloor$ has the desirable property that the bootstrap weights will never be negative.

For simplicity, the case of stratified three-stage sampling is presented, but the proposed procedure can easily be extended to any number of stages. The without replacement rescaled bootstrap procedure for stratified three-stage sampling is as follows:

(a) Draw a simple random sample of n_{1h}^* PSU's without replacement from the n_{1h} PSU's in the sample. Let δ_{1hi} be equal to 1 if PSU i in stratum h is selected and 0 otherwise. Calculate the PSU bootstrap weights:

$$w_{1hi}^* = w_{1hi}\left(1 - \lambda_{1h} + \lambda_{1h}\frac{n_{1h}}{n_{1h}^*}\delta_{1hi}\right)$$

where $\lambda_{1h} = \sqrt{n_{1h}^*(1 - f_{1h})/(n_{1h} - n_{1h}^*)}$.

(b) Within each of the PSU's in the sample, draw a simple random sample of n_{2hi}^* SSU's without replacement from the n_{2hi} SSU's in the sample. Let δ_{2hi} be equal to 1 if SSU j in PSU i in stratum h is selected and 0 otherwise. Calculate the conditional SSU bootstrap weights:

$$w^*_{2hij} =$$

$$w_{2hij}\frac{w_{1hi}}{w^*_{1hi}}\Bigg(1-\lambda_{1h}+\lambda_{1h}\frac{n_{1h}}{n^*_{1h}}\delta_{1hi} - \lambda_{2hi}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi}+\lambda_{2hi}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi}\frac{n_{2hi}}{n^*_{2hi}}\delta_{2hij}\Bigg)$$

where $\lambda_{2hi} = \sqrt{n^*_{2hi}f_{1h}(1-f_{2hi})/(n_{2hi}-n^*_{2hi})}$.

(c) Within each of the SSU's in the sample, draw a simple random sample of n^*_{3hij} USU's without replacement from the n_{3hij} USU's in the sample. Let δ_{3hijk} be equal to 1 if USU k in SSU j in PSU i in stratum h is selected and 0 otherwise. Calculate the conditional USU bootstrap weights:

$$w^*_{3hijk} =$$

$$w_{3hijk}\frac{w_{1hi}}{w^*_{1hi}}\frac{w_{2hij}}{w^*_{2hij}}\Bigg(1-\lambda_{1h}+\lambda_{1h}\frac{n_{1h}}{n^*_{1h}}\delta_{1hi} - \lambda_{2hi}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi}+\lambda_{2hi}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi}\frac{n_{2hi}}{n^*_{2hi}}\delta_{2hij} - \lambda_{3hij}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi}\sqrt{\frac{n_{2hi}}{n^*_{2hi}}}\delta_{2hij} + \lambda_{3hij}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi}\sqrt{\frac{n_{2hi}}{n^*_{2hi}}}\delta_{2hij}\frac{n_{3hij}}{n^*_{3hij}}\delta_{3hijk}\Bigg)$$

where $\lambda_{3hij} = \sqrt{n^*_{3hij}f_{1h}f_{2hi}(1-f_{3hij})/(n_{3hij}-n^*_{3hij})}$.

(d) Calculate the bootstrap estimates:

$$\hat{Y}^* = \sum_{h=1}^{H}\sum_{i=1}^{n_{1h}}\sum_{j=1}^{n_{2hi}}\sum_{k=1}^{n_{3hij}} w^*_{hijk}y_{hijk}, \quad \hat{\theta} = g(\hat{Y}^*)$$

where $w^*_{hijk} = w^*_{1hi}w^*_{2hij}w^*_{3hijk}$.

(e) Independently repeat steps (a) to (d) a large number of times, B, and calculate the bootstrap estimates, $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$, ..., $\hat{\theta}^{(B)}$.

(f) The bootstrap variance estimator of $\hat{\theta}$ is given by:

$$\mathrm{Var}(\hat{\theta}) = E_*(\hat{\theta}-E_*(\hat{\theta}))^2 \qquad (3.1)$$

or the Monte Carlo approximation:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)}-\bar{\hat{\theta}})^2$$

where $\bar{\hat{\theta}} = \sum_{b=1}^{B}\hat{\theta}^{(b)}/B$.

It is shown in the Appendix that the multistage rescaled bootstrap variance estimator for stratified three-stage sampling as defined by (3.1) reduces to the standard unbiased three-stage variance estimator (2.1) in the case of $\hat{\theta}$ being a linear estimator. The choice of $n^*_{1h} = \lfloor n_{1h}/2 \rfloor$, $n^*_{2hi} = \lfloor n_{2hi}/2 \rfloor$ and $n^*_{3hij} = \lfloor n_{3hij}/2 \rfloor$ will be optimal and will have the desirable property that the bootstrap weights will never be negative.

The proposed procedure can easily be extended to any number of stages by adding terms of the form $-\lambda_R(\prod_{r=1}^{R-1}\sqrt{(n_r/n^*_r)}\delta_r) + \lambda_R(\prod_{r=1}^{R-1}\sqrt{(n_r/n^*_r)}\delta_r)(n_R/n^*_R)\delta_R$ at each stage, R, to the bootstrap weight adjustments, where $\lambda_R = \sqrt{n^*_R(\prod_{r=1}^{R-1}f_r)(1-f_R)/(n_R-n^*_R)}$.

Yeo, Mantel and Liu (1999) presented an enhancement to the rescaled bootstrap which accounted for adjustments made to the design weights, such as post-stratification. For example, consider a simple case of non-integrated calibration using auxiliary information for two-stage stratified sampling (Estevao and Särndal 2006), which has the dual objectives of producing estimates for both a first-stage variable of interest $Y_1 = \sum_{(hi)\in U}y_{1hi}$ as well as a second-stage variable of interest, $Y_2 = \sum_{(hij)\in U}y_{2hij}$. Assume there exists:

(i) a set of p first-stage auxiliary variables $\boldsymbol{x}_{1hi}$ for which the population totals $\boldsymbol{X}_1 = \sum_{(hi)\in U}\boldsymbol{x}_{1hi}$ are known, and where the population totals are generated from a list frame of PSU's for which the $\boldsymbol{x}_{1hi}$ are known for every PSU in the population; and

(ii) a set of q second-stage auxiliary variables $\boldsymbol{x}_{2hij}$ for which the population totals $\boldsymbol{X}_2 = \sum_{(hij)\in U}\boldsymbol{x}_{2hij}$ are known, where the population totals are acquired from an external source.

The auxiliary variables can be used to form the first-stage and second-stage calibration estimators:

$$\hat{Y}_{\mathrm{CAL1}} = \sum_{(hi)\in s_1}\tilde{w}_{1hi}y_{1hi}$$

$$\hat{Y}_{\mathrm{CAL2}} = \sum_{(hij)\in s_2}\tilde{w}_{12hij}y_{2hij}$$

where the first-stage calibration weights, $\tilde{w}_{1hi}$, and the combined first-stage and second-stage calibration weights, $\tilde{w}_{12hij}$, are given by:

$$\tilde{w}_{1hi} = w_{1hi}\Bigg(1+\Bigg(\boldsymbol{X}_1-\sum_{(hi)\in s_1}w_{1hi}\boldsymbol{x}_{1hi}\Bigg)^T \Bigg(\sum_{(hi)\in s_1}w_{1hi}\boldsymbol{x}_{1hi}\boldsymbol{x}^T_{1hi}\Bigg)^{-1}\boldsymbol{x}_{1hi}\Bigg)$$

$$\tilde{w}_{12hij} = w_{1hi}w_{2hij}\Bigg(1+\Bigg(\boldsymbol{X}_2-\sum_{(hij)\in s_2}w_{1hi}w_{2hij}\boldsymbol{x}_{2hij}\Bigg)^T \Bigg(\sum_{(hij)\in s_2}w_{1hi}w_{2hij}\boldsymbol{x}_{2hij}\boldsymbol{x}^T_{2hij}\Bigg)^{-1}\boldsymbol{x}_{2hij}\Bigg).$$

Then the multistage rescaled bootstrap method can easily be modified in a similar manner to handle these calibration estimators by replacing step (d) in the procedure as follows:

(d) Calculate the first-stage and second-stage calibrated bootstrap weights in the same manner as the first-stage and second-stage calibrated weights:

$$\tilde{w}^*_{1hi} = w^*_{1hi}\left(1+\left(\boldsymbol{X}_1 - \sum_{(hi)\in s_1} w^*_{1hi}\boldsymbol{x}_{1hi}\right)^T \left(\sum_{(hi)\in s_1} w^*_{1hi}\boldsymbol{x}_{1hi}\boldsymbol{x}^T_{1hi}\right)^{-1}\boldsymbol{x}_{1hi}\right)$$

$$\tilde{w}^*_{12hij} = w^*_{1hi}w^*_{2hij}\left(1+\left(\boldsymbol{X}_2 - \sum_{(hij)\in s_2} w^*_{1hi}w^*_{2hij}\boldsymbol{x}_{2hij}\right)^T \left(\sum_{(hij)\in s_2} w^*_{1hi}w^*_{2hij}\boldsymbol{x}_{2hij}\boldsymbol{x}^T_{2hij}\right)^{-1}\boldsymbol{x}_{2hij}\right).$$

The first-stage and second-stage calibrated bootstrap estimates are calculated as:

$$\hat{Y}^*_{\mathrm{CAL1}} = \sum_{(hi)\in s_1} \tilde{w}^*_{1hi}y_{1hi}$$

$$\hat{Y}^*_{\mathrm{CAL2}} = \sum_{(hij)\in s_2} \tilde{w}^*_{12hij}y_{2hij}.$$

This procedure can easily be modified to any type of calibration and extended to any number of stages. This modification of the rescaled bootstrap takes into account adjustments made to the design weights due to calibration. Ideally all adjustments made to the design weights, including adjustments due to provider non- response and population under-coverage should also be made to the bootstrap weights.

4. Simulation study

A Monte Carlo simulation study was conducted to examine the performance of the multistage rescaled bootstrap variance estimator. The study was restricted to stratified two-stage sampling. The simulation study was based on ten artificial populations, each of which was stratified into $H = 5$ strata, with $N_{1h} = 50$ first-stage units within each stratum, and $N_{2hi} = 40$ second-stage units within each first-stage unit.

Firstly, the first-stage auxiliary variable x_{1hi} for each first-stage unit i in stratum h was generated from the normal distribution $N(\mu_{x1h}, (1-\rho_{x1b})\,\sigma^2_{x1b}/\rho_{x1b})$. Secondly, the second-stage auxiliary variable, x_{2hij}, and the second-stage target variables, y_{2hij} and z_{2hij}, for each second-stage unit j within first-stage unit i in stratum h were then generated from the multivariate normal distribution $N_3(\boldsymbol{\mu}_{2hi}, \boldsymbol{\Sigma}_{2hi})$ where $\boldsymbol{\mu}_{2hi}$ is the mean vector:

$$\boldsymbol{\mu}_{2hi} = \begin{bmatrix} \mu_{x2hi} \\ \mu_{y2hi} \\ \mu_{z2hi} \end{bmatrix}$$

with $\mu_{x2hi} = \mu_{y2hi} = \mu_{z2hi} = x_{1hi}$, and $\boldsymbol{\Sigma}_{2hi}$ is the variance-covariance matrix:

$$\boldsymbol{\Sigma}_{2hi} = \begin{bmatrix} \sigma^2_{x2hi} & \rho_{xy2hi}\sigma_{x2hi}\sigma_{y2hi} & \rho_{xz2hi}\sigma_{x2hi}\sigma_{z2hi} \\ \rho_{xy2hi}\sigma_{x2hi}\sigma_{y2hi} & \sigma^2_{y2hi} & \rho_{yz2hi}\sigma_{y2hi}\sigma_{z2hi} \\ \rho_{xz2hi}\sigma_{x2hi}\sigma_{z2hi} & \rho_{yz2hi}\sigma_{y2hi}\sigma_{z2hi} & \sigma^2_{z2hi} \end{bmatrix}$$

with $\sigma^2_{x2hi} = \sigma^2_{y2hi} = \sigma^2_{z2hi} = (1-\rho_{w2hi})\,\sigma^2_{w2hi}/\rho_{w2hi}$.

The parameter values that were kept stable across all ten populations were $\mu_{x1h} = 25\times(h+1)$, $\sigma^2_{b1h} = 10$, $\sigma^2_{w2hi} = 100$, $\rho_{xy2hi} = \rho_{xz2hi} = 0.75$ and $\rho_{yz2hi} = 0.50$. The parameter values that were varied across the ten populations were f_{1h}, the first-stage sampling fractions, f_{2hi}, the second-stage sampling fractions, ρ_{b1h} and ρ_{w2hi}. These parameter values are presented in Table 1.

Table 1
Characteristics of simulation populations

	f_{1h}	f_{2hi}	ρ_b	ρ_w
Pop I	0.1	0.1	0.75	0.75
Pop II	0.1	0.1	0.25	0.75
Pop III	0.1	0.5	0.75	0.75
Pop IV	0.1	0.5	0.25	0.75
Pop V	0.1	0.5	0.25	0.25
Pop VI	0.5	0.1	0.75	0.75
Pop VII	0.5	0.1	0.75	0.25
Pop VIII	0.5	0.1	0.25	0.25
Pop IX	0.3	0.3	0.75	0.25
Pop X	0.3	0.3	0.25	0.25

The parameters of interest used in the simulation study were the population mean, μ_y, the population ratio, $R_{yz} = \mu_y/\mu_z$, the population correlation coefficient, $\rho_{yz} = \sigma_{yz}/\sigma_y\sigma_z$, the population regression coefficient, $\beta_{yz} = \sigma_{yz}/\sigma^2_y$, and the population median, M_y.

In order to estimate these parameters of interest using the multistage bootstrap variance estimators, a total of $S = 20{,}000$ independent two-stage simple random samples were selected without replacement from each of the ten artificial populations. In addition, a grand total of $T = 100{,}000$ independent two-stage simple random samples were selected without replacement from each of the ten artificial

populations in order to estimate the true population variances for the parameters of interest. The multistage bootstrap variance estimators were calculated using $B = 100$ bootstrap samples.

The accuracy of the multistage bootstrap variance estimators were compared using the relative biases (RB) and the relative root mean square error (RRMSE). These measures were calculated as:

$$\text{RB} = \frac{1}{\hat{\text{V}}\text{ar}(\hat{Y})}\left[\frac{1}{S}\sum_{s=1}^{S}(\text{Var}_*(\hat{Y}_s) - \hat{\text{V}}\text{ar}(\hat{Y}))\right]$$

$$\text{RRMSE} = \frac{1}{\hat{\text{V}}\text{ar}(\hat{Y})}\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\text{Var}_*(\hat{Y}_s) - \hat{\text{V}}\text{ar}(\hat{Y}))^2}$$

where $\hat{\text{V}}\text{ar}(\hat{Y}) = T^{-1}\sum_{t=1}^{T}(\hat{Y}_t - Y)^2$ is the estimated true population variance, and $\text{Var}_*(\hat{Y}_s)$ are the multistage bootstrap variance estimators for the s^{th} simulation sample.

The multistage rescaled bootstrap variance estimator (MRBE) was compared against the single-stage rescaled bootstrap variance estimator (SRBE) and the multistage general Bernoulli bootstrap variance estimator (BBE) proposed by Funaoka *et al.* (2006), with bootstrap samples using the non-calibration estimation weights, $w_{hij} = w_{1hi}w_{2hij}$. The relative biases and relative root mean square errors of MRBE, SRBE and BBE using the non-calibration estimation weights for the ten artificial populations are given in Tables 2 and 3.

In the case of linear functions, such as means, and non-linear functions, such as ratios, correlation coefficients and regression coefficients, the MRBE performed better than the SRBE and BBE with respect to relative bias and relative root mean square error. While the MRBE performed consistently well across all ten artificial populations, the SRBE only performed well for artificial populations III, IV and V, where the first-stage sampling fractions were small ($f_{1h} = 0.1$) and the second-stage sampling fractions were large ($f_{2hi} = 0.5$), and the BBE only performed well for artificial populations VI, VII and VIII, where the first-stage sampling fractions were large ($f_{1h} = 0.5$) and the second-stage sampling fractions were small ($f_{2hi} = 0.1$). These sampling fractions were similar to the first-stage and second-stage sampling fractions used in the simulation study presented in Funaoka *et al.* (2006). The different levels of correlation between the first-stage units, and between the second-stage units within the first-stage units, controlled by varying the parameters ρ_b and ρ_w, had little impact on the performance of the variance estimators.

Table 2
Relative bias (%) of variance estimators

	Mean (μ_y)			Mean (μ_z)			Ratio (R_{yz})		
	MRBE	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**
Pop I	-0.28	-6.73	27.10	0.42	-6.63	27.32	0.00	-9.07	36.22
Pop II	-0.05	-2.21	11.83	0.59	-1.64	12.54	-0.43	-9.26	36.40
Pop III	-0.79	-2.63	3.62	-0.93	-2.66	3.40	-0.17	-5.30	5.19
Pop IV	-0.23	-0.52	3.60	-0.18	-0.46	3.61	0.53	-4.65	5.98
Pop V	0.15	-1.60	4.55	0.15	-1.64	4.54	0.52	-4.85	6.41
Pop VI	0.70	-39.18	-0.34	0.65	-39.36	-0.28	1.57	-46.40	1.30
Pop VII	0.19	-46.19	-0.26	-0.06	-46.48	-0.57	-0.27	-48.19	-0.73
Pop VIII	0.37	-38.62	-0.41	0.23	-39.36	-0.46	-0.26	-47.93	-0.62
Pop IX	0.42	-20.85	-7.76	-0.51	-20.03	-8.41	0.13	-23.13	-8.87
Pop X	-0.56	-12.35	-6.08	0.70	-10.87	-6.93	-0.72	-23.70	-9.51
	Correlation Coefficient (ρ_{yz})			**Regression Coefficient (β_{yz})**			**Median (M_y)**		
	MRBE	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**
Pop I	-2.31	-10.23	32.17	-0.08	-9.05	36.41	19.04	-19.86	33.21
Pop II	-1.51	-8.41	29.65	0.05	-8.74	36.41	19.29	2.42	40.85
Pop III	0.36	-4.37	5.69	0.05	-5.12	5.42	7.50	4.28	9.72
Pop IV	2.18	-0.60	7.17	0.28	-5.05	5.70	17.40	16.17	34.37
Pop V	0.79	-2.71	5.95	0.26	-5.40	6.34	8.29	4.78	11.49
Pop VI	0.32	-46.67	0.14	0.89	-46.59	0.69	13.57	-33.56	9.15
Pop VII	-0.07	-46.78	-0.39	-0.21	-47.85	-0.60	14.68	-38.16	11.86
Pop VIII	0.31	-44.25	-0.27	-0.09	-47.54	-0.55	2.09	-38.90	-0.64
Pop IX	-0.93	-23.02	-9.30	-0.20	-23.48	-9.20	8.08	-17.23	-1.97
Pop X	-0.82	-19.35	-8.24	-1.02	-23.89	-9.75	2.10	-13.84	-5.46

Note: The largest simulation error on the relative biases was less than 0.7%.

In the case of non-smooth statistics, such as medians, both the MRBE and the BBE tended to overestimate the true population variances, while the SRBE tended to underestimate the true population variances. Furthermore, the relative root mean square errors for medians were up to 3 times larger than the relative root mean square errors for means. The MRBE performed better than the BBE for the artificial populations I to V where the first-stage sampling fractions were smaller ($f_{1h} = 0.1$), while the BBE performed slightly better than the MRBE for the artificial populations VI to X where the first-stage sampling fractions were larger ($f_{1h} = 0.3$ or 0.5).

This overestimation of the multistage rescaled bootstrap for medians was similar to the findings shown in the studies by Kovar *et al.* (1988) and Rao *et al.* (1992) for the single-stage rescaled bootstrap. It should be noted that the original rescaled bootstrap introduced by Rao and Wu (1988) was developed only for smooth statistics, such as means, ratios, and correlation and regression coefficients.

The MRBE was examined using the calibration estimation weights, $\tilde{w}_{hij} = w_{1hi}\tilde{w}_{2hij}$, which satisfy the calibration constraint $\sum_{(hij)\in s_2} w_{1hi}\tilde{w}_{2hij}x_{2hij} = X_2$, where $X_2 = \sum_{(hij)\in U} x_{2hij}$ is the population total for the second-stage auxiliary variable. The relative biases and relative root mean square errors of the MRBE using the calibration estimation weights for the four artificial populations II, IV, VII and IX are given in Table 4.

