Journal of Official Statistics
vol. 32, i. 3 (2016)

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 549–578, http://dx.doi.org/10.1515/JOS-2016-0029

Weighting Strategies for Combining Data from Dual-Frame Telephone Surveys: Emerging Evidence from Australia

Bernard Baffour[1], *Michele Haynes*[1], *Mark Western*[1], *Darren Pennay*[2,3], *Sebastian Misson*[3], *and Arturo Martinez*[1]

Until quite recently, telephone surveys have typically relied on landline telephone numbers. However, with the increasing popularity and affordability of mobile phones, there has been a surge in households that do not have landline connections. Additionally, there has been a decline in the response rates and population coverage of landline telephone surveys, creating a challenge to collecting representative social data. Dual-frame telephone surveys that use both landline and mobile phone sampling frames can overcome the incompleteness of landline-only telephone sampling. However, surveying mobile phone users introduces new complexities in sampling, nonresponse measurement and statistical weighting. This article examines these issues and illustrates the consequences of failing to include mobile-phone-only users in telephone surveys using data from Australia. Results show that there are significant differences in estimates of populations' characteristics when using information solely from the landline or mobile telephone sample. These biases in the population estimates are significantly reduced when data from the mobile and landline samples are combined and appropriate dual-frame survey estimators are used. The optimal choice of a dual-frame estimation strategy depends on the availability of good-quality information that can account for the differential patterns of nonresponse by frame.

Key words: Dual-frame telephone surveys; mobile phone sampling; nonresponse; weighting.

1. Introduction

The implementation of national social surveys is important for measuring social phenomena. In many countries, computer-assisted telephone interviewing (CATI) has become the most common mode for conducting such surveys, chiefly because of the relatively lower costs than face-to-face interviewing (Keeter et al. 2000; Keeter et al. 2006; Steeh 2008). However, telephone ownership is not universal and specific segments of the population, such as lower-income and ethnic-minority people, are at risk of being systematically excluded (Tucker et al. 2007; Brick et al. 2011; Busse and Fuchs 2012).

[1] The University of Queensland - Institute for Social Science Research, Building 39A Campbell Road St Lucia, Brisbane, Queensland, 4067, Australia. Emails: b.baffour@uq.edu.au, m.haynes@uq.edu.au, m.western@uq.edu.au and amartinezjr@adb.org.
[2] Australian National University - Australian Centre for Applied Social Research Methods, Canberra, Australian Capital Territory, Australia. Email: darren.pennay@srcentre.com.au
[3] The Social Research Centre - Research Methodology, Melbourne, Victoria, Australia. Email: sebastian.misson@srcentre.com.au
Acknowledgments: We would like to thank the editor, associate editor and the three referees for their insightful comments and suggestions which considerably improved the article. This research was supported under Australian Research Council's Linkage Projects funding scheme (project number LP130100744 "Enhancing social research in Australia using dual-frame telephone surveys").

On average, current response rates of traditional landline telephone surveys have fallen to less than 60% for surveys conducted by national statistical institutes (Groves and Peytcheva 2008). For nongovernment surveys, response rates can be as low as ten percent (Pew Research 2012). These declines have implications for the representativeness of the sample with regard to the target population.

With mobile telephone use becoming increasingly prevalent in the population, including mobile telephone owners in the sampled population has the potential to address the coverage bias associated with traditional landline telephone surveys. This is because those who are more likely to be excluded from landline-based surveys often own a mobile telephone (Keeter et al. 2007; Pennay 2010; Brick 2011; Busse and Fuchs 2012). For instance, about 95% of adult Australians own a mobile telephone, compared to only 80% who own a landline (Australian Communications and Media Authority 2011). Additionally, there is an increasing trend for individuals to discard their landlines and rely solely on mobiles. In Australia, the proportion of adults who own a mobile telephone and live in a household without a landline telephone connection on which they receive calls has grown from five percent in 2005 to 29% in 2014 (Australian Communications and Media Authority 2015). This mobile-only population is excluded from surveys that rely solely on landline telephone sampling frames. Australian patterns mirror the experience in the United States, as shown in Figure 1.