Table 3
Relative root mean square error (%) of variance estimators

	Mean (μ_y)			Mean (μ_z)			Ratio (R_{yz})		
	MRBE	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**
Pop I	31.9	32.1	44.6	31.7	31.8	44.4	31.8	32.3	51.4
Pop II	33.9	33.8	38.2	33.4	33.3	38.1	32.2	32.9	51.7
Pop III	33.8	33.8	35.9	33.0	33.0	35.0	33.0	33.1	35.1
Pop IV	35.3	35.3	37.4	35.2	35.2	37.3	32.8	32.8	35.0
Pop V	32.0	31.9	34.2	34.3	34.2	36.5	33.0	33.1	35.6
Pop VI	16.4	40.7	16.5	16.4	40.9	16.5	16.5	47.5	16.5
Pop VII	16.1	47.4	16.4	16.1	47.8	16.4	16.1	49.0	16.1
Pop VIII	16.3	40.3	16.5	16.7	40.9	16.3	16.2	48.8	16.1
Pop IX	19.2	26.7	20.0	19.3	26.3	20.0	19.2	28.6	20.2
Pop X	19.8	22.4	20.2	19.9	21.6	20.3	19.1	29.0	20.6
	Correlation Coefficient (ρ_{yz})			**Regression Coefficient (β_{yz})**			**Median (M_y)**		
	MRBE	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**	**MRBE**	**SRBE**	**BBE**
Pop I	47.8	46.3	68.7	36.6	37.2	55.3	88.7	80.1	89.8
Pop II	48.4	47.1	66.6	37.4	37.9	55.6	93.4	91.0	115.9
Pop III	35.9	35.6	38.4	37.5	37.6	39.9	80.4	80.3	81.1
Pop IV	42.6	42.2	45.4	38.0	38.0	40.3	97.5	96.6	127.3
Pop V	40.3	40.0	43.3	37.3	37.5	40.1	31.5	30.7	63.3
Pop VI	21.6	48.4	21.7	16.9	47.8	17.0	55.3	51.4	52.0
Pop VII	21.4	48.4	21.3	16.9	49.0	16.8	53.5	51.4	51.4
Pop VIII	21.6	46.3	21.5	17.0	48.6	16.9	41.8	49.7	40.3
Pop IX	21.5	29.4	22.5	20.5	29.9	21.6	46.1	42.7	42.7
Pop X	22.7	27.8	23.4	20.6	30.2	21.9	39.7	38.9	37.9

Table 4
Relative bias (%) and relative root mean square error (%) of rescaled bootstrap variance estimator

	μ_y	R_{yz}	ρ_{yz}	β_{yz}	M_y
			Relative Bias (%)		
Pop II	-0.42	-0.29	-1.51	-0.08	20.98
Pop IV	0.40	0.49	1.83	0.08	18.28
Pop VII	-0.22	-0.24	-0.03	-0.28	12.24
Pop IX	0.62	0.19	-1.00	-0.20	7.24
			Relative Root Mean Square Error (%)		
Pop II	32.6	32.4	48.4	37.3	97.8
Pop IV	32.8	32.8	44.6	37.9	99.4
Pop VII	16.2	16.1	21.5	16.9	50.0
Pop IX	19.1	19.2	21.6	20.5	43.8

Note: The largest simulation error on the relative biases was less than 0.6%.

The relative biases and relative root mean square errors of the MRBE using the calibration estimation weights were similar to those using the non-calibration estimation weights.

5. Conclusion

This paper extends the rescaled bootstrap procedure to multistage sampling where units are selected using simple random sampling without replacement at each stage. Under the proposed multistage rescaled bootstrap method, the bootstrap samples are selected without replacement and rescaling factors are applied to the survey weights. This proposed method is relatively simple to implement and requires considerably less random number generations than the multistage general Bernoulli bootstrap method. The proposed method is also suitable for a wide range of reweighting techniques, including calibration, and adjustments due to provider non-response and population undercoverage. Furthermore, the results of the Monte Carlo simulation study indicate that the multistage rescaled bootstrap performs much better than the single-stage rescaled bootstrap and the multistage Bernoulli bootstrap for smooth statistics, such as means, ratios, and correlation and regression coefficients.

Appendix

In this Appendix it is shown that the multistage rescaled bootstrap variance estimator for stratified three-stage sampling reduces to the standard unbiased three-stage variance estimator (2.1) in the case of $\hat{\theta}$ being the linear estimator, $\hat{Y}^* = \sum_{h=1}^{H}\sum_{i=1}^{n_{1h}}\sum_{j=1}^{n_{2hi}}\sum_{k=1}^{n_{3hij}} w^*_{hijk} y_{hijk}$. The bootstrap variance estimator of $\hat{Y}^*$ is given by:

$$\mathrm{Var}(\hat{Y}^*) = \mathrm{Var}_{1*}(E_{2*}(E_{3*}(\hat{Y}^*))) + E_{1*}(\mathrm{Var}_{2*}(E_{3*}(\hat{Y}^*))) + E_{1*}(E_{2*}(\mathrm{Var}_{3*}(\hat{Y}^*))).$$

Using standard results on the expectation and variance with respect to the SRSWOR bootstrap sampling and some tedious but straightforward algebra, the components of bootstrap variance estimator are given below. The conditional expectation of $\hat{Y}^*$ given s_3 is

$$E_{3*}(\hat{Y}^*) = \sum_{h=1}^{H}\sum_{i=1}^{n_{1h}}\sum_{j=1}^{n_{2hi}} w_{1i}w_{2ij}\left(1 - \lambda_{1h} + \lambda_{1h}\frac{n_{1h}}{n^*_{1h}}\delta_{1hi} - \lambda_{2hi}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi} + \lambda_{2hi}\sqrt{\frac{n_{1h}}{n^*_{1h}}}\delta_{1hi}\frac{n_{2hi}}{n^*_{2hi}}\delta_{2hij}\right)\hat{Y}_{ij}$$

and the conditional variance of $\hat{Y}^*$ given s_3 is

$$\mathrm{Var}_{3*}(\hat{Y}^*) = \sum_{h=1}^{H}\frac{N_{1h}}{n^*_{1h}}\sum_{i=1}^{n_{1h}}\frac{N_{2hi}}{n^*_{2hi}}\sum_{j=1}^{n_{2hi}}\delta_{1hi}\,\delta_{2hij}\frac{N^2_{3hij}}{n_{3hij}}(1 - f_{3hij})s^2_{3hij}.$$

The conditional expectation of $E_{3*}(\hat{Y}^*)$ and $\mathrm{Var}_{3*}(\hat{Y}^*)$ given s_2 are

$$E_{2*}(E_{3*}(\hat{Y}^*)) = \sum_{h=1}^{H}\sum_{i=1}^{n_{1h}} w_{1i}(1 - \lambda_{1h} + \lambda_{1h}\frac{n_{1h}}{n^*_{1h}}\delta_{1hi})\,\hat{Y}_{hi}$$

$$E_{2*}(\mathrm{Var}_{3*}(\hat{Y}^*)) = \sum_{h=1}^{H}\frac{N_{1h}}{n^*_{1h}}\sum_{i=1}^{n_{1h}}\frac{N_{2hi}}{m_{2hi}}\sum_{j=1}^{n_{2hi}}\delta_{1hi}\frac{N^2_{3hij}}{n_{3hij}}(1 - f_{3hij})\,s^2_{3hij}$$

and the conditional variance of $E_{3*}(\hat{Y}^*)$ given s_2 is

$$\mathrm{Var}_{2*}(E_{3*}(\hat{Y}^*)) = \sum_{h=1}^{H}\frac{N_{1h}}{n^*_{1h}}\sum_{i=1}^{n_{1h}}\delta_{1hi}\frac{N^2_{2hi}}{n_{2hi}}(1 - f_{2hi})s^2_{2hi}.$$

Finally, the conditional expectation of $E_{2*}(\mathrm{Var}_{3*}(\hat{Y}^*))$ and $\mathrm{Var}_{2*}(E_{3*}(\hat{Y}^*))$ given s_1 are

$$E_{1*}(E_{2*}(\mathrm{Var}_{3*}(\hat{Y}^*))) = \sum_{h=1}^{H}\frac{N_{1h}}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{N_{2hi}}{n_{2hi}}\sum_{j=1}^{n_{2hi}}\frac{N^2_{3hij}}{n_{3hij}}(1 - f_{3hij})\,s^2_{3hij}$$

$$E_{1*}(\mathrm{Var}_{2*}(E_{3*}(\hat{Y}^*))) = \sum_{h=1}^{H}\frac{N_{1h}}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{N^2_{2hi}}{n_{2hi}}(1 - f_{2hi})\,s^2_{2hi}$$

which are equal to the third and second terms of (2.1) respectively, and the conditional variance of $E_{2*}(E_{3*}(\hat{Y}^*))$ given s_1 is

$$\mathrm{Var}_{1*}(E_{2*}(E_{3*}(\hat{Y}^*))) = \sum_{h=1}^{H}\frac{N^2_{1h}}{n_{1h}}(1 - f_{1h})\,s^2_{1h}$$

which is equal to the first term of (2.1).

References

Australian Bureau of Statistics (ABS) (2008). Employee Earnings and Hours, Catalogue Number 6306.0.

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics*, 12, 470-482.

Chipperfield, J., and Preston, J. (2007). Efficient bootstrap for business surveys. *Survey Methodology*, 33, 167-172.

Estevao, V., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74, 127-147.

Funaoka, F., Saigo, H., Sitter, R.R. and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, 32, 151-156.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, 181-184.

Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, 25-45.

McCarthy, P.J., and Snowden, C.B. (1985). The bootstrap and finite population sampling. Vital and Health Statistics (Series 2 No 95), Public Health Service Publication 85-1369, Washington, DC: U.S. Government Printing Office.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 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, 209-217.

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

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

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

White, V., and Hayman J. (2006). Smoking behaviours of Australian secondary students in 2005**.** National Drug Strategy Monograph Series No. 59. Canberra: Australian Government Department of Health and Ageing.

Yeo, D., Mantel, H. and Liu T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778-783.

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 235-245
Statistics Canada, Catalogue No. 12-001-X

Use of within-primary-sample-unit variances to assess the stability of a standard design-based variance estimator

Donsig Jang and John L. Eltinge [1]

Abstract

In analysis of sample survey data, degrees-of-freedom quantities are often used to assess the stability of design-based variance estimators. For example, these degrees-of-freedom values are used in construction of confidence intervals based on t distribution approximations; and of related t tests. In addition, a small degrees-of-freedom term provides a qualitative indication of the possible limitations of a given variance estimator in a specific application. Degrees-of-freedom calculations sometimes are based on forms of the Satterthwaite approximation. These Satterthwaite-based calculations depend primarily on the relative magnitudes of stratum-level variances. However, for designs involving a small number of primary units selected per stratum, standard stratum-level variance estimators provide limited information on the true stratum variances. For such cases, customary Satterthwaite-based calculations can be problematic, especially in analyses for subpopulations that are concentrated in a relatively small number of strata. To address this problem, this paper uses estimated within-primary-sample-unit (within PSU) variances to provide auxiliary information regarding the relative magnitudes of the overall stratum-level variances. Analytic results indicate that the resulting degrees-of-freedom estimator will be better than modified Satterthwaite-type estimators provided: (a) the overall stratum-level variances are approximately proportional to the corresponding within-stratum variances; and (b) the variances of the within-PSU variance estimators are relatively small. In addition, this paper develops errors-in-variables methods that can be used to check conditions (a) and (b) empirically. For these model checks, we develop simulation-based reference distributions, which differ substantially from reference distributions based on customary large-sample normal approximations. The proposed methods are applied to four variables from the U.S. Third National Health and Nutrition Examination Survey (NHANES III).

Key Words: Complex sample design; Degrees of freedom; Errors-in-variables regression; Satterthwaite approximation; Stratified multistage sample survey; Two-PSU-per-stratum design; U.S. Third National Health and Nutritional Examination Survey (NHANES III).

1. Introduction

1.1 Motivating example: Inference for special subpopulations in NHANES III

This work arose from a study of inference for geographically concentrated subpopulations in the U.S. Third National Health and Nutrition Examination Survey (NHANES III). For some general background on NHANES III, see National Center for Health Statistics (1996). In many analyses, NHANES III data are treated as arising from a stratified multistage sample design that uses 49 strata and two primary sample units (PSUs) per stratum. Consequently, formal inferences from NHANES III data (*e.g.*, construction of confidence intervals) often use the assumption that the associated variance estimators are based on approximately 49 degrees of freedom and are thus relatively stable.

However, the Mexican-American subpopulation is concentrated in a relatively small number of strata, so associated variance estimators may be less stable (*i.e.*, have greater sampling variability) than would be indicated by the nominal 49 degrees of freedom term. Consequently, it is important to use an appropriate estimator of the true degrees of freedom associated with variance estimators for such subpopulations, and to modify confidence interval calculations accordingly. Development of an appropriate degrees-of-freedom estimator can be complicated by moderate or severe heterogeneity in the underlying stratum-level variances. Such complications arose in the analysis of the four NHANES III variables listed in Table 1.1. Section 5 will consider inference for the means of these four variables for the subpopulation of Mexican-Americans aged 20-29.

Table 1.1
Four NHANES III variables

Variable Name	Description
BMPWT	Weight (*kg*)
HAR3	Do you smoke cigarettes now? (0/1)
TCRESULT	Serum total cholesterol (*mg/dL*)
HDRESULT	HDL cholesterol (*mg/dL*)

1.2 Stability of design-based variance estimators

Suppose we have a population partitioned into L strata, with N_h PSUs in stratum h for $h = 1, 2, ..., L$. Under a

1. Donsig Jang, Mathematica Policy Research, 600 Maryland Avenue SW, Suite 550, Washington, DC 20024-2512, U.S.A. E-mail: DJang@Mathematica-mpr.com; John L Eltinge, U.S. Bureau of Labor Statistics, PSB 1950, 2 Massachusetts Avenue NE, Washington, DC 20212-0001, U.S.A. E-mail: Eltinge_J@bls.gov.

stratified multistage sampling design, we select n_h PSUs, with replacement, and with per-draw selection probability p_{hi} for PSU i within stratum h where $\sum_{i=1}^{N_h} p_{hi} = 1$. Thus, a total of $n = \sum_{h=1}^{L} n_h$ PSUs are selected. Within selected PSU (h, i), n_{hi} secondary sample units (SSUs) are selected with replacement and with per-draw selection probabilities p_{hij}, where $\sum_{j=1}^{N_{hi}} p_{hij} = 1$ and N_{hi} is the number of SSUs in PSU (h, i). For a given survey item, let Y_h be the population total for stratum h, and define the overall population total $Y = \sum_{h=1}^{L} Y_h$. The total Y may correspond to a total either for the full population or for a specified subpopulation.

Our goal is to construct a confidence interval for the total Y. Let $\hat{Y}_{hij}$ be an unbiased estimator of Y_{hij}, the population total for secondary unit j in primary unit i in stratum h. Then a customary design-based estimator of Y is $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$, where $\hat{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} \hat{Y}_{hi}$; $p_{hi}^{-1} \hat{Y}_{hi}$ is a design unbiased estimator of Y_h based on data obtained from PSU i in stratum h; and $\hat{Y}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} p_{hij}^{-1} \hat{Y}_{hij}$ is an unbiased estimator of Y_{hi}, the population total for PSU i in stratum h.

Under the standard condition that sampling is independent across strata, the variance of $\hat{Y}$ can be written, $V(\hat{Y}) = \sum_{h=1}^{L} V_h$ where $V_h = \mathrm{Var}(\hat{Y}_h)$. Throughout the remainder of this paper, we will call the V_h terms the stratum-level variances, and we will assume that $n_h \geq 2$ for all $h = 1, 2, ..., L$. Note that V_h depends on the sample design used within stratum h, and is distinct from the within-stratum variance of element-level Y values. A simple unbiased estimator for $V(\hat{Y})$ is $\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}_h$ where $\hat{V}_h = n_h^{-1}(n_h - 1)^{-1} \sum_{h=1}^{n_h} (p_{hi}^{-1} \hat{Y}_{hi} - \hat{Y}_h)^2$; see, *e.g.*, Wolter (1985, page 44). Note that the estimator $\hat{V}_h$ is a multiple of a sum of squared differences among the terms $p_{hi}^{-1} \hat{Y}_{hi}$. In addition, under regularity conditions the random variables $p_{hi}^{-1} \hat{Y}_{hi}$ will be approximately normally distributed for a given stratum h. Consequently, the overall stratum-level variance estimators $\hat{V}_h$ generally will approximately satisfy the following condition.

(C.1) For $h = 1, 2, ..., L$, the terms $V_h^{-1}(n_h - 1)\hat{V}_h$ are distributed as independent chi-square random variables with $n_h - 1$ degrees of freedom, respectively, where $n_h \geq 2$.

Under condition (C.1), $\{V(\hat{Y})\}^{-1} d\hat{V}(\hat{Y})$ has the same first and second moments as a chi-square random variable with d degrees of freedom, where d is the solution to the equation,

$$2\{V(\hat{Y})\}^2 - V\{\hat{V}(\hat{Y})\}\, d = 0 \qquad (1.1)$$

or equivalently

$$d \overset{\text{def}}{=} \left\{ \sum_{h=1}^{L} (n_h - 1)^{-1} V_h^2 \right\}^{-1} \{V(\hat{Y})\}^2 \qquad (1.2)$$

where $V\{\hat{V}(\hat{Y})\} = \sum_{h=1}^{L} 2(n_h - 1)^{-1} V_h^2$. Direct substitution of $\hat{V}_h$ for V_h and $\hat{V}(\hat{Y})$ for $V(\hat{Y})$ in expression (1.2) leads to the Satterthwaite (1946)-type degrees-of-freedom estimator,

$$\hat{d}_S = \left\{ \sum_{h=1}^{L} (n_h - 1)^{-1} \hat{V}_h^2 \right\}^{-1} \{\hat{V}(\hat{Y})\}^2. \qquad (1.3)$$

For some general background on $\hat{d}_S$ and related estimators, see, *e.g.*, Smith (1936), Satterthwaite (1941, 1946), Cochran (1977, page 96) and Kendall, Stuart and Ord (1983, pages 91-92). In constructing confidence intervals for a subpopulation parameter, Casady, Dorfman and Wang (1998) use Bayesian ideas to develop related degrees-of-freedom measures for a Student's t-statistic.

For designs in which n_h is large for all h, the error in estimation of V_h is relatively small, and $\hat{d}_S$ can provide a satisfactory estimator of expression (1.2). However, many large-scale surveys use small n_h, *e.g.*, $n_h = 2$. For small-n_h cases, condition (C.1) and routine algebra lead to the expectation result $E(\hat{V}_h^2) = (n_h - 1)^{-1}(n_h + 1)V_h^2$. This implies that the standard Satterthwaite degrees-of-freedom estimator $\hat{d}_S$ can severely underestimate d, and that the corresponding confidence interval $\hat{Y} \pm t_{\hat{d}_S, 1-\alpha/2} \{\hat{V}(\hat{Y})\}^{1/2}$ may have a true coverage rate substantially below the nominal rate $1 - \alpha$. Consequently, Jang (1996) considered an alternative degrees-of-freedom estimator,

$$\hat{d}_{mS} = (3L + 14)^{-1}(9L)\, \hat{d}_S. \qquad (1.4)$$

for the two-PSUs-per-stratum design.

1.3 Use of auxiliary stratum-level data

For cases in which there is moderate heterogeneity among the V_h terms, simulation work by Jang (1996) indicated that $\hat{d}_{mS}$ performs relatively well. However, if there is substantial heterogeneity among the stratum variances (*i.e.*, if $L^{-1}d$ is relatively small), then $\hat{d}_{mS}$ may be unsatisfactory. The fundamental problem is that when the n_h values are relatively small, the estimators $\hat{V}_h$, by themselves, do not provide sufficient information regarding the relative magnitudes of the true stratum-level variances V_h. In some cases, a variance estimator based on auxiliary data is expected to be more stable than the customary design-based estimator; see *e.g.*, Isaki (1983). Similarly, auxiliary sources of information can be used to evaluate the relative magnitudes of the variances V_h.

The remainder of this paper will focus on auxiliary information provided by relationships between the overall stratum-level variances V_h and associated within-PSU variances. Recall from Wolter (1985, page 41) the decomposition,

$$\mathrm{Var}(\hat{Y}_h) = V_{Bh} + V_{Wh}, \qquad (1.5)$$

where $V_{Bh} = \mathrm{Var}\{\sum_{i=1}^{n_h}(n_h p_{hi})^{-1} Y_{hi}\}$ is the between-PSU variance, $V_{Wh} = \sum_{i=1}^{N_h}(n_h p_{hi})^{-1}\sigma_{2hi}^2$ is the within-PSU variance, $Y_{hi} = E(\hat{Y}_{hi} \mid \text{PSU } i, \text{ stratum } h)$ and $\sigma_{2hi}^2 = \mathrm{Var}(\hat{Y}_{hi} \mid \text{PSU } i, \text{ stratum } h)$. In addition, define $\overline{V}_W = L^{-1}\sum_{h=1}^{L} V_{Wh}$.

Estimators of V_{Wh} can provide useful auxiliary information on the relative magnitudes of V_h for two reasons. First, for designs with a small n_h and relatively large n_{hi}, the within-PSU variance estimators $\hat{V}_{Wh}$ may be considerably more stable than $\hat{V}_h$. Second, in some applications (*e.g.*, some of the examples presented in Section 5 below), observed variance estimates are consistent with a model under which V_h is proportional to V_{Wh}, *i.e.*,

$$V_h = \beta_1 V_{Wh} \text{ for all } h = 1, \ldots, L, \tag{1.6}$$

where β_1 is a fixed constant. The proportionality relationship (1.6) would arise if both V_{Bh} and V_{Wh} are proportional to a common scale factor, *e.g.*, $(\overline{Y}_h)^\alpha$ for some power α. Under relationship (1.6), expression (1.2) may be rewritten,

$$d = \left\{\sum_{h=1}^{L}(n_h - 1)^{-1} V_{Wh}^2\right\}^{-1}\left\{\sum_{h=1}^{L} V_{Wh}\right\}^2. \tag{1.7}$$

Consequently, given a set of stable within-PSU variance estimators $\hat{V}_{Wh}$ and associated variance-of-variance-estimators $\widehat{\mathrm{Var}}(\hat{V}_{Wh})$,

$$\hat{d}_{WS} = \left\{\sum_{h=1}^{L}(n_h - 1)^{-1}[\hat{V}_{Wh}^2 - \widehat{\mathrm{Var}}(\hat{V}_{Wh})]\right\}^{-1}\left(\sum_{h=1}^{L}\hat{V}_{Wh}\right)^2 \tag{1.8}$$

is an alternative estimator of d.

Section 2 considers some of the properties of $\hat{d}_{WS}$. Section 3.1 uses errors-in-variables tests to check the adequacy of the proportionality condition (1.6). Section 3.2 presents two related diagnostics for the relationship between V_h and auxiliary variables, and for the magnitude of the error in the observed auxiliary variables $\hat{V}_{Wh}$.

A simulation study in Section 4 explores conditions under which the proposed new estimator $\hat{d}_{WS}$ may perform better than $\hat{d}_{mS}$. This assessment considers both the estimation of d as such, and the performance of confidence intervals for Y. Section 5 applies the proposed estimator to four variables from NHANES III, with emphasis on cases for which differences between the proposed estimators $\hat{d}_{WS}$ and $\hat{d}_{mS}$ have a substantial practical effect on assessment of the stability of the variance estimator $\hat{V}(\hat{Y})$. Section 6 reviews the methods developed in this paper and considers some possible extensions.

2. An estimator based on auxiliary information

2.1 A within-PSU variance estimator

A simple estimator of V_{Wh} is

$$\hat{V}_{Wh} = n_h^{-2}\sum_{i=1}^{n_h} p_{hi}^{-2}\,\hat{\sigma}_{2hi}^2, \tag{2.1}$$

where $\hat{\sigma}_{2hi} = n_{hi}^{-1}(n_{hi} - 1)^{-1}\sum_{j=1}^{n_{hi}}(p_{hij}^{-1}\hat{Y}_{hij} - \hat{Y}_{hi})^2$. Note that $\hat{\sigma}_{2hi}^2$ is approximately unbiased for σ_{2hi}^2 under a with-replacement sampling design within PSU i in stratum h; or under simple random sampling without replacement and with a small sampling fraction, $f_{hi} = N_{hi}^{-1} n_{hi}$. Standard sampling theory shows that $\hat{V}_{Wh}$ is approximately unbiased for V_{Wh}. Then an approximately unbiased estimator of $\mathrm{Var}(\hat{V}_{Wh})$ is

$$\widehat{\mathrm{Var}}(\hat{V}_{Wh}) = n_h^{-1}(n_h - 1)^{-1}\sum_{i=1}^{n_h}(\hat{V}_{Whi} - \hat{V}_{Wh})^2, \tag{2.2}$$

where $\hat{V}_{Whi} = n_h^{-1} p_{hi}^{-2}\hat{\sigma}_{2hi}^2$; see, *e.g.*, Eltinge and Jang (1996) and references cited therein. Note that the overall stratum-level variance estimators $\hat{V}_h$ are functions of the sample means of $p_{hij}^{-1}\hat{Y}_{hij}$ over PSUs in stratum h. In addition, the estimators $\hat{V}_{Wh}$ are functions of sample variances of the $p_{hij}^{-1}\hat{Y}_{hij}$ within the PSU (h, i). Thus, for variables Y for which $p_{hij}^{-1}\hat{Y}_{hij}$ are approximately normally distributed within stratum h, the estimators $\hat{V}_h$ and $\hat{V}_{Wh}$ are approximately independent.

2.2 Properties of $\hat{d}_{WS}$

In the remainder of this paper, the estimator $\hat{d}_{WS}$ defined in expression (1.8) will use $\widehat{\mathrm{Var}}(\hat{V}_{Wh})$ as defined in expression (2.2). Also, the remainder of this paper will use several asymptotic results. These results will use the condition that the number of strata, L, is increasing, while stratum-level PSU and SSU sample sizes n_h and m_h are allowed to remain small. This is in keeping with many practical multi-stage designs that use $n_h = 2$ and moderate values of m_h. See, *e.g.*, Krewski and Rao (1981) for a detailed development of large-L asymptotic results. The proof of Result 2.1 is routine and is thus omitted.

Result 2.1. Assume that $E(\hat{V}_{Wh}^r) = O(1)$ for $r = 1, 2, 3, 4$ and define

$$\hat{\overline{V}}_W = L^{-1}\sum_{h=1}^{L}\hat{V}_{Wh}$$

and (2.3)

$$\hat{\overline{V}}_{w(2)} = L^{-1}\sum_{h=1}^{L}(n_h - 1)^{-1}\{\hat{V}_{Wh}^2 - \widehat{\mathrm{Var}}(\hat{V}_{Wh})\}.$$

Then $\hat{\bar{V}}_W$ and $\hat{\bar{V}}_{W(2)}$ are consistent estimators of $\bar{V}_W$ and $L^{-1}\sum_{h=1}^{L}(n_h-1)^{-1}V_{Wh}^2$, respectively. In addition, $L^{-1}\hat{d}_{WS}$ is a consistent estimator of $L^{-1}d_{WS}$.

Section 1 suggested that in some cases, the auxiliary-data based estimator $\hat{d}_{WS}$ might be more stable than the modified Satterthwaite estimator $\hat{d}_{mS}$. To examine this idea, we will compare the variances of $\hat{d}_{WS}$ and $\hat{d}_{mS}$ under condition (C.1) and the following additional assumptions.

(C.2) For $h = 1, 2, ..., L$, $V_{Wh}^{-1}(m_h-1)\hat{V}_{Wh}$ are distributed as independent chi-square random variables with $m_h - 1$ degrees of freedom, respectively, where m_h is the number of SSUs in stratum h; and are mutually independent of $\hat{V}_h$.

(C.3) For all $h = 1, 2, ..., L$, $n_h = 2$; and $m_h = m_0$ for some fixed positive integer $m_0 \geq 2$.

Arguments similar to those for condition (C.1) indicate that condition (C.2) may be satisfied approximately if within a given PSU (h, i), the m_h random variables $p_{hij}^{-1}\hat{Y}_{hij}$ are approximately independent and identically distributed normal random variables. Condition (C.3) restricts attention to the common case $n_h = 2$. In addition, condition (C.3) requires that an equal number, m_0, of secondary units be selected within each selected PSU. This allows simplification of the resulting approximations for the variances of $\hat{d}_{WS}$, as presented in Result 2.2.

Result 2.2. Assume conditions (C.1), (C.2), (C.3), and (1.6), and define $a = 4\mu_{A_2}^2\mu_{B_2}^{-2}\mathrm{Var}(A_2)$, $b = 4\mu_{A_2}^3\mu_{B_2}^{-3}\mathrm{Cov}(A_2, B_2)$, and $c = \mu_{A_2}^4\mu_{B_2}^{-4}\mathrm{Var}(B_2)$, where $A_2 = L^{-1}\sum_{h=1}^{L}\hat{V}_{Wh}$, $B_2 = L^{-1}\sum_{h=1}^{L}\{\hat{V}_{Wh}^2 - \widehat{\mathrm{Var}}(\hat{V}_{Wh})\}$, $\mu_{A2} = \bar{V}_W$ and $\mu_{B2} = L^{-1}\sum_{h=1}^{L}V_{Wh}^2$. Then

(i) the variances of the leading terms in Taylor expansions of $L^{-1}(\hat{d}_{WS} - d)$ and $L^{-1}(\hat{d}_{mS} - d)$ are, respectively,

$$\mathrm{V}_{LW} = a - b + c$$

and

$$\mathrm{V}_{Lm} = \frac{1}{9}\left(\frac{9L}{3L+14}\right)^2(m_0-1)\left\{a - b + \frac{4(m_0-1)}{3(m_0+2)}c\right\}.$$

(ii) for all $m_0 \geq \lim_{L\to\infty} g(a,b,c)$, $\lim_{L\to\infty} V_{Lm} \geq \lim_{L\to\infty} V_{LW}$ where

$$g(a, b, c) = \{2(3a-3b+4c)\}^{-1}\left\{11c + \sqrt{144a^2+144b^2+153c^2-288ab+216ac-216bc}\right\}.$$

(iii) for $m_0 \geq 10$, $\lim_{L\to\infty} V_{Lm} \geq \lim_{L\to\infty} V_{LW}$ regardless of the values of the limiting moments $\lim_{L\to\infty}(\mu_{A_2}, \mu_{B_2}, L^{-1}\sum_{h=1}^{L}V_{Wh}^3, L^{-1}\sum_{h=1}^{L}V_{Wh}^4)$.

Result 2.2 indicates that for large L, $\hat{d}_{WS}$ may be preferable to $\hat{d}_{mS}$, provided: (1) the proportionality condition (1.6) is satisfied; and (2) the secondary unit sample size m_0 exceeds the lower bound given by $g(a,b,c)$ (thus ensuring relatively small variances of the $\hat{V}_{Wh}$). This motivates the use of within-PSU variances to assess the stability of survey variance estimators, especially under sample designs with small numbers of PSUs per stratum. For some additional discussion of this point, and some specific diagnostics to check the stability of $\hat{V}_{Wh}$, see Eltinge and Jang (1996) and references cited therein. For the four cases considered in Table 1.1 and studied further in Section 4 below, $g(a,b,c)$ is equal to 4.7, 4.3, 4.6, and 4.8 respectively, while the NHANES III application had the mean of the m_h values approximately equal to 22. In addition, we are treating V_{Wh} values as fixed, and Result 2.2 depends on the limiting moments of these V_{Wh} terms. Suppose that $V_{Wh}/\bar{V}_W$ had the same moments as F/f, where F follows a chi-square distribution on f degrees of freedom. Then $f = \infty$ corresponds to the case in which $V_{Wh} = \bar{V}_W$ for all h, which corresponds to the case in which the true d in (1.1) equals the customary value of $n - L$.

3. Testing the proportionality condition

3.1 An errors-in-variables model for V_h and V_{Wh}

Development of the alternative estimator $\hat{d}_{WS}$ in Section 1, and evaluation of its properties in Section 2, depended heavily on the proportionality condition (1.6). One may test the adequacy of this condition through the following steps. First, note that condition (1.6) is a special case of the following model,

(C.4) For all $h = 1, 2, ..., L$,

$$V_h = \beta_0 + \beta_1 V_{Wh} + q_h \qquad (3.1)$$

where β_0 and β_1 are constants, and q_h is an equation error with mean zero and variance σ_{qqh}.

Second, recall that V_h and V_{Wh} are unknown quantities, for which we have the unbiased estimators $\hat{V}_h$ and $\hat{V}_{Wh}$, respectively. Using the errors-in-variables model notation in Fuller (1987), define the estimation errors

$$e_h = \hat{V}_h - V_h \quad \text{and} \quad u_h = \hat{V}_{Wh} - V_{Wh}. \qquad (3.2)$$

Under conditions (C.1) and (C.2), the vector $(e_h, u_h)'$ is distributed with a mean vector equal to $(0,0)'$ and a variance-covariance matrix equal to $\text{diag}(\sigma_{eeh}, \sigma_{uuh})$ where $\sigma_{eeh} = (n_h - 1)^{-1} 2V_h^2$ and $\sigma_{uuh} = (m_h - 1)^{-1} 2V_{wh}^2$. Under the additional condition (C.3), these variance terms simplify to $\sigma_{eeh} = 2V_h^2$ and $\sigma_{uuh} = (m_0 - 1)^{-1} 2V_{Wh}^2$.

Expressions (3.1) and (3.2) define an errors-in-variables regression model with heterogeneous measurement error variances and non-normal errors. In addition $\widehat{\text{Var}}(\hat{V}_{Wh})$ defined in expression (2.2) is an unbiased estimator of σ_{uuh}, and thus provides identifying information for the parameters β_0, β_1 and σ_{qqh} in model (3.1) – (3.2). A direct application of Fuller (1987, pages 187-189) with equal weights then gives the consistent estimators (for increasing L),

$$\hat{\beta}_0 = L^{-1} \sum_{h=1}^{L} \hat{V}_h - \hat{\beta}_1 \hat{\bar{V}}_W,$$

$$\hat{\beta}_1 = \left[\sum_{h=1}^{L} (\hat{V}_{Wh} - \hat{\bar{V}}_W)^2 - \hat{\sigma}_{uu.} \right]^{-1} \sum_{h=1}^{L} (\hat{V}_{Wh} - \hat{\bar{V}}_W) \hat{V}_h, \quad (3.3)$$

and

$$\hat{\bar{\sigma}}_{qq.} = \max \Big[0,\ L^{-1} \sum_{h=1}^{L} (n_h - 1)^{-1} \{ (L-2)^{-1} L (\hat{V}_h - \hat{\beta}_0 - \hat{\beta}_1 \hat{V}_{Wh})^2 - (\hat{\sigma}_{eeh} + \hat{\beta}_1^2 \hat{\sigma}_{uuh}) \} \Big], \quad (3.4)$$

where

$$\hat{\sigma}_{uu.} = \textstyle\sum_{h=1}^{L} \widehat{\text{Var}}(\hat{V}_{Wh}),\ \hat{\bar{V}}_W = L^{-1} \textstyle\sum_{h=1}^{L} \hat{V}_{Wh},$$

and (3.5)

$$\hat{\sigma}_{eeh} = 2(n_h + 1)^{-1} \hat{V}_h^2$$

from condition (C.1). In addition, direct application of Fuller (1987, page 188) leads to variance estimators $\hat{V}(\hat{\beta}_0)$ and $\hat{V}(\hat{\beta}_1)$, say; details are available from the authors.

3.2 Two related diagnostics

In keeping with condition (C.4), the proposed estimator $\hat{d}_{WS}$ is intended for cases in which the $\hat{V}_{Wh}$ provide useful auxiliary information on the relative magnitudes of the overall stratum-level variances V_h. To identify such cases, one simple diagnostic is the ratio $\{\hat{V}(\hat{V}_h)\}^{-1} \{\hat{\beta}_1^2 \hat{V}(\hat{V}_{Wh}) + \hat{\sigma}_{qqh}\}$, *i.e.*, the ratio of estimators of the variances of the approximate distributions of $\hat{V}_h - V_h$ and $\beta_1 \hat{V}_{Wh} - V_h$, respectively, under model (3.1) – (3.2). If this ratio is substantially less than unity, then use of $\hat{d}_{WS}$ may be indicated.

In addition, the performance of the estimator $\hat{d}_{WS}$ depends heavily on the magnitude of $\hat{\sigma}_{uu.}$ relative to the variability of the true within-PSU variances V_{Wh}. Define an estimator of the reliability ratio (Fuller 1987, page 3)

$$\hat{\kappa}_{xx} = \max \left\{ 0, \left[\sum_{h=1}^{L} (\hat{V}_{Wh} - \hat{\bar{V}}_W)^2 \right]^{-1} \left[\sum_{h=1}^{L} (\hat{V}_{Wh} - \hat{\bar{V}}_W)^2 - \hat{\sigma}_{uu.} \right] \right\}.$$

The values of $\hat{\kappa}_{xx}$ are between 0 and 1; and values of $\hat{\kappa}_{xx}$ close to unity indicate relatively small errors in the estimation of within-PSU variances. Conversely, small values of $\hat{\kappa}_{xx}$ (*e.g.*, $\hat{\kappa}_{xx} < 0.7$) may indicate that the methods of Sections 3.1 – 3.2 may not perform well, due to the relatively large sampling errors in the auxiliary information $\hat{V}_{Wh}$. The numerical work in Sections 4 and 5 below will consider these diagnostics further.

The work in this section is based on the assumption that $\sigma_{qq.} > 0$. One may develop related diagnostics applicable to the case of no equation errors, *i.e.*, $\sigma_{qq.} = 0$; details are available from the authors.

4. A simulation study

4.1 Design of the study

We now use a simulation study to evaluate the properties of our degrees-of-freedom estimators, and related variates, under moderate-sample-size conditions. We set up the simulation procedure as follows.

We considered four sets of V_h values from the NHANES III example for the Mexican-American subpopulation introduced in Section 1.1. Those four sets of V_h are the estimated $\hat{V}_h$ values from the variables BMPWT, HAR3, TCRESULT and HDRESULT, respectively, and are listed in Table 4.1. For each case, we used $(\beta_0, \beta_1) = (0, 1)$ and $\sigma_{qq} = 0$, in keeping with the results of Section 3, and thus $V_{Wh} = V_h$. Then, for each $h = 1, ..., L$, we obtained 10,000 realizations of the initial estimators $(\hat{Y}_{h1}, \hat{Y}_{h2}, \hat{V}_{Wh1}, \hat{V}_{Wh2})$ by assuming that the $\hat{Y}_{hi}$ are distributed as a normal random variable with mean zero and variance $2^{-1} V_h$; that $V_{Wh}^{-1}(m_{hi} - 1)\hat{V}_{Whi}$ is distributed as a chi-square random variable with $m_{hi} - 1$ degrees of freedom, where $m_{hi} = 11$ for all h and i; and the $\hat{Y}_{hi}$ and $\hat{V}_{Whi}$ are mutually independent. Note that in our data from NHANES III, the average number of secondary units for each PSU i in stratum h is about 11. For each replication, we computed $\hat{V}_h = (\hat{Y}_{h1} - \hat{Y}_{h2})^2$ and $\hat{V}_{Wh} = 2^{-1}(\hat{V}_{Wh1} + \hat{V}_{Wh2})$, and then carried out an errors-in-variables regression of $\hat{V}_h$ on $\hat{V}_{Wh}$ with measurement error variance $\hat{\sigma}_{uuh} = \widehat{\text{Var}}(\hat{V}_{Wh})$ using formula (2.2). This produced the coefficient estimators $(\hat{\beta}_0, \hat{\beta}_1)$, and the degrees-of-freedom estimators $\hat{d}_{mS}$ and $\hat{d}_{WS}$.

Table 4.1
"True" variances V_h used in simulation studies

Stratum	Case 1	Case 2	Case 3	Case 4
1	0.00E+00	0.00E+00	0.00E+00	0.00E+00
2	0.00E+00	0.00E+00	0.00E+00	0.00E+00
3	1.56E-04	7.67E-05	1.45E-02	1.76E-02
4	2.01E-04	3.57E-06	5.60E-02	4.55E-03
5	2.82E-04	4.88E-07	1.54E-03	2.91E-03
6	4.36E-04	0.00E+00	3.73E-03	8.60E-04
7	7.30E-04	2.14E-06	1.69E-02	1.13E-05
8	8.80E-04	1.30E-05	2.72E-02	1.40E-03
9	1.65E-03	1.16E-06	9.24E-03	1.35E-04
10	1.70E-03	9.46E-07	2.24E-03	1.77E-03
11	2.73E-03	0.00E+00	2.54E-04	1.32E-03
12	2.91E-03	5.40E-06	2.75E-02	6.40E-03
13	4.95E-03	3.73E-07	1.15E-02	5.38E-03
14	7.25E-03	2.90E-04	3.75E-02	6.97E-02
15	9.06E-03	9.81E-05	3.46E-01	7.58E-01
16	1.14E-02	7.47E-06	1.54E-02	4.75E-03
17	2.69E-02	9.65E-05	7.99E-02	1.01E-03
18	4.00E-02	1.12E-04	1.44E-01	1.77E-01
19	4.27E-02	2.68E-06	8.59E-02	3.88E-02
20	6.05E-02	7.57E-06	2.68E+00	7.18E-02
21	6.45E-02	1.17E-04	1.65E-01	4.52E-04
22	1.08E-01	1.05E-04	5.41E-01	1.98E-03

4.2 Coverage rates of *t*-based confidence intervals

For the four specified cases, Table 4.2 presents the simulated non-coverage probabilities obtained for t-based confidence intervals for the population mean $\overline{Y}$ that used the corresponding $\hat{d}$. For the severely heterogeneous cases (Cases 3 and 4), none of the degrees of freedom measures (not even the true d) leads to confidence intervals with coverage rates meeting the nominal rates $1 - \alpha$. That is, in extreme cases, the general Satterthwaite approach can be problematic for construction of confidence intervals, regardless of whether d, $\hat{d}_{mS}$, or $\hat{d}_{WS}$ is used to determine the t multiplier.

For Cases 1 and 2, the V_h values display less severe heterogeneity than in Cases 3 and 4. Table 4.2 shows that the simulated coverage probabilities with the true d for these two cases are slightly above 0.95. This overcoverage may be attributable to the fact that the variance estimator $\hat{V}(\hat{Y})$ is not distributed exactly as a multiple of a χ^2_d random variable, due to the heterogeneity of the V_h. Use of the standard degrees-of-freedom term $n - L$ or the modified estimator $\hat{d}_{mS}$ produces confidence intervals with coverage rates below the nominal level of 95%. On the other hand, use of our auxiliary-data-based term $\hat{d}_{WS}$ gives simulation based coverage rates close to the nominal 0.95 level.

Tables 4.3a and 4.3b display the empirical distributions of $\hat{d}$ and $2t_{\hat{d}}$ for the estimators $\hat{d}_{mS}$ and $\hat{d}_{WS}$. The simulated standard deviation of $t_{\hat{d}_{WS}}$ is smaller than that of $t_{\hat{d}_{mS}}$. In addition, the mean and median of $t_{\hat{d}_{WS}}$ are slightly larger than those of $t_{\hat{d}_{mS}}$. This is consistent with the undercoverage of the intervals based on $t_{\hat{d}_{mS}}$. Thus, under conditions similar to those for Cases 1 and 2 (or under conditions with less heterogeneity of V_h), it is worthwhile to consider the use of $\hat{d}_{WS}$ as a degrees-of-freedom estimator.

5. Application to a health survey

5.1 Preliminary model checks

We applied our proposed methods to the NHANES III data described in Section 1. It is important to check the modeling assumptions before we apply the proposed stability measures. First, for the Mexican-American subpopulation described in Section 1, Table 5.1 gives values of $\hat{\kappa}_{xx}$ for the four variables which all have $\hat{\kappa}_{xx}$ values greater than 0.7.

Second, Figure 5.1 displays the scatter plots of $\hat{V}_h$ against $\hat{V}_{Wh}$ for the four variables with equal scales used for the horizontal and vertical axes. It shows that a linear relationship for the corresponding variables is plausible even if the relation would not be perfect and there are some outliers. Consequently, those four variables might be appropriate for the auxiliary-data-based method developed in Sections 2 and 3.

Table 4.2
Observed non-coverage rates for nominal 95% confidence intervals with $V_h = V_{Wh}$ in simulation study

	Case 1	Case 2	Case 3	Case 4
True d_S	6.26	6.04	2.38	2.20
Non-Coverage with t_{d_s}	0.0428	0.0443	0.0162	0.0164
Non-Coverage with t_{n-L}	0.0744	0.0788	0.1220	0.1263
Non-Coverage with $t_{\hat{d}_{mS}}$	0.0552	0.0567	0.0911	0.0905
Non-Coverage with $t_{\hat{d}_{WS}}$	0.0428	0.0466	0.0224	0.0220

Table 4.3a
Means and quantiles of degrees-of-freedom estimators $\hat{d}_{mS}$ and $\hat{d}_{WS}$: Cases 1 and 2

Cases	True d	Est.	[1]Mean $\hat{d}$	SD($\hat{d}$)	[2]Q(0.05)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.95)
1	6.26	$\hat{d}_{mS}$	9.33	3.33	4.45	6.86	9.01	11.41	15.30
		$\hat{d}_{WS}$	6.52	0.82	5.06	5.99	6.57	7.10	7.78
2	6.04	$\hat{d}_{mS}$	8.87	2.95	4.35	6.69	8.72	10.97	13.99
		$\hat{d}_{WS}$	6.34	0.96	4.67	5.69	6.42	7.06	7.80

[1] Mean denotes the average of the estimates, taken across all 10,000 replications.
[2] Q (·) indicates the quantile of the estimator, taken across all 10,000 replications.

Table 4.3b
Simulated non-coverage probabilities; and means and quantiles of *t*-multipliers for nominal 95% confidence intervals: Unequal true variances, cases 1 and 2

Cases	Est.	[1] $1-\hat{\alpha}$	[2] M($2t_\alpha$)	$SD(2t_\alpha)$	[3]Q(0.05)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.95)
1	$\hat{d}_{mS}$	0.0552	4.62	0.36	4.26	4.38	4.52	4.75	5.37
	$\hat{d}_{WS}$	0.0428	4.83	0.16	4.64	4.72	4.80	4.90	5.13
	n - L	0.0744	4.15						
	True d_S	0.0428	4.85						
2	$\hat{d}_{mS}$	0.0567	4.66	0.36	4.29	4.41	4.55	4.78	5.41
	$\hat{d}_{WS}$	0.0466	4.87	0.21	4.64	4.72	4.83	4.97	5.28
	n - L	0.0788	4.15						
	True d_S	0.0443	4.89						

[1] $1-\hat{\alpha}$ is the simulated non-coverage probability of confidence intervals computed using estimated d.f.'s
[2] M($2t_{0.975}$) is the average of twice of the 97.5% *t*-percentile value
[3] Q (.) indicates the quantile of $2t_{0.975,\hat{d}}$, taken across all replications.

Table 5.1
$\hat{\kappa}_{xx}$, estimates of model parameters, model diagnostics, and degrees of freedom estimates for four NHANES III variables (Mexican-American (Age 20-29) subgroup)

Variables	$\hat{\kappa}_{xx}$	$\tilde{\beta}_0$	se($\tilde{\beta}_0$)	$\tilde{\beta}_1$	se($\tilde{\beta}_1$)	Simulation based p-value for H_0: $\beta_0 = 0$	Simulation based p-value for H_0: $\beta_1 = 1$	$\hat{\sigma}_{qq.}$	$\hat{r}_{qq}$	$\hat{d}_{mS}$	$\hat{d}_{WS}$
BMPWT	0.75	-0.0013	0.0039	1.135	0.5429	0.3815	0.3541	-0.000	-0.43	15.49	10.04
HAR3	0.75	-0.000009	0.000012	1.095	0.3991	0.4229	0.3400	0.000	-0.83	14.94	8.30
TCRESULT	0.88	-0.146	0.0493	2.879	0.6252	0.0606	0.2259	-0.178	-0.77	5.88	6.59
HDRESULT	0.90	-0.042	0.0098	6.650	0.9988	<0.0001	0.1506	-0.017	-0.91	5.45	5.93

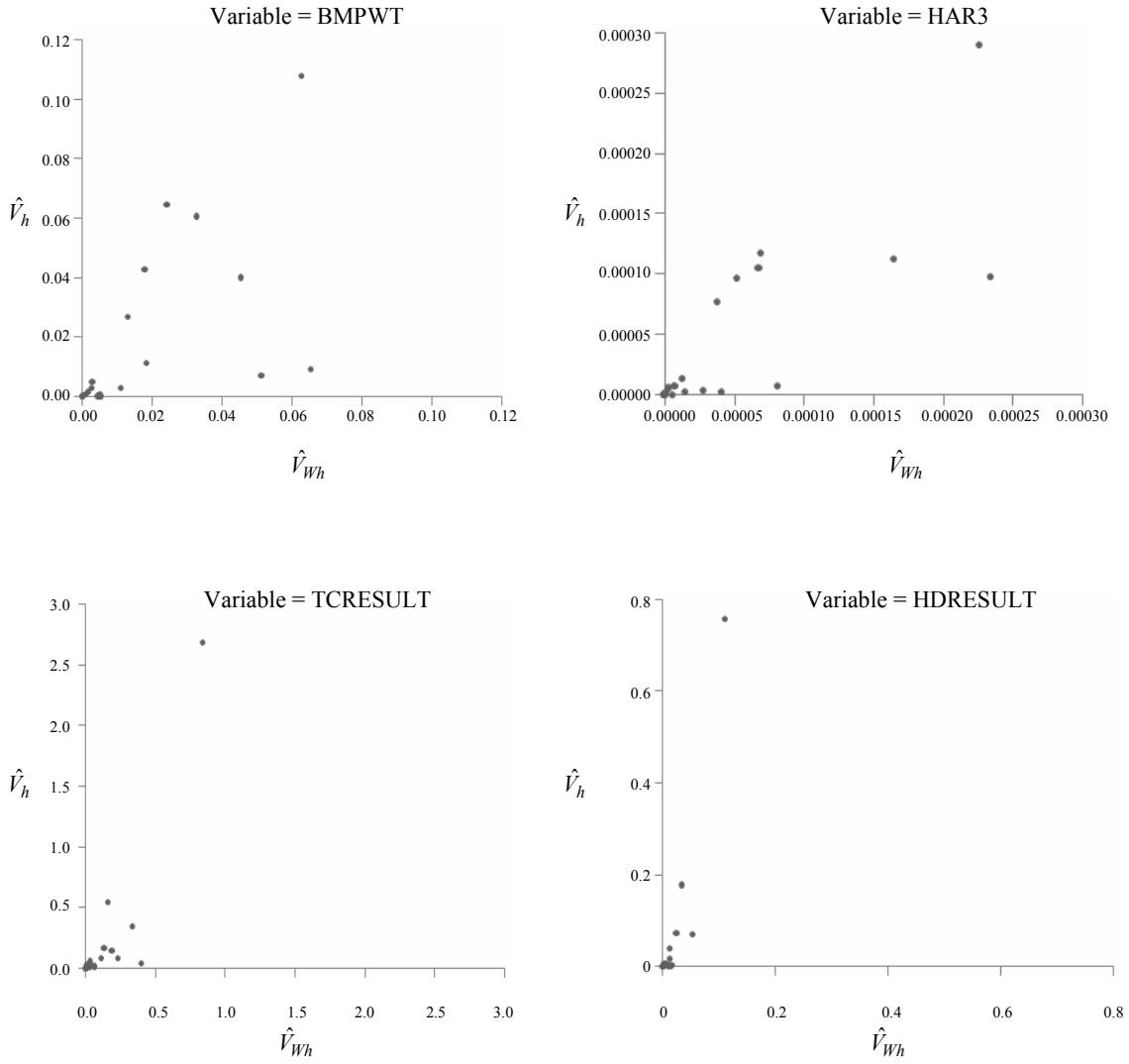

Figure 5.1 Plot of $\hat{V}_{Wh}$ vs. $\hat{V}_h$ for M-A (Age 20-29), Variable = BMPWT

5.2 An *ad hoc* test of $\bar{\sigma}_{qq.} = 0$ under condition (C.1)

For all four variables considered in Table 5.1, the direct estimates $\hat{\bar{\sigma}}_{qq}$ of equation error variance (3.4) were negative or close to zero. That suggests that our χ^2-based estimator of σ_{eeh} as given in Section 3.1 might be too conservative or that $\bar{\sigma}_{qq.}$ is indeed close to zero. This suggests that we need to re-examine the distributional assumption (C.1) in the NHANES III example. To do this, we considered the simulated distribution of $\hat{r}_{qq} \overset{\text{def}}{=} \hat{\bar{\sigma}}_{qq}/\hat{\sigma}_{ee}$, where division by $\hat{\sigma}_{ee}$ is used to avoid scale problems. The conditions and simulation design were as described in Section 4.1.

Table 5.2 reports results for $\hat{\sigma}_{ee}$ from expression (3.5), and $\hat{\bar{\sigma}}_{qq}$ computed from expression (3.4) with $\hat{\beta}_0$ set equal to zero and with $\hat{\beta}_1$, computed from expression (3.3). Table 5.2 reports the mean, standard deviation and selected quantiles of the simulated distribution of $\hat{r}_{qq.}$ for the four variables. Table 5.3 reports the corresponding quantities for $\hat{r}_{gq}$, computed from $\hat{\bar{\sigma}}_{qq}$ given by expression (3.4) and with $\hat{\beta}_0$ and $\hat{\beta}_1$ computed from expression (3.3).

The results reported in Tables 5.2 and 5.3 lead to an ad hoc test of H_0: $\sigma_{qq} = 0$. Specifically, if the observed ratio $\hat{r}_{qq}$ falls above the upper 0.95 simulated quantile, then the assumption that $\bar{\sigma}_{qq} = 0$ may be problematic. Conversely, an observed $\hat{r}_{qq}$ below the .05 simulated quantiles in Tables 5.2 or 5.3 might indicate that $\hat{\sigma}_{eeh}$ is conservative, or may indicate violation of other parts of condition (C.1).

From Table 5.1, the values of $\hat{r}_{qq}$ for the variables are between -0.91 to -0.43. Except for HDRESULT, we do not have any strong evidence of violation of the model assumptions. However, for HDRESULT, the ratio $\hat{r}_{qq}$ = -0.91 falls between the 0.01 and 0.05 quantiles reported in Table 5.2 and 5.3 for case 4. In general, values of $\hat{r}_{qq}$ that fall above the 0.95 or 0.99 quantiles of Tables 5.2 or 5.3 would be consistent with values of $\bar{\sigma}_{qq.}$ greater than zero. The observed value $\hat{r}_{qq}$ = -0.91 is not necessarily consistent with $\bar{\sigma}_{qq.} > 0$, but may indicate violation of one or more conditions in (C.1)-(C.4).

Table 5.2
Means and quantiles of $\hat{r}_{qq} = \hat{\sigma}_{ee}^{-1}\hat{\sigma}_{qq.}$ ($\hat{\beta}_0 = 0$)

Cases	[1]M($\hat{r}_{qq}$)	SD($\hat{r}_{qq}$)	[2]Q(0.01)	Q(0.05)	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)	Q(0.95)	Q(0.99)
1	-0.50	0.66	-1.71	-1.30	-1.15	-0.99	-0.79	0.16	0.54	0.60	0.65
2	-0.48	0.68	-1.72	-1.32	-1.16	-0.99	-0.76	0.23	0.57	0.62	0.66
3	-0.19	0.42	-1.01	-0.84	-0.74	-0.53	-0.20	0.17	0.38	0.46	0.55
4	-0.20	0.39	-1.00	-0.82	-0.72	-0.51	-0.20	0.11	0.34	0.44	0.56

[1] M denotes the average of the estimates, taken across all 10,000 replications.
[2] Q(.) indicates the quantile of the estimator, taken across all 10,000 replications.

Table 5.3
Means and quantiles of $\hat{r}_{qq}=\hat{\sigma}_{ee}^{-1}\hat{\bar{\sigma}}_{qq.}$

Cases	[1]M($\hat{r}_{qq}$)	SD($\hat{r}_{qq}$)	[2]Q(0.01)	Q(0.05)	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)	Q(0.95)	Q(0.99)
1	-0.56	0.62	-1.85	-1.34	-1.17	-1.00	-0.80	0.05	0.38	0.44	0.52
2	-0.56	0.62	-1.91	-1.37	-1.18	-1.00	-0.78	0.06	0.35	0.42	0.50
3	-0.24	0.42	-1.16	-0.90	-0.79	-0.57	-0.22	0.12	0.29	0.36	0.45
4	-0.24	0.38	-1.09	-0.87	-0.75	-0.53	-0.22	0.06	0.25	0.33	0.44

[1] M denotes the average of the estimates, taken across all 10,000 replications.
[2] Q(.) indicates the quantile of the estimator, taken across all 10,000 replications.

5.3 Coefficient estimates and degrees-of-freedom estimates

Because our data were consistent with $\bar{\sigma}_{qq.} = 0$ for all four cases, we used the methods of Fuller (1987, page 124) to produce estimates of β_0 and β_1 appropriate for a model (3.1) – (3.2) with no equation error; details are available from the authors. Table 5.1 also reports the resulting coefficient estimates $\tilde{\beta}_0$ and $\tilde{\beta}_1$, and their standard errors, $se(\tilde{\beta}_0)$ and $se(\tilde{\beta}_1)$. Recall from Section 3.1 that under model (3.1) – (3.2), if $\beta_0 = 0$ and $\beta_1 \neq 0$, then each stratum variance V_h is a constant multiple of the within-PSU variance V_{Wh}, and $\hat{d}_{WS}$ in (1.8) may be an appropriate estimator of d. Section 5.2 already considered the condition $\bar{\sigma}_{qq.} = 0$. To test the null hypothesis H_0: $\beta_0 = 0$, we use the test statistic, $t_0 = \tilde{\beta}_0 / se(\tilde{\beta}_0)$. In some practical errors-in-variables work, quantities like t_0 are compared with a standard normal or t reference distribution. However, simulation work based on the four cases from Section 4.1 indicated that the null distribution of t_0 deviated substantially from these customary reference distributions. This is due to the very skewed distributions of the response variables $\hat{V}_h$ used in the errors-in-variables regression. Consequently, we used standard methods to develop a simulation-based reference distribution for t_0. Column 7 of Table 5.1 reports the resulting left-tailed p-value. (Due to negative point estimates $\tilde{\beta}_0$, we have chosen to report the left-tailed p-values here. In other cases, it may be of interest to report right-tailed or two-tailed p-values for β_0). There is strong evidence against H_0: $\beta_0 = 0$ for the variable HDRESULT, and the moderate evidence against H_0: $\beta_0 = 0$ for TCRESULT. Thus, it may not be appropriate to use $\hat{d}_{WS}$ for these two variables. Now consider the slope coefficient β_1, and suppose that $\sigma_{qqh} = 0$ so $q_h = 0$ with probability one. Then expressions (1.5) and (3.1), and the nonnegativity of V_{Bh} implies that $0 \leq V_{Bh} = V_h - V_{Wh} = \beta_0 + (\beta_1 - 1)V_{Wh}$. Consequently, if $\beta_0 = 0$, then $\beta_1 \geq 1$ and $\beta_1 = 1$ is equivalent to $V_h = V_{Wh}$. This final condition is of practical interest because some authors have noted cases in which V_{Bh} is small relative to V_{Wh}, or equivalently, $V_h \doteq V_{Wh}$. See for example, Wolter (1985, page 46). To test H_0: $\beta_1 = 1$ against the one-sided alternative H_1: $\beta_1 > 1$, we used the statistic $t_1 = (\tilde{\beta}_1 - 1)/se(\tilde{\beta}_1)$. For reasons similar to those for t_0, we developed simulation-based reference distributions for t_1 under each of Cases 1 through 4. Column 8 of Table 5.1 reports the resulting one-tailed p-values.

The last two columns of Table 5.1 report the degree-of-freedom estimators $\hat{d}_{mS}$ and $\hat{d}_{WS}$. For HAR3 and BMPWT, $\hat{d}_{mS}$ gives substantially larger values than $\hat{d}_{WS}$.

6. Discussion

This paper has considered estimation of a degrees-of-freedom term d used to quantify the variability of a standard design based variance estimator $\hat{V}(\hat{Y})$. The fundamental issue is that under a design involving heterogeneous stratum-level variances and small numbers of primary sample units selected per stratum, the Satterthwaite-type estimator $\hat{d}_{mS}$ may perform poorly. We developed an alternative estimator $\hat{d}_{WS}$ based on within-primary-sample unit variance estimators $\hat{V}_{Wh}$. This alternative estimator is a solution to an unbiased estimating equation (1.1) for d, provided the proportionality condition (1.6) is satisfied. Also, the variance of the approximate distribution of $\hat{d}_{WS}$ is smaller than that of $\hat{d}_{mS}$, provided the number of secondary sample units selected within each primary unit is large, in the sense defined by Result 2.2.

Section 3 developed errors-in-variables methods for testing the adequacy of the proportionality condition (1.6), and suggested some related diagnostics. The simulation study in Section 4, in conjunction with the data analysis in Section 5, indicated that under moderate amounts of heterogeneity, $\hat{d}_{WS}$ can perform better than $\hat{d}_{mS}$, in terms of the distributional properties of these estimators of d, and in terms of the coverage rates and widths of associated confidence intervals for the population totals Y. However, as one would expect from standard large-sample theory, neither estimator performs well under severe heterogeneity.

One could in principle consider use of the errors-in-variables estimators ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}_{qq.}$), in conjunction with the $\hat{V}_h$ and $\hat{V}_{Wh}$, to construct an alternative estimator of d that will be consistent under the general errors-in-variables model (3.1) - (3.2), and will not require the restrictive condition (1.6). However, simulation results in Jang (1996) indicated that the resulting estimator $\hat{d}_{EIV}$, say, did not perform well under the design conditions used in Section 5.

The principal results of Sections 1 through 3 extend readily from the within-primary-unit variances V_{Wh} to more general auxiliary variables X_h. For such extensions, the principal issues remain the adequacy of the proportionality approximation (1.6); and the amount of sampling error in the auxiliary estimators $\hat{X}_h$, say, relative to the error in the basic stratum-level variance estimator $\hat{V}_h$.

Acknowledgements

The authors thank the U.S. National Center for Health Statistics for providing access to the NHANES III dataset, and thank V.L. Parsons, C. Johnson and L.R. Curtin for sharing a wealth of information regarding the NHANES III. This research was supported in part by the U.S. National Center for Health Statistics. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the U.S. National Center for Health Statistics or the U.S. Bureau of Labor Statistics.

Appendix A

Proof of result 2.2

Consider a nonlinear function $B^{-1}A^2$ of two estimators A and B with means μ_A and μ_B, respectively. Then, the variance of the leading term of a Taylor expansion of $B^{-1}A^2$ is

$$\frac{4\mu_A^2}{\mu_B^2}\mathrm{Var}(A) - 4\frac{\mu_A^3}{\mu_B^3}\mathrm{Cov}(A,B) + \frac{\mu_A^4}{\mu_B^4}\mathrm{Var}(B) \quad \text{(A.1)}$$

Now we define the following two estimators: $L^{-1}\hat{d}_{S1} = B_1^{-1}A_1^2$ and $L^{-1}\hat{d}_{S2} = B_2^{-1}A_2^2$, where $A_1 = L^{-1}\sum_{h=1}^{L}\hat{V}_h$, $B_1 = L^{-1}\sum_{h=1}^{L}\hat{V}_h^2$, $A_2 = L^{-1}\sum_{h=1}^{L}\hat{V}_{Wh}$, and $B_2 = L^{-1}\sum_{h=1}^{L}\{\hat{V}_{Wh}^2 - \widehat{\mathrm{Var}}(\hat{V}_{Wh})\}$.

Assume conditions (C.1), (C.2) and (C.3). In addition, define $\hat{F}_{L\hat{d}_{S1}}$ and $\hat{F}_{L\hat{d}_{S2}}$ to be the leading terms of Taylor expansions of $L^{-1}\hat{d}_{S1} - \mu_{B_1}^{-1}\mu_{A_1}^2$ and $L^{-1}\hat{d}_{S2} - \mu_{B_2}^{-1}\mu_{A_2}^2$, respectively. Also, recall that if D is distributed as a chi-square random variable on d degrees of freedom, then $V(D) = 2d$, $E(D^3) = d(d+2)(d+4)$, and $V(D^2) = 8d(d+2)(d+3)$. Then the corresponding components of $\mathrm{Var}(\hat{F}_{L\hat{d}_{S1}})$ and $\mathrm{Var}(\hat{F}_{L\hat{d}_{S2}})$ in (A.1) are

$$\begin{aligned}
\mathrm{Var}(A_1) &= 2L^{-2}\sum_{h=1}^{L}V_h^2,\\
\mathrm{Var}(A_2) &= 2(m_0-1)^{-1}L^{-2}\sum_{h=1}^{L}V_{Wh}^2\\
\mathrm{Var}(B_1) &= 96L^{-2}\sum_{h=1}^{L}V_h^4,\\
\mathrm{Var}(B_2) &= 8(m_0-1)^{-2}(m_0+1)L^{-2}\sum_{h=1}^{L}V_{Wh}^4\\
\mathrm{Cov}(A_1,B_1) &= 12L^{-2}\sum_{h=1}^{L}V_h^3,
\end{aligned}$$

and (A.2)

$$\mathrm{Cov}(A_2,B_2) = 4(m_0-1)^{-1}L^{-2}\sum_{h=1}^{L}V_{Wh}^3.$$

Since we assume $n_h = 2$ and $m_h = m_0$ for all $h = 1, 2, \ldots, L$, we have

$$L^{-1}\hat{d}_{mS} = L^{-1}(3L+14)^{-1}(9L)\hat{d}_{S1} \quad \text{(A.3)}$$

and

$$L^{-1}\hat{d}_{WS} = L^{-1}\hat{d}_{S2}. \quad \text{(A.4)}$$

Under condition (1.6), $\mu_{A1} = \beta_1 \mu_{A2}$,

$$\mu_{B1} = 3\beta_1^2 \mu_{B2},$$
$$\mathrm{Var}(A_1) = (m_0 - 1)\beta_1^2 \mathrm{Var}(A_2),$$
$$\mathrm{Var}(B_1) = 12(m_0 + 1)^{-1}(m_0 - 1)^2 \beta_1^4 \mathrm{Var}(B_2)$$

and

$$\mathrm{Cov}(A_1, B_1) = 3(m_0 - 1)\beta_1^3 \mathrm{Cov}(A_2, B_2) \qquad \text{(A.5)}$$

Substituting (A.5) into (A.1) leads to,

$$\begin{aligned}
&\mathrm{Var}(\hat{F}_{L\hat{d}_{S1}}) \\
&= \frac{4}{9}(m_0 - 1)\frac{\mu_{A_2}^2}{\mu_{B_2}^2}\mathrm{Var}(A_2) \\
&\quad - \frac{4}{9}\frac{\mu_{A_2}^3}{\mu_{B_2}^3}(m_0 - 1)\mathrm{Cov}(A_2, B_2) \\
&\quad + \frac{4(m_0 - 1)^2}{27(m_0 + 1)}\frac{\mu_{A_2}^4}{\mu_{B_2}^4}\mathrm{Var}(B_2) \\
&= \frac{1}{9}(m_0 - 1)a - \frac{1}{9}(m_0 - 1)b + \frac{4(m_0 - 1)^2}{27(m_0 + 2)}c \qquad \text{(A.6)}
\end{aligned}$$

where $\mathrm{Var}(L^{-1}\hat{d}_{WS}) = a - b + c$. With large L, $\mathrm{Var}(\hat{F}_{L\hat{d}_{mS}}) = (m_0 - 1)a - (m_0 - 1)b + \{3(m_0 + 2)\}^{-1}4(m_0 - 1)^2 c$. Thus for large L, $V(\hat{F}_{L\hat{d}_{mS}}) - V(\hat{F}_{L\hat{d}_{WS}}) \doteq (m_0 - 2)a - (m_0 - 2)b + \{3(m_0 + 2)\}^{-1}(4m_0^2 - 11m_0 - 2)c$. Therefore, $\lim_{L\to\infty} V_{Lm} - \lim_{L\to\infty} V_{LW} \geq 0$ if $m_0 \geq \lim_{L\to\infty}\{2(3a - 3b + 4c)\}^{-1}\{11c + \sqrt{144a^2 + 144b^2 + 153c^2 - 288ab + 216ac - 216bc}\}$. In particular, $\lim_{L\to\infty} V_{Lm} - \lim_{L\to\infty} V_{LW}$ becomes greater than or equal to zero when $m_0 = 10$ regardless of values of a, b, and c. Because it is an increasing function in m_0, for all values of $m_0 \geq 10$, $\lim_{L\to\infty} V_{Lm} \geq \lim_{L\to\infty} V_{LW}$.

References

Casady, R., Dorfman, A.H. and Wang, S. (1998). Confidence intervals for sub-domain parameters when the sub-domain sample size is random. *Survey Methodology*, 24, 57-67.

Cochran, G.C. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons, Inc.

Eltinge, J.L., and Jang, D. (1996). Stability measures for variance component estimators under a stratified multistage design. *Survey Methodology*, 22, 157-165.

Fuller, W.A. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc.

Isaki, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78, 117-123.

Jang, D. (1996). *Stability of Variance Estimators Under Complex Sampling Designs*. Unpublished Ph.D. dissertation, Department of Statistics, Texas A&M University, College Station, Texas.

Kendall, M., Stuart, A. and Ord, J.K. (1983). *The Advanced Theory of Statistics*, *Volume* 3: *Design and Analysis*, *and Time Series*. New York: Macmillan.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jacknife and balanced repeated replication methods. *Ann. Statist*., 9, 1010-1019.

National Center for Health Statistics (1996). NHANES III Reference Manuals and Reports, CD-ROM GPO, 017-022-1358-4. Washington, D.C.: United States Government Printing Office.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.

Smith, H.F. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, 9, 211-212.

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

Survey Methodology, December 2009
Vol. 35, No. 2, pp. 247-260
Statistics Canada, Catalogue No. 12-001-X

Semiparametric regression model for complex survey data

Zilin Wang and David R. Bellhouse [1]

Abstract

A semiparametric regression model is developed for complex surveys. In this model, the explanatory variables are represented separately as a nonparametric part and a parametric linear part. The estimation techniques combine nonparametric local polynomial regression estimation and least squares estimation. Asymptotic results such as consistency and normality of the estimators of regression coefficients and the regression functions have also been developed. Success of the performance of the methods and the properties of estimates have been shown by simulation and empirical examples with the Ontario Health Survey 1990.

Key Words: Complex survey; Domain estimates; Nonparametric regression; Smoothing.

1. Introduction

In practice, many surveys are used to explore a relationship between a response variable and explanatory variables and to build predictive models. Hence, it is necessary to develop techniques that apply stochastic regression models to survey data. Although nonparametric regression techniques have been widely applied in many fields of statistics, not much attention has been paid to them in the field of complex surveys due to the complexity of the data structure. The correlation induced by clustering and unequal probabilities of selection of the sample cause survey data to be neither independent nor identically distributed. As a result, standard nonparametric regression methods are often inappropriate for analyzing sample survey data.

There is some work, for instance Breidt and Opsomer (2000), Montanari and Ranalli (2005), and Zheng and Little (2004), on nonparametric regression techniques that have been developed for survey data. However, as in the conventional way of applying regression techniques, most of this work uses model-assisted approaches to estimate descriptive population quantities and parameters related to the descriptive quantities. In this paper, we are interested in the application of nonparametric regression techniques to exploring the relationship between the response variable and covariates, as well as prediction using auxiliary information. Bellhouse and Stafford (2001) extended a local polynomial regression technique to conduct flexible regression modelling for complex survey data. However, their paper dealt only with a simple nonparametric regression function. Here we extend their enquiry to a case of several independent variables, including indicator variables that often appear in regression analysis for survey data.

We consider a partially linear semi-parametric regression function defined as $E(\mathbf{y} \mid \mathbf{X}, \mathbf{z}) = \mathbf{X}\boldsymbol{\beta} + G(\mathbf{z})$, where $G(\cdot)$ is an arbitrary function and $\boldsymbol{\beta}$ is an unknown $p-$dimensional parameter vector. In this semi-parametric regression model, the explanatory variables are represented separately in two parts: a nonparametric part and a parametric linear part. It is of interest to estimate both the functional form of the nonparametric part of the model and the parameters that are included in the parametric part of the model. We put the categorical explanatory variables and continuous variables with assumed linear dependence in the parametric part of the model, $\mathbf{X}\boldsymbol{\beta}$, and a variable with little information on the functional form in the nonparametric part of the model, $G(\mathbf{z})$. This partial linear semi-parametric model not only has a priori motivation as a data analytic tool and retains an important interpretive feature, it also eases the high dimensional problem created by factors and some covariates by including them in the parametric part of the model.

A similar model has been developed for independently and identically distributed data independently by Robinson (1988) and Speckman (1988). In these papers, the estimation is conducted in three steps. In the first step, the means of the response variable and the parametric independent variables, conditional on the nonparametric variable, are treated as a function of that variable and smoothed; in the second step, the linear coefficients are estimated by regressing the residuals from the smoothed response variable on the residuals from the smoothed parametric covariates; finally, the difference between the response variable and its prediction from the regression model is smoothed in a similar manner to provide an estimate of the nonparametric part of the regression function. It has been shown in Robinson (1988) and Speckman (1988) that the resulting estimators are root-n consistent when the model is correct and the data points are independent and identically distributed. The objective of our paper is to apply this smoothing technique to survey data while allowing for a complex sampling scheme.

1. Zilin Wang, Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada, N2L 3C5. E-mail: zwang@wlu.ca; David R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON, Canada, N6A 5B7. E-mail: bellhouse@stats.uwo.ca.

We use the local polynomial regression estimation technique developed in Bellhouse and Stafford (2001) to conduct all the smoothing during the estimation process. A key element in accomplishing the local polynomial regression technique from Bellhouse and Stafford (2001) is binning, which follows the work of Bellhouse and Stafford (1999) in density estimation. In many survey data sets, a continuous variable may be naturally binned; for example age may be recorded as age last birthday. In general, bins correspond to the disjoint sets of values of a continuous covariate, and thus can be regarded as domains. At the level of the sample, we estimate the domain mean of the variable of interest by dividing the weighted sum of the variable within the domain by the sum of the weights within the domain. In Bellhouse and Stafford (2001), the response variable is binned according to the values of the covariate, and discretized, and the domain means of the response variable are smoothed to obtain the regression function. When the sample size is large and the number of bins is relatively small, then estimators based on binning are functions of domain estimators whose inferential properties can be readily derived from results in Shao (1996) and Serfling (1980). One of the practical advantages to binning is that it can reveal information on an obscured trend in a complex survey, which is sometimes quite important when the scale of the complex survey data set is large. There are, usually, multiple observations at each set of covariate values in these data sets.

An example that illustrates these features of binned data is taken from the Ontario Health Survey. The survey was conducted by Statistics Canada in 1990 with 61,239 individuals living in Ontario, Canada. The data were obtained by a stratified two-stage clustered design. The strata were the urban and rural areas covered by each of the public health units in the province of Ontario. Within each stratum enumeration areas were randomly selected, as were households within each enumeration area. The purpose of this survey is to measure the health status of the people of Ontario and to collect data relating to the risk factors of major causes of mortality and morbidity in Ontario. In this example, we examine people's weight as a function of age. In the Ontario Health Survey, age was given only to age last birthday. The measurement we use for a proxy of weight here is called body mass index (BMI) which is calculated as weight in kilograms divided by the square of height in meters. BMI is used as one of the indicator of a person's obesity level. Normally, a person with a BMI below 18 is considered underweight and a BMI greater than 30 suggests obesity. BMI is used as an appropriate measure only for all persons between the ages of 18 and 64 with the exception of pregnant and breast feeding women. Consequently the sample size is reduced to 44,457 eligible respondents that have 47 distinct possible ages or bins.

In the left panel of Figure 1, the age trend of body mass index is plotted. It is readily seen that the "black cloud"-like scatterplot masks the relationship between age and body mass index. Now, if we calculate mean of the body mass index at each distinct point of age, and plot the binned mean estimates of the body mass index versus age, we can obtain the plot in the right panel of Figure 1. It is obvious that a binned mean provides more visual information than the raw data does. Large-scale data sets not only can result in non-informative plots, they also make the estimation process computationally very cumbersome. Hence, it is natural in complex survey data analysis to bin the data into domains according to distinct values of a discretized covariate. Further, estimators from binning are functions of domain estimators.

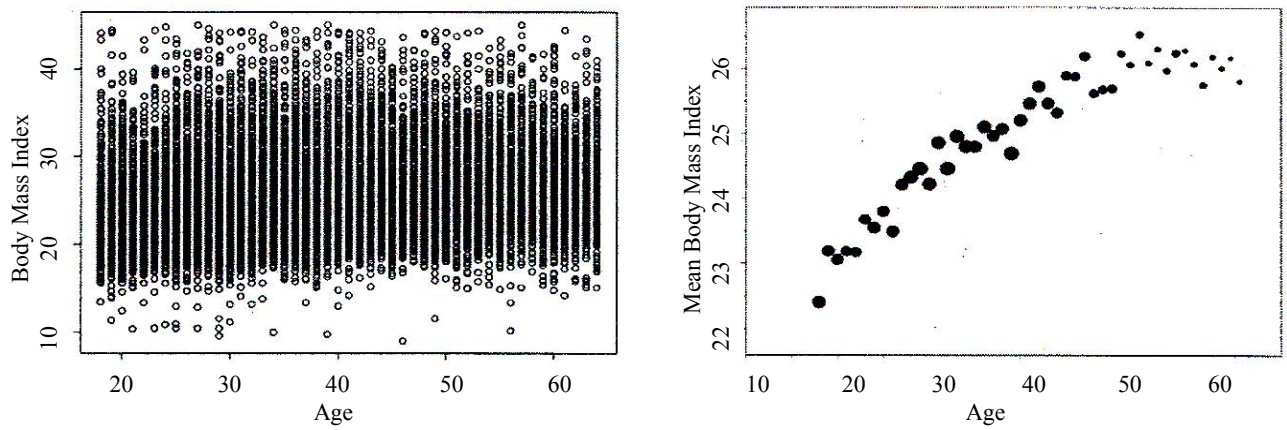

Figure 1 Comparison of the scatter plots of the binned and unbinned data from the Ontario health survey

One drawback to binning is that number of bins cannot grow asymptotically with the population if the data are naturally binned, as with the age variable in the above example. In such a case, the population level nonparametric estimators will remain biased as estimators of super-population functions due to a fixed bin size. In our framework, we assume that the bins induced by the distinct values of the covariate are the same in the population as in the sample; similarly in smoothing we will take the bandwidth to be the same at the population level as at the sampling level. We will show that the sample estimators are design consistent estimators of the corresponding finite population parameters and functions, though not of their superpopulation counterparts. In the Ontario Health Survey data example, the same set of distinct ages appears in both the population and the sample.

The paper is organized as follows. Superpopulation working models leading to the estimation procedures in survey data are introduced in Section 2. In Section 3, we derive all the moments of the estimates obtained and establish some asymptotic results. A simulation study and an empirical illustration of the estimation method carried out using the 1990 Ontario Health Survey (1992) appear in Section 4 and Section 5. Section 6 concludes with a discussion of assumptions made and some future work. The Proofs of all lemmas and theorems in Section 3 are given in an appendix.

2. Semiparametric regression model and its estimation

We take a typical approach to complex survey data analysis. First, we assume a working model on the finite population under the assumption of independent observations. Model parameter estimates then become the finite population parameters, or census parameters, to be estimated from the survey sample. Once the finite population target parameters have been defined we assume a more realistic model on the finite population in order to obtain inferences about these parameters. This is done in the next section. Consider a finite population of size N with a vector of measurements $(y_k, \mathbf{x}_k, z_k)$ attached to unit k, $k = 1, ..., N$, where y_k represents an observation of the response variable and $(\mathbf{x}_k, z_k)$ represents a vector of observations of the explanatory variables with length $p + 1$. As a working model we imagine that the response variable is generated by the following partial linear regression model,

$$\mathbf{Y} = G(\mathbf{z}) + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{Y}$ is the vector of responses and $\boldsymbol{\varepsilon}$ has entries that are independent and identically distributed with mean zero and constant variance. The function $G(\cdot)$ is an arbitrary function of $\mathbf{z}$ and $\boldsymbol{\beta}$ is an unknown p-dimensional parameter vector. The $N \times p$ matrix $\mathbf{X}$ corresponds to the linear part of the model and contains either continuous or discrete explanatory variables which are random. The term $G(\mathbf{z})$ is the nonparametric part of the model. We assume that z is non-stochastic and measured on a continuous scale, discretized into D distinct values. Additionally, it is imagined that $E(\boldsymbol{\varepsilon} \mid \mathbf{z}, \mathbf{X}) = \mathbf{0}$. There is no interaction between $\mathbf{X}$ and $\mathbf{z}$ in the model.

We are interested in estimating population level versions of $G(\cdot)$ and the parameters $\boldsymbol{\beta}$. We first develop expressions for these, guided by the estimation procedures in Robinson (1988) and Speckman (1988). In particular, we begin by taking the expectation of both sides of (1) conditional on $\mathbf{z}$:

$$E(\mathbf{Y} \mid \mathbf{z}) = E(\mathbf{X} \mid \mathbf{z})\boldsymbol{\beta} + G(\mathbf{z}). \tag{2}$$

Then we subtract (2) from (1) to obtain

$$\mathbf{Y} - E(\mathbf{Y} \mid \mathbf{z}) = \big(\mathbf{X} - E(\mathbf{X} \mid \mathbf{z})\big)\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

To define the population version of $\boldsymbol{\beta}$ in (3), we will replace $E(\mathbf{Y} \mid \mathbf{z})$ and $E(\mathbf{X} \mid \mathbf{z})$ in (3) by their population level estimates and estimate $\boldsymbol{\beta}$ by the method of least squares.

For the population level estimates of $E(\mathbf{Y} \mid \mathbf{z})$ and $E(\mathbf{X} \mid \mathbf{z})$, we adopt the local polynomial smoother in Jones (1989), in which binning is an essential part of the operation. Let the discretized Z variable take values $z_1, ..., z_D$; let the vectors of means in the bins of $z_1, ..., z_D$ be $\overline{\mathbf{Y}} = (\overline{Y}_1, ..., \overline{Y}_D)$ and $\overline{\mathbf{X}}_j = (\overline{X}_{j1}, ..., \overline{X}_{jD})$ for $j = 1, ..., p$, respectively. Also, let P_d be the population proportion of observations in the d^{th} bin for $d = 1, ..., D$. Then denote the population smoothed conditional expectations of $\mathbf{Y}$ and $\mathbf{X}_j$ at the point z_d by $m_{\mathbf{y}}(z_d)$ and $m_j(z_d)$, respectively. Given that $K(\cdot)$ is a kernel function satisfying $\int K(t)\,dt = 1$ and $\int K(t)^2\,dt < \infty$ and h is the bandwidth and using the principle of local polynomial regression technique, we minimize

$$\sum_{d'=1}^{D} \frac{P_d}{h} \{\overline{Y}_d - \alpha_0 - \alpha_1(z'_d - z_d), ..., -\alpha_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \tag{4}$$

and

$$\sum_{d'=1}^{D} \frac{P_d}{h} \{\overline{X}_{jd} - \gamma_0 - \gamma_1(z'_d - z_d), ..., -\gamma_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \tag{5}$$

with respect to α's and γ's so that the population estimated (smoothed) conditional expectations of y and X_j

on z_d, $m_{\mathbf{y}}(z_d)$ and $m_j(z_d)$, are the solutions of α_0 and γ_0 for equations (4) and (5). Specifically,

$$m_j(z_d) = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\bar{\mathbf{X}}_j$$

and

$$m_{\mathbf{y}}(z_d) = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\bar{\mathbf{Y}}$$

where q is the degree of the polynomial smoother, $\mathbf{e}$ is a $(q+1)\times 1$ vector in the form of $(1, 0, 0, ..., 0)^T$, and $\mathbf{Z}$ and $\mathbf{K}_{\mathbf{W}}$ are respectively defined as

$$\mathbf{Z} = \begin{pmatrix} 1 & z_1 - z_d & \dots & (z_1 - z_d)^q \\ 1 & z_2 - z_d & \dots & (z_2 - z_d)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_D - z_d & \dots & (z_D - z_d)^q \end{pmatrix} \tag{6}$$

and $\mathbf{K}_{\mathbf{W}} = \text{diag}(\hat{P}_1 K((z_1 - z_d)/h), ..., \hat{P}_D K((z_D - z_d)/h))/h$.

With the census estimators of the conditional expectations $m_j(z_d)$ and $m_{\mathbf{y}}(z_d)$, we define a $N \times p$ matrix $\mathbf{M}_{\mathbf{X}}$ and a $N \times 1$ vector $\mathbf{M}_{\mathbf{y}}$ as,

$$\mathbf{M}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} m_1(z_1) & m_2(z_1) & \cdots & m_p(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(z_1) & m_2(z_1) & \cdots & m_p(z_1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} m_1(z_D) & m_2(z_D) & \cdots & m_p(z_D) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(z_D) & m_2(z_D) & \cdots & m_p(z_D) \end{pmatrix} \end{pmatrix}$$

and

$$\mathbf{M}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} m_{\mathbf{y}}(z_1) \\ \vdots \\ m_{\mathbf{y}}(z_1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} m_{\mathbf{y}}(z_D) \\ \vdots \\ m_{\mathbf{y}}(z_D) \end{pmatrix} \end{pmatrix}. \tag{7}$$

Note that the d^{th} blocks of $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{y}}$ are of the dimensions of $N_d \times p$ and $N_d \times 1$, respectively, where N_d be number of observations that fall in the d^{th} bin and $\sum N_d = N$. Replacing the conditional expectation matrix, $E(\mathbf{X} \mid \mathbf{z})$, and vector, $E(\mathbf{Y} \mid \mathbf{z})$, in (3) with their estimates, $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{y}}$, and using the general estimating equations framework suggested by Godambe and Thompson (1986) for the least squares estimation, we can obtain the finite population versions parameters (census estimators) of $\boldsymbol{\beta}$, namely $\mathbf{B}$, by solving

$$\begin{aligned} \mathbf{u}(\boldsymbol{\theta}) &= \sum_{k=1}^{N} (\mathbf{x}_k - \mathbf{M}_{xk})^T (y_k - M_{yk}) \\ &\quad - \sum_{k=1}^{N} (\mathbf{x}_k - \mathbf{M}_{xk})^T (\mathbf{x}_k - \mathbf{M}_{xk})\mathbf{B} \\ &= \mathbf{0}_{p\times 1}, \end{aligned} \tag{8}$$

where $\mathbf{M}_{xk}$ is the k^{th} row of the $N \times p$ matrix $\mathbf{M}_{\mathbf{X}}$ and M_{yk} is the k^{th} element of the $N \times 1$ vector $\mathbf{M}_{\mathbf{y}}$. The finite population parameter vector $\boldsymbol{\theta}^T$ is composed of $(\mathbf{B}^T, \mathbf{m}_{\mathbf{x}}(\mathbf{z}), \mathbf{m}_{\mathbf{y}}(\mathbf{z})^T)$, where $\mathbf{m}_{\mathbf{x}}(\mathbf{z})$ is a vector of the form $(\mathbf{m}_1(\mathbf{z})^T, ..., \mathbf{m}_p(\mathbf{z})^T)$ with $\mathbf{m}_j(\mathbf{z}) = (m_j(z_1), ..., m_j(z_D))$ for $j = 1, ..., p$ and $\mathbf{m}_{\mathbf{y}}(\mathbf{z}) = (m_{\mathbf{y}}(z_1), ..., m_{\mathbf{y}}(z_D))$. Hence, the closed form expression for the estimator (census parameter) $\mathbf{B}$ is

$$\mathbf{B} = ((\mathbf{X} - \mathbf{M}_{\mathbf{X}})^T(\mathbf{X} - \mathbf{M}_{\mathbf{X}}))^{-1}(\mathbf{X} - \mathbf{M}_{\mathbf{X}})^T(\mathbf{Y} - \mathbf{M}_{\mathbf{y}}).$$

Once $\mathbf{B}$ is obtained, the difference between the response variable $\mathbf{Y}$ and the product $\mathbf{XB}$ is treated as the dependent random variable and the function $G(\cdot)$ is estimated in accordance with the following model

$$\mathbf{Y} - \mathbf{XB} = G(\mathbf{z}) + \boldsymbol{\varepsilon}.$$

The finite population version of $G(\mathbf{z})$ at z_d, namely $g(z_d)$, is

$$g(z_d) = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}(\bar{\mathbf{Y}} - \bar{\mathbf{X}}\mathbf{B}),$$

where $\bar{\mathbf{X}}$ is a $D \times p$ matrix of the form $(\bar{\mathbf{X}}_1, ..., \bar{\mathbf{X}}_p)$.

Realistically, we cannot access the whole population. Instead, we can only observe a sample drawn from the population using a certain probability sampling design. Let $\mathbf{s}$ be the set of n sample units with sample $(y_k, \mathbf{x}_k, z_k, w_k)$ for $k \in \mathbf{s}$, where w_k is the sampling weight for unit k. Additionally, we assume that there is complete response so that the inclusion probability is equal to the reciprocal of the sampling weight. We assume further that the bins induced by the distinct values of $\mathbf{z}$ are preserved from the population to the sample. This is appropriate in a variable such as age recorded to age last birthday.

Using the local polynomial regression technique for complex survey data in Bellhouse and Stafford (2001), we use the sampling versions of the objective functions in (4) and (5) as follows,

$$\sum_{d'=1}^{D} \frac{\hat{p}_d}{h}\{\bar{y}_d - \alpha_0 - \alpha_1(z'_d - z_d), ..., -\alpha_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right) \tag{9}$$

and

$$\sum_{d'=1}^{D} \frac{\hat{p}_d}{h} \{\overline{x}_{jd} - \gamma_0 - \gamma_1(z'_d - z_d), \ldots, -\gamma_q(z'_d - z_d)^q\}^2 \times K\left(\frac{z'_d - z_d}{h}\right), \quad (10)$$

where $\overline{\mathbf{y}}$ and $\overline{\mathbf{x}}_j$ are sample estimators of $\overline{\mathbf{Y}}$ and $\overline{\mathbf{X}}_j$ and are of the forms $(\overline{y}_1, \ldots, \overline{y}_D)^T$ and $(\overline{x}_{j1}, \ldots, \overline{x}_{jD})^T$, respectively, and $\hat{p}_d$ is the weighted sample proportion of observations in bin d. Consequently, we have the survey estimator of $m_{\mathbf{y}}(z)$ and $m_j(z)$ at z_d, given by

$$\hat{m}_j(z_d) = \mathbf{e}^T(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\overline{\mathbf{x}}_j$$

and (11)

$$\hat{m}_{\mathbf{y}}(z_d) = \mathbf{e}^T(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\overline{\mathbf{y}},$$

where $\mathbf{Z}$ has the same form as in (6) and $\hat{\mathbf{K}}_{\mathbf{W}}$ is defined as

$$\hat{\mathbf{K}}_{\mathbf{W}} = \frac{1}{h}\text{diag}(\hat{p}_1 K((z_1 - z_d)/h), \ldots, \hat{p}_D K((z_D - z_d)/h)).$$

We can also construct the $n \times p$ matrix $\hat{\mathbf{M}}_{\mathbf{X}}$ and $n \times 1$ vector $\hat{\mathbf{M}}_{\mathbf{y}}$ using the same method that we used to construct $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{y}}$ in equations (7). That is, we use sampling estimators $\hat{m}_j(z_d)$ and $\hat{m}_{\mathbf{y}}(z_d)$ that are shown in (11) to obtain

$$\hat{\mathbf{M}}_{\mathbf{X}} = \begin{pmatrix} \begin{pmatrix} \hat{m}_{\mathbf{x}_1}(z_1) & \hat{m}_{\mathbf{x}_2}(z_1) & \cdots & \hat{m}_{\mathbf{x}_p}(z_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{m}_{\mathbf{x}_1}(z_1) & \hat{m}_{\mathbf{x}_2}(z_1) & \cdots & \hat{m}_{\mathbf{x}_p}(z_1) \end{pmatrix} \\ \vdots \quad\quad\quad \vdots \\ \begin{pmatrix} \hat{m}_{\mathbf{x}_1}(z_D) & \hat{m}_{\mathbf{x}_2}(z_D) & \cdots & \hat{m}_{\mathbf{x}_p}(z_D) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{m}_{\mathbf{x}_1}(z_D) & \hat{m}_{\mathbf{x}_2}(z_D) & \cdots & \hat{m}_{\mathbf{x}_p}(z_D) \end{pmatrix} \end{pmatrix}$$

and

$$\hat{\mathbf{M}}_{\mathbf{y}} = \begin{pmatrix} \begin{pmatrix} \hat{m}_{\mathbf{y}}(z_1) \\ \vdots \\ \hat{m}_{\mathbf{y}}(z_1) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \hat{m}_{\mathbf{y}}(z_D) \\ \vdots \\ \hat{m}_{\mathbf{y}}(z_D) \end{pmatrix} \end{pmatrix}.$$

Let n_d be the number of observations in the d^{th} bin such that $\sum n_d = n$. Similar to $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{y}}$ in (7), the d^{th} blocks of $\hat{\mathbf{M}}_{\mathbf{X}}$ and $\hat{\mathbf{M}}_{\mathbf{y}}$ are of the dimensions of $n_d \times p$ and $n_d \times 1$, respectively.

Analogous to the population estimating equation (8), the sampling estimating equation for $\mathbf{B}$ is

$$\begin{aligned} \hat{\mathbf{u}}(\hat{\boldsymbol{\theta}}) &= \sum_{k \in \mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{xk})^T (y_k - \hat{M}_{xk}) w_k \\ &\quad - \sum_{k \in \mathbf{s}} (\mathbf{x}_k - \hat{\mathbf{M}}_{xk})^T (\mathbf{x}_k - \hat{\mathbf{M}}_{xk}) \hat{\mathbf{B}} w_k \\ &= \mathbf{0}, \end{aligned} \quad (12)$$

where $\hat{\boldsymbol{\theta}}^T = (\hat{\mathbf{B}}^T, \hat{\mathbf{m}}_{\mathbf{x}}(\mathbf{z}), \hat{\mathbf{m}}_{\mathbf{y}}(\mathbf{z})^T)$ is the sampling estimator of $\boldsymbol{\theta}^T = (\mathbf{B}^T, \mathbf{m}_{\mathbf{x}}(\mathbf{z}), \mathbf{m}_{\mathbf{y}}(\mathbf{z})^T)$. Note that a similar approach was considered by Fuller (1975) and Binder (1983). Nevertheless, the solution to (12) provides the closed form of $\hat{\mathbf{B}}$ as

$$\hat{\mathbf{B}} = ((\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{X}})^T \mathbf{W}_n (\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{X}}))^{-1} (\mathbf{x} - \hat{\mathbf{M}}_{\mathbf{X}})^T \mathbf{W}_n (\mathbf{y} - \hat{\mathbf{M}}_{\mathbf{y}}),$$

where $\mathbf{W}_n$ is an $n \times n$ weight matrix with design weights w_k on the diagonal entry for $k \in \mathbf{s}$, $\mathbf{y}$ is an $n \times 1$ vector containing the sample observations of the response variable and $\mathbf{x}$ is an $n \times p$ matrix consisting of the sample observations of the covariates.

Using the sample estimates of $\mathbf{B}$ and denoting $\overline{\mathbf{x}}$ as a $D \times p$ matrix of the form $(\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_p)$, we can obtain the sampling estimate of $g(z_d)$ as

$$\hat{g}(z_d) = \mathbf{e}^T(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}(\overline{\mathbf{y}} - \overline{\mathbf{x}}\hat{\mathbf{B}}).$$

Again, if q and h are the same as for $\hat{m}_j(z_d)$, the expression for $\hat{g}(z_d)$ simplifies.

When applying local polynomial regression techniques to obtain the estimators of conditional expectations as well as the arbitrary function $G(\cdot)$, we need to choose an appropriate bandwidth h. Because binning is involved in all aspects of the estimation process and since we assume that bins induced by the distinct values of $\mathbf{z}$ are preserved from the population to the sample, we argue that the same bandwidth should be used for obtaining both the census estimators and the sample estimators. Since we do not have all the observations of the finite population, we use the sample to choose the appropriate band width. In this paper, we adopt the method in Fan and Gijbels (1995), where the authors developed a data-driven bandwidth selector that combines the ideas of the plug-in and the cross-validation methods for the identically and independently distributed data. When applying this data-driven method to our case, criteria, such as the residual sum of squares and mean square error, of the resulting estimates of the conditional expectations are needed. By noting that those criteria depend on the estimated conditional expectations or regression functions and the derivatives of the regression functions, we

can use the objective functions defined in (9) and (10) to obtain not only the survey estimates of regression functions, but also the derivatives of the regression functions. For more details, see Wang (2004).

3. Design properties of sampling estimators

3.1 Notation and assumptions

In showing design properties of the estimators, we follow Särndal, Swensson and Wretman (1992) and Isaki and Fuller (1982) in considering a nested sequence of populations U_ν, for $\nu = 1, 2, ...$, such that $U_1 \subset U_2 \subset U_3 \subset \ldots$. All population quantities, sample sizes and values, and survey estimators are indexed by ν. However, for ease of notation we drop ν as a subscript for these quantities. We denote the expectation and variance with respect to sampling design as E_p and Var_p, respectively, and in accordance with the above nested populations, we define design-based consistency and asymptotic unbiasedness as in Thompson (1997, page 167).

In what follows, the development of the asymptotic results for the estimators will depend on the asymptotic normality and consistency of the estimates of means and totals. We will not restrict ourselves to specific sampling designs; instead, we assume that all the survey totals that appear in the estimators are of the Horvitz-Thompson type. Hence, the consistency and asymptotic normality of estimators are subject to the standard regularity conditions on the sampling designs for the consistency and normality of Horvitz-Thompson type estimators, which have been studied by Madow (1948), Hàjek (1960), Bickel and Freedman (1983), Krewski and Rao (1981) and Shao (1996). The aforementioned literature shares some restrictions on the sampling design. An implication of these restrictions is that no survey weight is disproportionately large, the total number of first stage sampled clusters or primary sampling units is increasing, but with a growing gap between sample and population. In addition, a Liapunov - type condition ensures that the variables z, $\boldsymbol{x}$ and y develop in a regular manner as ν tends to infinity.

We will use the result that any vector of estimators of totals from binned data is asymptotically multivariate normal, provided that the conditions in the previous paragraph are met and the number of domains is fixed. This is obtained through application of results in Shao (1996, page 211) and Serfling (1980, page 18). Shao (1996) shows that in this framework any smooth function of estimates of totals is asymptotically normal. An estimate of a domain mean is one such smooth function. Likewise, any linear combination of different domain mean estimates is a smooth function of survey estimates of totals. For our purposes, the bins form the domains and hence any vector of bin means is asymptotically multivariate normal. The asymptotic result used here depends on having a fixed number of bins. However, it can be incorporated in principle into a theory of the superpopulation parameters, as for example in the approach of Buskirk and Lohr (2005).

Define $\hat{\mathbf{m}}_\xi(\mathbf{z}) = (\hat{\mathbf{m}}_\mathbf{x}(\mathbf{z}), \hat{\mathbf{m}}_\mathbf{y}(\mathbf{z})^T)^T$ as the survey estimator of $\mathbf{m}_\xi(\mathbf{z}) = (\mathbf{m}_\mathbf{x}(\mathbf{z}), \mathbf{m}_\mathbf{y}(\mathbf{z})^T)^T$. Using a Taylor linearization technique on (12) and letting ε denote a quantity approaching 0 and as $\hat{\boldsymbol{\theta}}$ approached to $\boldsymbol{\theta}$, we have

$$-\hat{\mathbf{u}}_B(\boldsymbol{\theta})(\hat{\mathbf{B}} - \mathbf{B}) \doteq \hat{\mathbf{u}}(\boldsymbol{\theta}) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})(\hat{\mathbf{m}}_\xi(\mathbf{z}) - \mathbf{m}_\xi(\mathbf{z})) + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|\,\varepsilon, \qquad (13)$$

where $\hat{\mathbf{u}}(\boldsymbol{\theta})$ is a linear sampling estimator of $\mathbf{u}(\boldsymbol{\theta})$ in (8) and is of the form

$$\hat{\mathbf{u}}(\boldsymbol{\theta}) = \sum_{k\in\mathbf{s}} (\mathbf{x}_k - \mathbf{M}_{xk})^T (y_k - M_{yk})\, w_k - \sum_{k\in\mathbf{s}} (\mathbf{x}_k - \mathbf{M}_{xk})^T (\mathbf{x}_k - \mathbf{M}_{xk})\mathbf{B} w_k; \qquad (14)$$

$\hat{\mathbf{u}}_\mathbf{B}(\boldsymbol{\theta})$ is the gradient of $\hat{\mathbf{B}}$ obtained from $\hat{\mathbf{u}}(\boldsymbol{\theta})$; and $\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ is a $p \times (p+1)D$ matrix whose components are the first derivatives of $\hat{\mathbf{u}}(\boldsymbol{\theta})$ with respect to $\mathbf{m}_\xi(\mathbf{z})$. Denote by $\mathbf{u}_B(\boldsymbol{\theta})$ and $\mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ the population parameters corresponding to $\hat{\mathbf{u}}_\mathbf{B}(\boldsymbol{\theta})$ and $\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$, respectively.

In addition to the aforementioned regularity conditions, we impose the following conditions, letting $\mathcal{N}$ denote a neighbourhood of the true value of the parameters of interest.

C1. $\lim_{\nu\to\infty} \mathbf{u}(\boldsymbol{\theta})/N$ exists and is finite for all $\boldsymbol{\theta}$ and $\mathcal{N}$.

C2. $\lim_{\nu\to\infty} \mathbf{u}_\mathbf{B}(\boldsymbol{\theta})/N = \mathbf{H}_\mathbf{B}$ and $\mathbf{H}_\mathbf{B}$ is of full rank and is invertible for all $\boldsymbol{\theta}$ and $\mathcal{N}$.

C3. $\lim_{\nu\to\infty} \mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})/N = \mathbf{H}_\xi(\boldsymbol{\theta})$ and $\mathbf{H}_\xi(\boldsymbol{\theta})$ has a finite determinant for all $\boldsymbol{\theta}$ and $\mathcal{N}$.

C4. $\lim_{\nu\to\infty} n\,\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta})/N) = \mathrm{V}(\hat{\mathbf{u}}(\boldsymbol{\theta}))$ where Var_p is the design-based variance and $V(\hat{\mathbf{u}}(\boldsymbol{\theta}))$ is a positive-definite variance matrix for all $\boldsymbol{\theta}$ and $\mathcal{N}$.

C5. $\lim_{\nu\to\infty} N_d/N = \omega_d$ and $\lim_{\nu\to\infty} n/N = f$ with both ω_d and f are constants between 0 and 1.

C6. Let $\mathbf{A}_d = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_\mathbf{W}$ be the population smoothing matrix; then $\lim_{\nu\to\infty} \mathbf{A}_d$ exists and is finite for $d = 1, ..., D$.

C7. $\lim_{\nu\to\infty} n\,\mathrm{Var}_p(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})) = \mathrm{V}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}))$.

C8. Matrices of population values $\mathbf{Z}^T\mathbf{K}_\mathbf{W}\mathbf{Z}$ and $\mathbf{u}_\mathbf{B}(\boldsymbol{\theta})$ are invertible, as well as their sampling estimators $\mathbf{Z}^T\hat{\mathbf{K}}_\mathbf{W}\mathbf{Z}$ and $\hat{\mathbf{u}}_\mathbf{B}(\hat{\boldsymbol{\theta}})$.

3.2 Asymptotic properties of $\hat{\mathbf{B}}$

The proofs of all lemmas and theorems in this and the following section may be found in the Appendix. From the Taylor linearization results in (13), we know that the properties of $\hat{\mathbf{B}}$ are dependent on those of $\hat{\mathbf{u}}(\boldsymbol{\theta})$, $\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})$, $\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ and $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$; their properties are stated in the following two Lemmas.

Lemma 1. *If conditions* $C1 - C4$ *are satisfied*, *we have as* $\nu \to \infty$:

1) $\sqrt{n}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta}))/N \longrightarrow N(\mathbf{0}, \mathrm{V}(\hat{\mathbf{u}}(\boldsymbol{\theta})))$;
2) $|\hat{\mathbf{u}}_B(\boldsymbol{\theta}) - \mathbf{u}_B(\boldsymbol{\theta})|/N$ *and* $|\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta}) - \mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})|$ *converge to* $\mathbf{0}$ *in probability for* $\boldsymbol{\theta}$ *and* $\mathcal{N}$;
3) $|\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta})|/N$ *converges to zero in probability.*

Lemma 2. *Under conditions* $C5$ *to* $C7$, $\sqrt{n}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})) = O_p(1)$.

Building on Lemmas 1 and 2, we have the asymptotic normality of $\hat{\mathbf{B}}$ in Theorem 1.

Theorem 1. *Under conditions* $C1$ *to* $C7$, *assuming the parameter space contains a neighbourhood of the parameter of interest*, *we have as* ν *goes to infinity*:

1) $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) \xrightarrow{d} N(\mathbf{0}, \mathrm{V}(\hat{\mathbf{B}}))$ *where* $\mathrm{V}(\hat{\mathbf{B}}) = \lim_{\nu \to \infty} n \mathrm{Var}_p(\hat{\mathbf{B}})$;
2) $|\hat{\mathbf{B}} - \mathbf{B}|$ *converges to zero in probability.*

To obtain approximate moments for $\hat{\mathbf{B}}$, we take expectations on both sides of equation (13), which yields

$$E_p(-\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})(\hat{\mathbf{B}} - \mathbf{B})) \doteq E_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) + E_p\{\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})[\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})]\} + E_p(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||)\varepsilon. \quad (15)$$

The assumption that the second moments of the estimates are bounded makes the last term of equation (15) vanish in the limit. Following along the lines of Binder (1983), we have

$$E_p(-\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta}))E_p((\hat{\mathbf{B}} - \mathbf{B})) \doteq E_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) + E_p(\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta}))E_p\{[\hat{\mathbf{m}}_{\xi}(\mathbf{z}) - \mathbf{m}_{\xi}(\mathbf{z})]\}.$$

The survey totals that define the vector $\hat{\mathbf{u}}(\boldsymbol{\theta})$ and matrix $\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})$ are Horvitz-Thompson-type estimators and they are unbiased (Thompson 1997). Hence, $E_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) = \mathbf{u}(\boldsymbol{\theta})$ and $E_p(\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})) = \mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})$. Since $\mathbf{u}(\boldsymbol{\theta})$ is the estimating equation for the partial linear coefficients defined in (8), it is equal to a $1 \times p$ zero vector. Further, it has been shown in Bellhouse and Stafford (2001) that $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ is an asymptotically unbiased estimator of $\mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})$. Hence, $-\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})E_p((\hat{\mathbf{B}} - \mathbf{B})) \doteq \mathbf{0}$, or, based on the conditions that $\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})$ is invertible and $\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})^{-1}$ is finite, we have $E_p(\hat{\mathbf{B}}) \doteq \mathbf{B}$.

Taking the variance of both sides of equation (13) and using the approximated variance-covariance matrices of $\hat{\mathbf{u}}(\boldsymbol{\theta})$ and $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$, we obtain the asymptotic variance of $\hat{\mathbf{B}}$ as

$$\mathrm{Var}_p(\hat{\mathbf{B}}) \doteq \mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})^{-1} (\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta})) + \mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})(\mathbf{A}(\mathbf{J} \otimes \mathrm{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}))\mathbf{A}^T)\mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})^T + 2(I_p \otimes \boldsymbol{\ell})(\boldsymbol{\ell} \otimes \mathbf{C})\mathbf{A}^T\mathbf{U}_{\boldsymbol{\xi}}(\boldsymbol{\theta})^T)(\mathbf{u}_{\mathbf{B}}(\boldsymbol{\theta})^T)^{-1}, \quad (16)$$

where $\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta}))$ is a $p \times p$ matrix composed of variances of totals in the vector $\hat{\mathbf{u}}(\boldsymbol{\theta})$ and $\mathrm{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is the variance-covariance matrix of the binned means of the parametric covariates and the response variable. The matrices $\mathbf{J}$ and $\boldsymbol{\ell}$ are the $D \times D$ unit matrix and the $1 \times D$ unit vector, respectively. Finally, we have

$$\mathbf{A} = \begin{pmatrix} I_{p+1} \otimes \mathbf{A}_1 & 0 & 0 & 0 \\ 0 & I_{p+1} \otimes \mathbf{A}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & I_{p+1} \otimes \mathbf{A}_D \end{pmatrix}$$

and

$$\mathbf{C} = \begin{pmatrix} \mathrm{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{x}}_1) & \cdots & \mathrm{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{x}}_p) & \mathrm{Cov}_p(\hat{\mathbf{t}}_1, \bar{\mathbf{y}}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{x}}_1) & \cdots & \mathrm{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{x}}_p) & \mathrm{Cov}_p(\hat{\mathbf{t}}_p, \bar{\mathbf{y}}) \end{pmatrix},$$

where, for $j = 1, ..., p$, $\hat{\mathbf{t}}_j$ is a $D \times 1$ vector whose d^{th} entry is $\sum_{k \in \mathbf{s}_d} w_{jk} u_{jk}(\boldsymbol{\theta})$ and $\mathbf{A}_d = \mathbf{e}^T(\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}_{\mathbf{W}}$ for $d = 1, ..., D$.

Replacing $\boldsymbol{\theta}$, $\mathrm{Var}_p(\hat{\mathbf{u}}(\boldsymbol{\theta}))$, $\mathrm{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, $\mathbf{A}$ and $\mathbf{C}$ by their sample estimators, we have the survey estimator of the variance of $\hat{\mathbf{B}}$:

$$\widehat{\mathrm{Var}}_p(\hat{\mathbf{B}}) = \hat{\mathbf{u}}_{\mathbf{B}}^{-1}(\hat{\boldsymbol{\theta}})(\widehat{\mathrm{Var}}_p(\hat{\mathbf{u}}(\hat{\boldsymbol{\theta}})) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}})(\hat{\mathbf{A}}(\mathbf{J} \otimes \widehat{\mathrm{Cov}}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}}))\hat{\mathbf{A}}^T)\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}})^T + 2(I_p \otimes \boldsymbol{\ell})(\boldsymbol{\ell} \otimes \hat{\mathbf{C}})\hat{\mathbf{A}}^T\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\hat{\boldsymbol{\theta}})^T)(\hat{\mathbf{u}}_{\mathbf{B}}^T(\hat{\boldsymbol{\theta}})^T)^{-1},$$

where $\hat{\mathbf{A}}$ is the survey estimator of $\mathbf{A}$ and is composed of $\hat{\mathbf{A}}_d = \mathbf{e}^T(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}$.

3.3 Asymptotic properties of $\hat{g}(\cdot)$

Define $\bar{\mathbf{r}} = \bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}$ to be the sample estimator of $\bar{\mathbf{R}} = \bar{\mathbf{Y}} - \bar{\mathbf{X}}\mathbf{B}$. A linearization around the population parameters, as well as design unbiasedness of domain means and $\hat{\mathbf{B}}$, leads to the asymptotic design unbiasedness of $\bar{\mathbf{r}}$. Reexpressing $\hat{g}(z_d)$, we have $\hat{g}(z_d) = \hat{\mathbf{A}}_d\bar{\mathbf{r}}$. In $\hat{\mathbf{A}}_d$, we can expand $(\mathbf{Z}^T\hat{\mathbf{K}}_{\mathbf{W}}\mathbf{Z})^{-1}$ using the Taylor series expansion that $(\mathbf{I} + \mathbf{G})^{-1} = \mathbf{I} - \mathbf{G} + \mathbf{G}^2 - \cdots$ given that $\mathbf{G}$ is a symmetric and invertible matrix. Using the first two terms of the expansion, we can show that $E_p(\hat{\mathbf{A}}_d)$ is approximately $\mathbf{A}_d$. Hence, we have the asymptotic design

unbiasedness of $\hat{g}(z_d)$. With the same technique, the approximate asymptotic design-based variance of $\hat{g}(z_d)$ is obtained as

$$\mathrm{Var}_p(\hat{g}(z_d)) = \mathbf{A}_d \mathrm{Var}_p(\bar{\mathbf{r}})\mathbf{A}_d^T,$$

where, given that $\mathbf{Q} = (1, -B_1, ..., -B_p)$,

$$\begin{aligned}\mathrm{Var}_p(\bar{\mathbf{r}}) \doteq\ & (\mathbf{Q}\otimes\mathbf{I}_D)\mathrm{Cov}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})(\mathbf{Q}\otimes\mathbf{I}_D)^T \\ & + \bar{\mathbf{X}}\mathrm{Var}_p(\hat{\mathbf{B}})\bar{\mathbf{X}}^T \\ & - 2(\mathbf{Q}\otimes\mathbf{I}_D)\mathrm{Cov}_p(\bar{\mathbf{y}}, \hat{\mathbf{B}})\bar{\mathbf{X}}^T \\ & - \sum_{j=1}^{p} 2(\mathbf{Q}\otimes\mathbf{I}_D)\mathrm{Cov}_p(\bar{\mathbf{x}}_j, \hat{\mathbf{B}})\bar{\mathbf{X}}^T.\end{aligned}$$

Given the estimated variance of $\bar{\mathbf{r}}$, namely

$$\begin{aligned}\widehat{\mathrm{Var}}_p(\bar{\mathbf{r}}) =\ & (\hat{\mathbf{Q}}\otimes\mathbf{I}_D)\widehat{\mathrm{Cov}}_p(\bar{\mathbf{x}}, \bar{\mathbf{y}})(\hat{\mathbf{Q}}\otimes\mathbf{I}_D)^T \\ & + \bar{\mathbf{x}}\widehat{\mathrm{Var}}_p(\hat{\mathbf{B}})\bar{\mathbf{x}}^T \\ & - 2(\mathbf{Q}\otimes\mathbf{I}_D)\widehat{\mathrm{Cov}}_p(\bar{\mathbf{y}}, \hat{\mathbf{B}})\bar{\mathbf{x}}^T \\ & - \sum_{j=1}^{p} 2(\hat{\mathbf{Q}}\otimes\mathbf{I}_D)\widehat{\mathrm{Cov}}_p(\bar{\mathbf{x}}_j, \hat{\mathbf{B}})\bar{\mathbf{x}}^T,\end{aligned}$$

the estimated variance of $\hat{g}(z_d)$ is $\widehat{\mathrm{Var}}_p(\hat{g}(z_d)) = \hat{\mathbf{A}}_d \widehat{\mathrm{Var}}_p(\bar{\mathbf{r}})\hat{\mathbf{A}}_d^T$.

The asymptotic normality of $\hat{g}(\cdot)$ is also dependent on the normality of $\bar{\mathbf{r}}$, which is shown in the following Lemma.

Lemma 3. Under conditions **C**1 *to* **C**7 *and assuming that the dimension of* $\bar{\mathbf{r}}$ *is finite, we have as* ν *goes to infinity*

$$\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) \xrightarrow{d} N(\mathbf{0}, \mathrm{V}(\bar{\mathbf{r}})),$$

where

$$\mathrm{V}(\bar{\mathbf{r}}) = \lim_{\nu\to\infty} n\,\mathrm{Var}_p(\bar{\mathbf{r}}).$$

Based on the asymptotic normality developed in Lemma 3 and the estimator of the variance of $\hat{g}(z_d)$, we establish the asymptotic properties of $\hat{g}(z_d)$ in the following Theorem.

Theorem 2. Under conditions **C**1 *to* **C**7, *we have as* ν *goes to infinity*:

1) $|\hat{g}(z_d) - g(z_d)| \xrightarrow{P} 0$;

2) $(\hat{g}(z_d) - g(z_d))/\sqrt{\widehat{\mathrm{Var}}_p(\hat{g}(z_d))} \xrightarrow{d} N(0, 1)$.

4. Simulation studies

4.1 Design of experiment

The simulation study implemented here was designed to illustrate the theoretical results in Theorems 1 and 2. We generated the data in a two-step process that mimicked a superpopulation approach to sampling. First, we generated the finite population and then the sample was selected from it. In particular, we considered a finite population of $L = 500$ clusters with $M(= M_i) = 20{,}000$ in each. The population observations for the measurement of interest y_{ij} were obtained from the model

$$\begin{aligned}y_{ij} =\ & \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \\ & + 0.5\exp\left(\frac{z_{ij} - 40}{10}\right) + \mu_i + \varepsilon_{ij} \qquad (17)\end{aligned}$$

for $i = 1, ..., L$ and $j = 1, ..., M$ where the error terms μ_i and ε_{ij} are mutually independent with $\mu_i \sim N(0, \sigma_\mu^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. We set $\sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$ so that the intracluster correlation coefficient is $\rho = \sigma_\mu^2/\sigma^2$. Among the covariates in the model, both x_{1ij} and x_{2ij} were treated as the parametric linear part of the model and z_{ij} as the nonparametric part. We generated the x_{1ij} from the Bernoulli(1/2) distribution and the x_{2ij} from the Uniform(0, 1) distribution. The z_{ij} were generated from the age distribution of the Canadian population (according to the 1996 census) for the 18 to 64 age range and were independent of the error terms. Results for the values $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 3$, $\sigma^2 = 3$ and $\rho = 0, 0.2, 0.5$ are reported in this study. A two-stage sampling design, with $l(= 10, 25, 50, 100)$ clusters chosen at random from L and $m(= 1{,}000)$ secondary sampling units chosen at random from each cluster of size M, was used for the study. For each sample size and value of ρ, the simulation was repeated 300 times. At the population level, we applied the bandwidth selection method from Fan and Gijbels (1995) and determined that the bandwidths for estimating the conditional expectations of X_1 and X_2 on z were 1.2 and 1.5 respectively. When smoothing the residuals to estimate $g(z)$, the bandwidth was 0.6.

4.2 Results

Using the generated finite population, we found that the census estimates were $B_1 = 2.01$ and $B_2 = 3.00$. To check the design unbiasedness and efficiency of $\hat{\mathbf{B}}$, we calculated the simulated squared bias (Bias^2), which is the square of the difference between the average of the simulated estimates and the census estimates. In addition, the ratio of the average variance estimates to the simulated variance of each estimator of a linear coefficients (RVar) is presented to show the validity of the variance estimator $\mathrm{Var}_p(\hat{\mathbf{B}})$. To

evaluate the normality of $\hat{\mathbf{B}}$, we standardized the estimates of linear coefficients using the empirical standard deviation and population value of $\mathbf{B}$ and graphed the quantile - quantile plots of the standardized values.

Applying the semiparametric technique in Speckman (1988) to the model (17), we obtained census estimates $g(z)$ for $z = 18, ..., 64$. To evaluate the design accuracy of $\hat{g}(z)$, we took the difference between $\hat{g}(z)$ and $g(z)$ at each distinct point. The average of the squares of the differences over 47 distinct values of z is then reported as ABias^2. Two mean square errors were computed to check the design efficiency of $\hat{g}(z)$ and convergence of $\widehat{\text{Var}}_p(\hat{g}(z))$. One of the mean square errors is the average of the estimates of the integrated mean square error (AIMSE), which is obtained by first summing the $\widehat{\text{Var}}_p(\hat{g}(z))$ over $z = 18, ..., 64$ for each simulation and then taking the average of the sums over the total number of simulations. The simulated integrated mean square error (IMSE) is another mean square error and was computed by summing up the simulated mean square error at each distinct point of z. The average of the ratios of the simulated mean of $\widehat{\text{Var}}_p(\hat{g}(z))$ to the simulated variance of $\hat{g}(z)$ (Reff) shows the convergence of $\widehat{\text{Var}}_p(\hat{g}(z))$. In addition, we computed the coverage of the pointwise 95% confidence interval at each distinct point of z.

The results on the properties of $\hat{\mathbf{B}}$, $\widehat{\text{Var}}_p(\hat{\mathbf{B}})$, $\hat{g}(z)$ and $\widehat{\text{Var}}_p(\hat{g}(z))$ are found in Tables 1 and 2 and Figures 2 and 3. Tables 1 and 2 show information about accuracy and precision of the simulated estimates of $\hat{\mathbf{B}}$ and $\hat{g}(\cdot)$. Figure 2 gives the quantile-quantile plots of the sample standardized value of $\hat{B}_2$. Note that the quantile-quantile plots for $\hat{B}_1$ behave in a similar way to those for $\hat{B}_2$. Figure 3 graphs the coverage of the 95% confidence intervals for $g(\cdot)$. In Figures 2 and 3, we only report the cases where $l = 10, 25, 100$ and $\rho = 0, 0.5$. The overall performance of the estimators agrees with the theory in Theorems 1 and 2.

Table 1 confirms the design unbiasedness of $\hat{\mathbf{B}}$. It also shows that as the sample size increases, the performances of the estimates of the linear coefficients improve for all the error structures. In particular, the squared bias and variance of $\hat{\mathbf{B}}$ decreases as the number of primary samples increases. The estimated variance of $\hat{\mathbf{B}}$ gets closer to the simulated variance of $\hat{\mathbf{B}}$ as the sample size increases; this confirms the consistency of the variance estimates of $\hat{\mathbf{B}}$. Comparing the variances and biases of $\hat{\mathbf{B}}$ in the cases that $\rho = 0.2$ and $\rho = 0.5$ to the case where $\rho = 0$, we found that the intracluster correlation (cluster effects) did not affect the performance of $\hat{\mathbf{B}}$. This may be because the within cluster sample size was large.

Observing Figure 2, we find that both the number of primaries sampled and the cluster effect play some role in the normality of $\hat{\mathbf{B}}$. In particular, when the primary sample size is low, for instance $l = 10$, normality of the standardized $\hat{\mathbf{B}}$ shows some deviation from the theory for both $\rho = 0$ and $\rho = 0.5$. When l increases to 25, we find that performance of $\hat{\mathbf{B}}$ for $\rho = 0$ starts to recover whereas, for $\rho = 0.5$, there is no improvement until $l = 100$. Empirically, this finding suggests that when the number of clusters is low, we should not rely on the theoretical normality of the estimates of the coefficient; instead, we may want to use t distribution to carry out the inference.

As for the results of the nonparametric part of the estimation, Table 2 shows that the average estimated integrated mean square errors are very close to the simulated integrated mean square errors for all the sample sizes and error structures. Design unbiasedness is again confirmed with the average squared bias (ABias^2). The values of average ratio of the estimated variance to the simulated variance (RVar), which are close to 1 for all cases, are in line with the design consistency of the estimator of the variance of $\hat{g}(z)$. The integrated mean square errors of $\hat{g}(\cdot)$ are influenced by the intracluster correlations. This can be shown by the fact that the approach to zero of both integrated mean square error and average estimated integrated mean square error is slower in the cases where $\rho = 0.2$ and $\rho = 0.5$ than in the case where $\rho = 0$.

Table 1
Simulation results for point estimators of $\hat{\mathbf{B}}$

		$\rho = 0$			$\rho = 0.2$			$\rho = 0.5$		
	l	Bias^2 ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar	Bias^2 ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar	Bias^2 ($\times 10^{-6}$)	Var ($\times 10^{-3}$)	Rvar
$\hat{B}_1$	10	5.77	1.07	1.13	3.12	1.1	1.01	0.23	1.19	1.33
	25	9.97	0.46	1.07	0.38	0.44	1.08	0.30	0.53	0.98
	50	0.54	0.21	1.08	0.13	0.27	0.93	0.026	0.21	1.18
	100	0.22	0.13	0.96	0.019	0.11	1.06	0.039	0.13	0.98
$\hat{B}_2$	10	0.36	3.32	1.13	1.54	3.74	0.92	1.26	3.5	1.78
	25	0.64	1.31	1.10	2.40	1.34	1.06	0.14	1.42	1.03
	50	0.31	0.75	0.94	1.27	0.85	0.94	0.16	0.76	0.97
	100	0.15	0.38	0.94	1.11	0.38	0.98	0.072	0.33	1.03

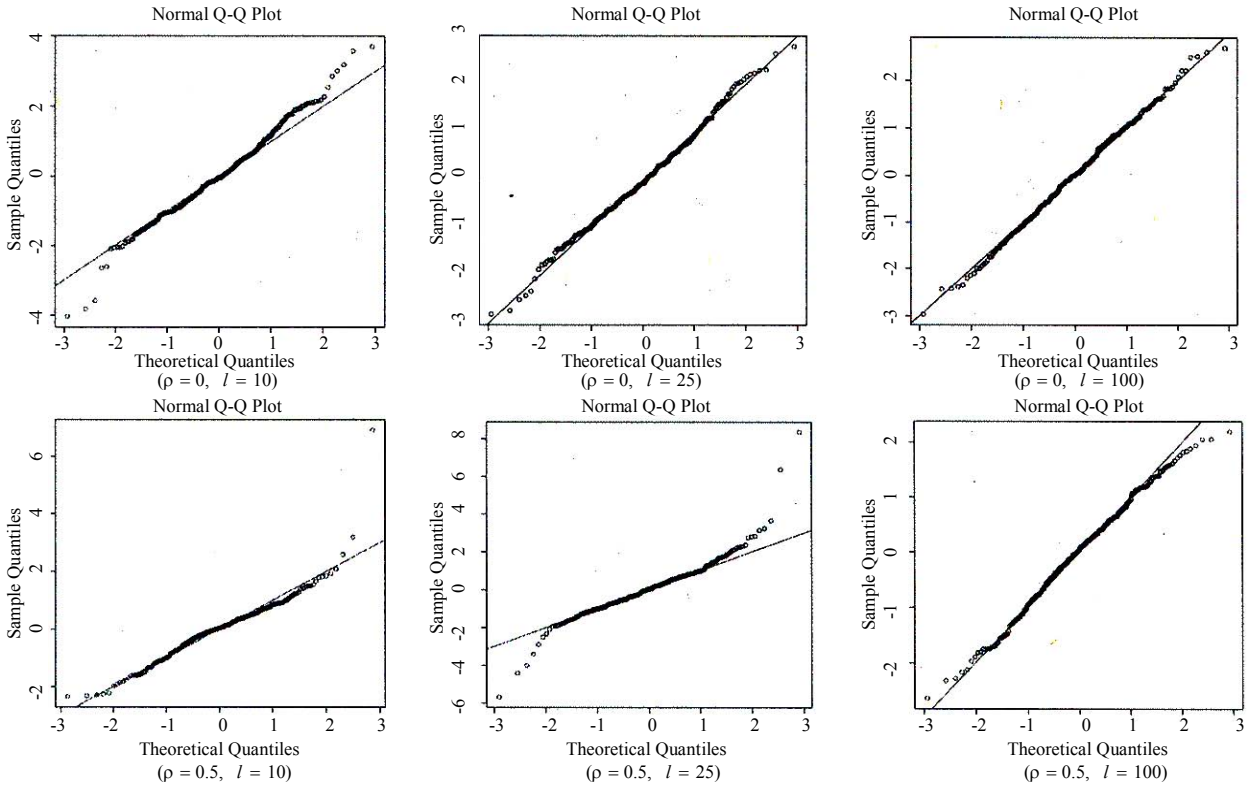

Figure 2 Quantile – quantile plots for standardized $\hat{B}_2$

Table 2
Bias and efficiency of $\hat{g}(z)$

ρ	l	AIMSE	IMSE	ABias^2 (×10-5)	RVar
0	10	0.37	0.42	5.29	1.27
	25	0.15	0.17	3.20	1.10
	50	0.074	0.086	3.29	1.09
	100	0.037	0.044	2.34	1.08
0.2	10	2.95	3.25	6.13	0.91
	25	1.22	1.17	3.71	1.04
	50	0.74	0.54	2.34	1.0
	100	0.26	0.27	7.08	0.98
0.5	10	8.143	8.877	3.73	0.92
	25	3.155	3.073	6.56	1.03
	50	1.461	1.599	2.86	1.15
	100	0.659	0.607	3.59	1.09

The coverage of the point-wise 95% confidence intervals for $g(\cdot)$ in Figure 3 varies between between 85% and 96%. The coverage improves as the sample size increases. The performance of $\hat{g}(\cdot)$ is, however, more sensitive to the lower effective sample size caused by the intracluster correlation. In particular, the coverages of the 95% confidence intervals in the cases of $\rho = 0.2$ and $\rho = 0.5$ are smaller than the 95% nominal confidence level when $l = 10$. The coverage improves as the number of primary sampling units increases for the cases of $\rho = 0$ and $\rho = 0.2$. For $\rho = 0.5$, the undercoverage is still present when the sample size increases to 100. It is also seen that at $z = 18$ or 64, the coverages are higher even than the nominal level; this is because the boundary effect of the local polynomial regression estimation causes larger bias at the two boundaries of the data. For $\rho = 0.5$, the effective sample size is low so that the boundary effect becomes severe, creating the downward spikes at 18 and 63.

It is worth pointing out that although the size of the primary sampling units is large (1,000), the sampling fraction is very small (0.05). Hence, this performance of the estimates would not change even though the size of the primary sampling units is small.

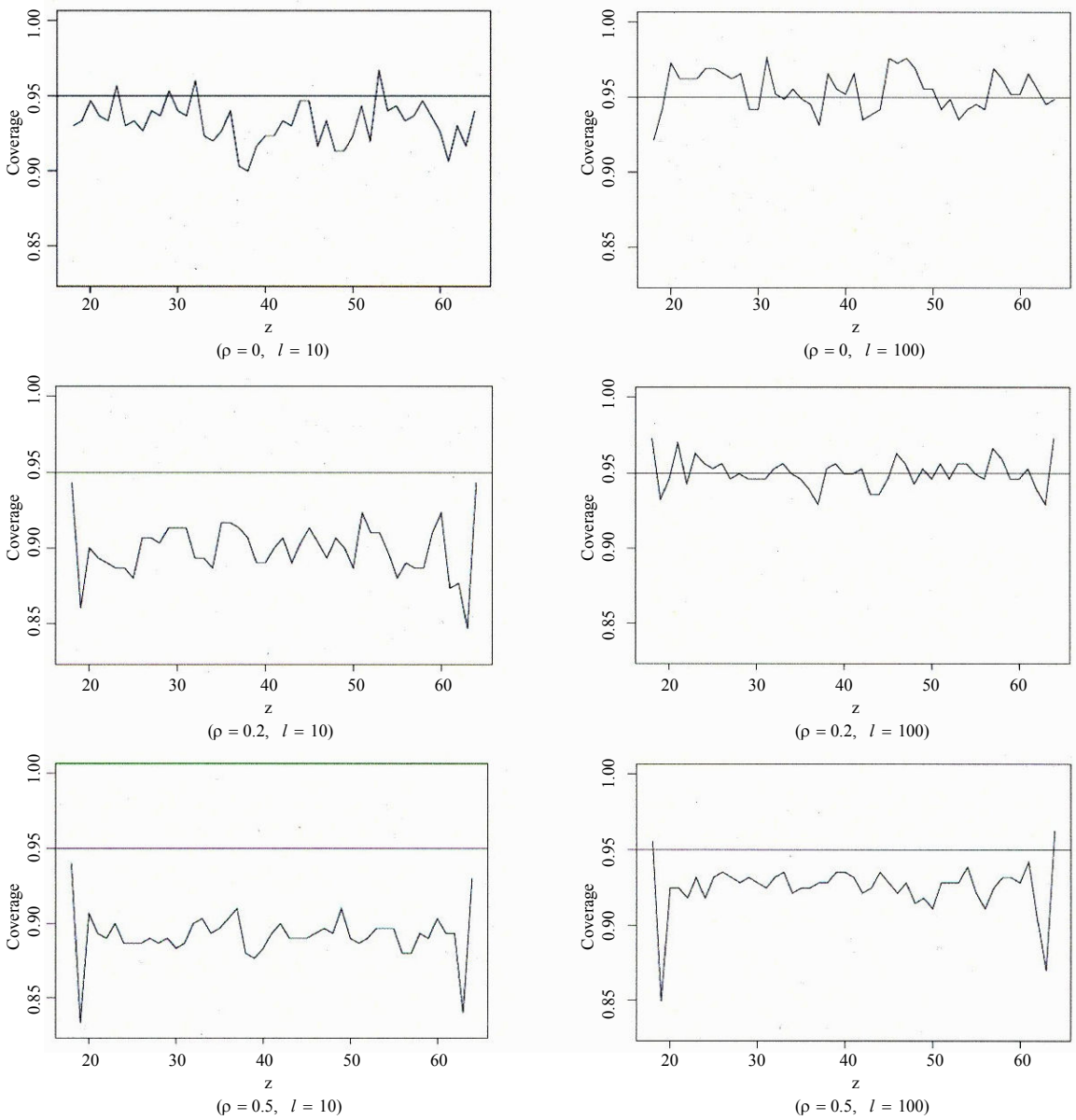

Figure 3 Coverage of the 95% confidence point-wise intervals for $g(z)$

5. Empirical illustrations

We now return to the example introduced in Section 1. For the purpose of illustrating the partial linear model, we examine the effects of age, gender, smoking status and physical activity on the body mass index (BMI) and the desired body mass index (DBMI). Similar to the measure BMI, DBMI is a derived variable for the question asking about the desired weight of a person. Since people stop growing for the age group for which we are interested, we use the actual height when calculating DBMI. We use age as the nonparametric covariate and treat the other factors as discrete variables. Since there are only 47 distinct points in the age variable, we bin the data set according to age. The bin size is set to unity such that there are 47 bins, with midpoints being 18, 19, …, 64. Among all the categorical explanatory variables, gender has two levels, male = 1 and female = 0; smoking status includes levels such as former smoker = 0, never smoked = 1, occasional smoker = 2, daily smoker = 3; and physical activeness is divided into three levels: active = 0, moderately active = 1 and inactive = 2. The regression models are (a) BMI = $g_1(\text{age}) + \mathbf{XB}_1 + \varepsilon_1$ and (b) DBMI = $g_2(\text{age}) + \mathbf{XB}_2 + \varepsilon_2$, where $\mathbf{X}$ is the design matrix including all the indicator variables.

Table 3 lists all the survey estimates of the linear coefficients in the models (a) and (b). On comparing BMI by gender, we found that male BMI is higher. Using former smoker as the base category, the coefficients of smoking status are all negative and significant, which suggests that former smokers tend to be heavier than people with other types of smoking status. The estimates also indicate that inactive people have higher BMI. With respect to the DBMI, p -values suggest that most of the life style related factors are not significant.

Table 3
Results for semiparametric regression models (a) and (b) (Values in the parenthesis are the standard errors)

Variable	$\hat{B}_1$	p-value	$\hat{B}_2$	p-value
Gender	1.45	0.00	2.80	0.00
	(0.05)		(0.05)	
Never Smoked	-0.45	0.00	-0.06	0.34
	(0.10)		(0.06)	
Occasional Smoker	-0.31	0.04	-0.00	0.96
	(0.17)		(0.10)	
Daily Smoker	-0.61	0.00	-0.12	0.03
	(0.09)		(0.06)	
Moderately Active	-0.33	0.00	-0.07	0.24
	(0.09)		(0.06)	
Active	-0.50	0.00	-0.14	0.07
	(0.09)		(0.09)	

In Figure 4, the estimated functions of age, $\hat{g}_1(\text{Age})$ and $\hat{g}_2(\text{Age})$, and their confidence bands are plotted versus different ages. It is found that, in both cases, the BMI and the DBMI are increasing functions of age.

Figure 5 gives the estimated functions of age, $\hat{g}_1(\text{Age})$ and $\hat{g}_2(\text{Age})$, for active and moderately active people. If we look at the age effect for female and male separately, we find that for females who are either active or moderately active on average the DBMI is lower than the BMI, whereas males with the same intensity of physical activity desire to be heavier before age 21. In addition, we also compare the age trends in the BMI and the DBMI for both the females and males. Due to the inconsistency between the female and male trends, we can conclude that there are interactions between the gender factor and age.

6. Conclusion

With the assistance of a partial linear model, we extend semi-parametric regression techniques to complex survey data. Asymptotic properties of the survey estimators are developed. Computation of the variance estimates of both the linear coefficients and the regression function rely on the variance estimates of survey totals and means. Provided that we obtain the required variance estimates of survey totals and means, we can apply this method using standard statistical packages.

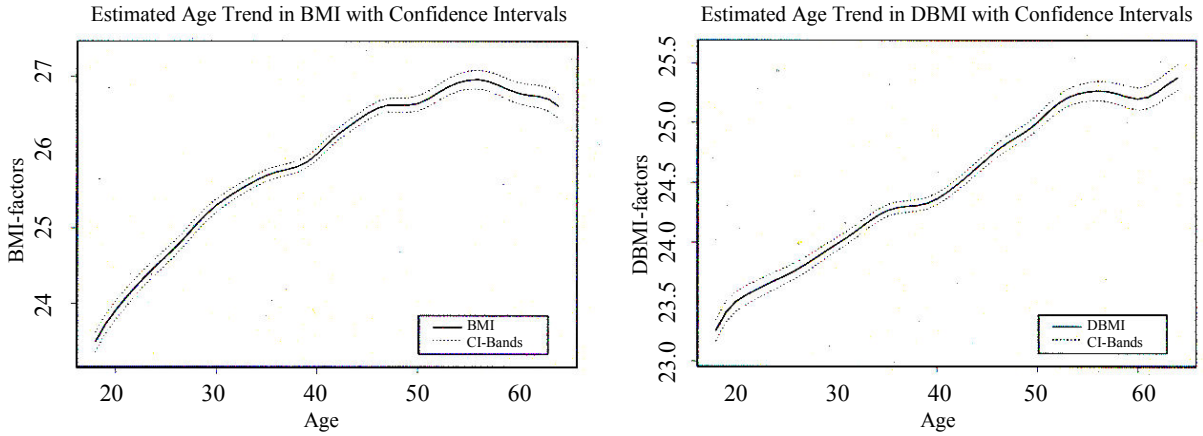

Figure 4 Estimated age trends in BMI and DBMI with 95% pointwise confidence intervals

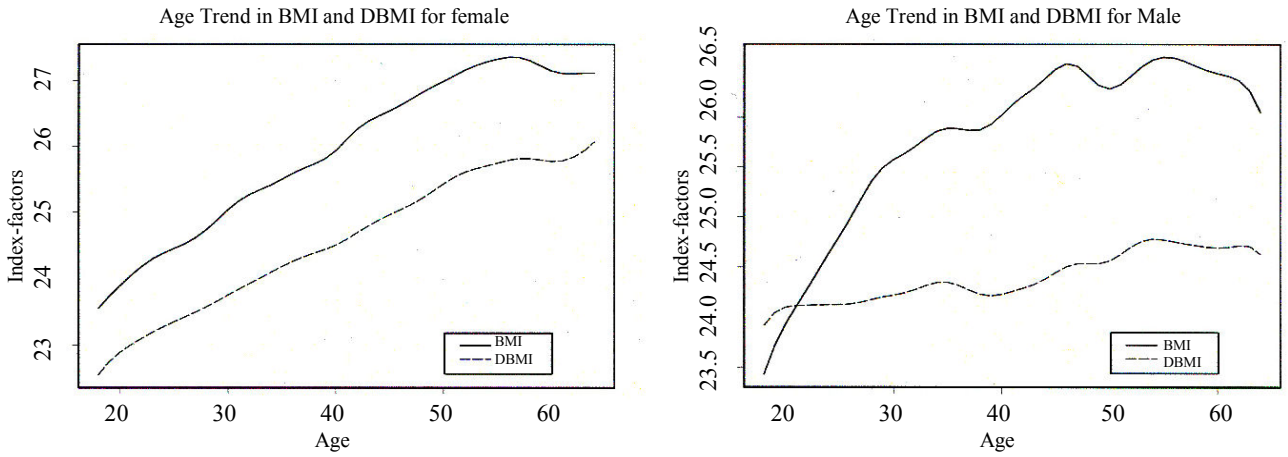

Figure 5 A comparison of estimated age trends in BMI to DBMI for both female and male who are active and moderately active

In the partial linear working model, we assume that there is no interaction between the parametric component and the nonparametric component. However, the empirical example of the age trends of the body mass index has illustrated that this assumption needs to be checked. In future work, we will relax the no interaction assumption. A direct approach to modelling interaction terms is to let the nonparametric component appear linearly in the interaction term. That is, we define the partial linear model as

$$\mathbf{y} = G(\mathbf{z}) + \mathbf{X}\boldsymbol{\beta} + \mathbf{X}H(\mathbf{z}) + \boldsymbol{\varepsilon}.$$

By testing the departure of $H(\mathbf{z})$ from zero, we can detect the existence of interaction.

When estimating conditional expectation on the nonparametric components for indicator discrete random variables, we propose to use generalized linear or additive models to conduct the estimation.

Appendix

A.1 Proof of lemma 1

Observing that entries of $\hat{\mathbf{u}}(\boldsymbol{\theta})$, $\hat{\mathbf{u}}_B(\boldsymbol{\theta})$ and $\hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ are either sample totals or ratios of sample totals, we can apply Lemmas 1.2.5 and 1.2.6 in Wang (2004) to establish this Lemma.

A.2 Proof of lemma 2

Each entry of $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ is just an estimated regression function with the local polynomial technique developed by Bellhouse and Stafford (2001). Theorem 2.2.1 in Wang (2004) shows that $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ is root-n consistent. Hence, since the dimension of $\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z})$ is finite, we can show that $\sqrt{n}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z}))$ is bounded in probability.

A.3 Proof of theorem 1

Since for the true $\boldsymbol{\theta}$, we have $\mathbf{u}(\boldsymbol{\theta}) = \mathbf{0}$, we can rewrite equation (13) as follows:

$$-\frac{\sqrt{n}\hat{\mathbf{u}}_B(\boldsymbol{\theta})}{N}(\hat{\mathbf{B}} - \mathbf{B}) \doteq$$

$$\left(\frac{\sqrt{n}}{N}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta})) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})\frac{\sqrt{n}}{N}\left(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})\right)\right)$$

$$+ \frac{\sqrt{n}}{N}||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||\varepsilon.$$

The standard argument in Rao (1973, page 387) yields

$$\sqrt{n}/N\,||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||\varepsilon \xrightarrow{P} 0.$$

Using the condition that the sampling fraction $f = n/N$ is constant as ν goes to infinity, we have,

$$\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) = -\left(\frac{\hat{\mathbf{u}}_B(\boldsymbol{\theta})}{N}\right)^{-1}$$

$$\left(\frac{\sqrt{n}}{N}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta})) + \hat{\mathbf{U}}_{\boldsymbol{\xi}}(\boldsymbol{\theta})\frac{f\sqrt{n}}{n}\left(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z})\right)\right).$$

Following from the results in Lemma 1, both $(\hat{\mathbf{u}}_{\mathbf{B}}(\boldsymbol{\theta})/N)^{-1}$ and $\hat{\mathbf{U}}_{\mathbf{m}_{\boldsymbol{\xi}}}(\boldsymbol{\theta})$ converge to their population values in probability. Lemma 2 indicates that the vector $\sqrt{n}(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(z) - \mathbf{m}_{\boldsymbol{\xi}}(z)) = O_p(1)$. Thus, $(f\sqrt{n}/n)(\hat{\mathbf{m}}_{\boldsymbol{\xi}}(\mathbf{z}) - \mathbf{m}_{\boldsymbol{\xi}}(\mathbf{z}))$ converges to a zero vector in probability as n goes to infinity. Finally, from the normality of $\sqrt{n}(\hat{\mathbf{u}}(\boldsymbol{\theta}) - \mathbf{u}(\boldsymbol{\theta}))/N$ stated in Lemma 1, we use the Slutsky Theorem to show the asymptotic normality of $\hat{\mathbf{B}}$.

A.4 Proof of lemma 3

Given that $\bar{\mathbf{r}} = \bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}$, we have $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) = \sqrt{n}[(\bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}) - \bar{\mathbf{R}}]$. Based on Theorem 1, we know that in the limit as ν goes to infinity, $\hat{\mathbf{B}}$ converges to $\mathbf{B}$ in probability. Hence, we have $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) \doteq \sqrt{n}((\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B}) - \bar{\mathbf{R}})$. The d^{th} entry of $(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B})$ is,

$$\bar{y}_d - \bar{x}_{1d} B_1 - \cdots - \bar{x}_{pd} B_p = \frac{1}{\hat{N}_d} \sum_{k \in \mathbf{s}_d} w_k (y_k - x_{1k} B_1 - \cdots - x_{pk} B_p).$$

That is, $(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B})$ is merely a vector of estimated binned means. Using the result from Shao (1996) on functions of sample means and "Cramer-Wold device" results found in Serfling (1980, page 18), we see that $\sqrt{n}(\bar{\mathbf{y}} - \bar{\mathbf{x}}\mathbf{B} - \bar{\mathbf{R}})$ converges to a random vector distributed normally. Thus, using this indirect Slutsky idea, we have proved the normality of $\sqrt{n}(\bar{\mathbf{r}} - \bar{\mathbf{R}}) = \sqrt{n}[(\bar{\mathbf{y}} - \bar{\mathbf{x}}\hat{\mathbf{B}}) - \bar{\mathbf{R}}]$.

A.5 Proof of theorem 2

The proof follows the same argument that $\hat{g}(z_d)$ is a function of domain mean and proportions as does in the proof of theorem 2.2.1 in Wang (2004).

Acknowledgements

This work is supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors are grateful to Mary Thompson for her valuable comments and suggestions on the early draft of this paper. The authors also wish to thank the Associate editor and two referees for their very helpful comments.

References

Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex survey. *Statistca Sinica*, 9, 407-424.

Bellhouse, D.R., and Stafford, J.E. (2001). Local polynomial regression in complex survey. *Survey Methodology*, 27, 197-203.

Bickel, P.J., and Freedman, D.A. (1983). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

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

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

Buskirk, D.T., and Lohr, L.S. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

Fan, J., and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society*, series B, 57, 371-394.

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

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

Hàjek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koelemenyei*, 5, 361-374.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

Jones, M.C. (1989). Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84, 733-741.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balance repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535–545.

Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.

Ontario Health Survey (1992). *Ontario Health Survey*: *User's Guide.* Ministry of Health, Toronto, Ontario, Canada.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd Ed.). New York: John Wiley & Sons, Inc.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56, 931-954.

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

Serfling, R.J. (1980). *Approximation Theorem of Mathematical Statistics*. New York: John Wiley & Sons, Inc.

Shao, J. (1996). Resampling methods in sample survey. *Statistics*, 27, 203-254.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society*, Series B, 50, 413-436.

Thompson, M.E. (1997). *Theory of Sample Survey* (1st Ed.). New York: Chapman and Hall.

Wang, Z. (2004). *Some Nonparametric Regression Techniques for Complex Survey Data*. Unpublished Ph.D. thesis, The University of Western Ontario, London, Ontario, Canada.

Zheng, H., and Little, R.J.A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, 30, 209-218.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2009.

R. Andridge, *Ohio State University*
J.-F. Beaumont, *Statistics Canada*
E. Berg, *Iowa State University*
J.M. Brick, *Westat, Inc.*
D. Cantor, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *University of Wollongong*, *Australia*
D. Chapman, *Federal Deposit Insurance Corporation*
A.-S. Charest, *Carnegie-Mellon University*
S. Chatterjee, *University of Minnesota*
M. Cohen, *National Academy of Sciences/Committee on National Statistics*
S. Cohen, *National Science Foundation*
M.P. Couper, *University of Michigan*
R. Curtin, *National Centre for Health Statistics*
E. Dagum, *University of Bologna*
G. Datta, *University of Georgia*
P.-P. de Wolf, *Statistics Netherlands*
P. Dick, *Statistics Canada*
J. Dixon, *Bureau of Labor Statistics*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
V. Estevao, *Statistics Canada*
E. Fabrizi, *University of Bergamo*, *Italy*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M. Ghosh, *Univerisy of Florida*
S. Godbout, *Statistics Canada*
C. Goga, *Université de Bourgogne*
B. Gross, *ABS*
R.M. Groves, *U.S. Census Bureau*
R. Harter, *National Opinion Research Centre*
S. Haslett, *Massey University, New Zealand*
D. Haziza, *Université de Montréal*
M.A. Hidiroglou, *Statistics Canada*
G. James, *Office for National Statistics*, *UK*
L. Jang, *Statistics Canada*
J. Jiang, *University of California*, *Davis*
D. Judkins, *Westat Inc.*
C. Julien, *Statistics Canada*
D. Kasprzyk, *Mathematica Policy Research*
R.S. Kenett, *KPA Ltd.*, *Raanana*, *Israel and University of Torino*, *Italy*
J.-K. Kim, *Iowa Stat University*
J.-M. Kim, *University of Minnesota-Morris*
P. Kokic, *CSIRO*
P. Kott, *National Agricultural Statistics Service*
S. Laaksonen, *University of Helsinki*
D. Ladiray, *INSEE*
P. Lahiri, *JPSM*, *University of Maryland*
P. Lavallée, *Statistics Canada*
C. Leon, *Statistics Canada*
R. Little, *University of Michigan*
B. Liu, *Westat Inc.*
L. Mach, *Statistics Canada*
T. Maiti, *Iowa State University*
H. Mantel, *Statistics Canada*
J. Maples, *U.S. Census Bureau*
A. Matei, *Université de Neuchâtel*, *Suisse*
C. McLaren, *Office for National Statistics*, *UK*
Y. McNab, *UBC*
F. Mecatti, *University of Milan-Bicocca*, *Italy*
S.M. Miller, *Bureau of Labor Statistics*
L. Mohadjer, *Westat Inc.*
G.E. Montinari, *University of Perugia*, *Italy*
F.A.S. Moura, *Universidade do Brasil-UFRJ*
Y. Mpetsheni, *Statistics South Africa*
G. Nathan, *Hebrew University*
T. Nayak, *George Washington University*
J. Opsomer, *Colorado State University*
S.P. Paben, *Bureau of Labor Statistics*
M. Park, *Korea University*
Z. Patak, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
M. Pratesi, *Università di Pisa*
L. Qualité, *Université de Neuchâtel*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
S. Rubin-Bleuer, *Statistics Canada*
A. Ruiz-Gazen, *Université des Sciences Sociales de Toulouse*
H. Saigo, *Waseda University*
N. Salvati, *Università di Pisa*
C.-E. Särndal, *Université de Montréal*
O. Sautory, *INSEE*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
G. Shapiro, Independent consultant
N. Shlomo, *University of Southampton*
D.B.N. Silva, *Office for National statistics*, *U.K.*
P. do N. Silva, *University of Southampton*
S. Sinha, *Carleton University*
C.J. Skinner, *University of Southampton*
E. Slud, *University of Maryland and US Census Bureau*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R. Vaillant, *University of Maryland*
V.J. Verma, *Università degli Studi di Siena*
C. Walker, *Statistics Canada*
D. Willimack, *U.S. Census Bureau*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard Medical School*
F. Zhang, *National Science Foundation*

Acknowledgements are also due to those who assisted during the production of the 2009 issues: Eric Rancourt of Corporate Planning and Evaluation Division, Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau of Household Survey Methods Division, Nick Budko and Carole Jean-Marie of Business Survey Methods Division, Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Denis Coutu, Darquise Pellerin and Isabelle Poliquin (Dissemination Division), Sheri Buck (Systems Development Division) and Sylvie Dupont (Official Languages and Translation Division).

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 25, No. 2, 2009

Contents
Volume 25, No. 3, 2009

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 37, No. 1, March/mars 2009

Volume 37, No. 2, June/juin 2009

Volume 37, No. 3, September/septembre 2009

GUIDELINES FOR MANUSCRIPTS

Before finalizing your text for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. 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 consecutively with arabic numerals on the right if they are to be referred to later.
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 Italics are used for emphasis.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables.

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.