Similar trends have been reported in Canada and in Europe (Brick et al. 2011; Mohorko et al. 2013). Exclusion of mobile-only individuals from social surveys has adverse consequences for survey estimates, as there are sociodemographic differences between individuals who own a mobile telephone and those who own a landline telephone (Brick et al. 2006 in the USA; Callegaro and Possio 2004 in Italy; Kuusela et al. 2008 in Finland; Vicente and Reis 2009 in Portugal; and Arcos et al. 2014 in Spain). Individuals living in mobile-only households are more likely to be younger, male, of a lower socioeconomic status, foreign-born, students, highly transient, in large cities, and in full-time employment (Blumberg and Luke 2014). These households are also more likely to experience poor health and adverse socioeconomic outcomes. (Barnes et al. 2015; Thomée et al. 2011;

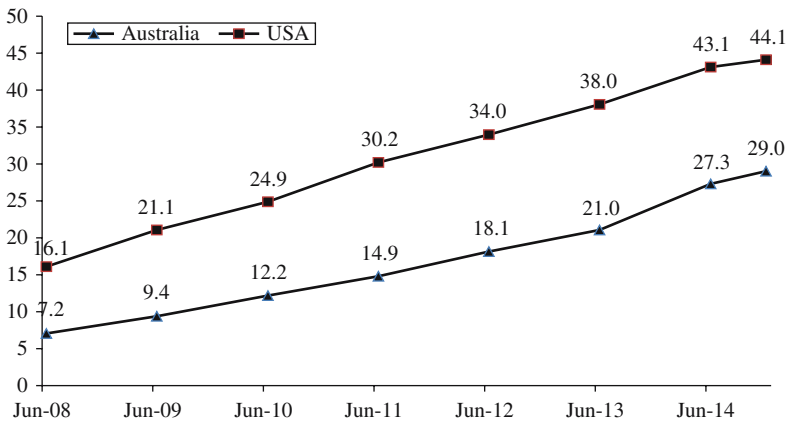


Fig. 1. The percentage of Australian and US adults with a mobile telephone and no fixed-line telephone service, June 2008 to June 2014. Sources: Australian data: ACMA (2015). USA data: Blumberg and Luke (2015).

Hu et al. 2011). There are also differences between these groups in health and behavioural risk factors, such as smoking and alcohol use (Barr et al. 2012; Livingston et al. 2013). Altogether, this literature suggests that any survey that is directed at either landline-only or mobile-only individuals will suffer from coverage bias. To address this, an effective approach is to redesign traditional telephone surveys to include mobile telephones. Researchers agree that using dual-sampling frames will be an integral part of telephone surveys in the future (AAPOR 2010; Brick 2011).

The dual-frame telephone sampling approach involves supplementing telephone numbers from a randomly generated landline sampling frame with an independent sample of randomly generated mobile telephone numbers. This introduces additional complexity into the survey design and analysis. A number of estimators have been proposed to estimate population characteristics using data from dual-frame surveys. While each estimator has its advantages and limitations, the existing literature offers limited guidelines on choosing the appropriate estimator in a specific research context. Another unresolved issue is the extent to which nonresponse adjustments based on auxiliary data can be used to improve the efficiency of dual-frame estimation. Many of these methodological problems still remain in Australia, due to the lack of official statistics on mobile telephone usage and the fact that dual-frame telephone interviewing is a relatively recent innovation in comparison to other countries.

This article makes two contributions to the emerging literature on dual-frame surveys. First, it provides an up-to-date review of the available dual-frame estimators and their suitability to estimating population quantities. Second, it provides an empirical assessment of these estimators using nationally representative Australian data from dual-frame surveys. The findings presented in this article, while set within an Australian survey-research context, will be informative to researchers in other countries facing similar design decisions.

2. Dual-Frame Sampling Theory

2.1. Background

The objective of the dual-frame approach is to draw subpopulation samples from different sampling frames that, when combined, provide full coverage of the target population. The concept of dual-frame sampling dates back to the 1950s (Hartley 1962), but has not been applied to sampling from mobile and landline telephone frames until very recently (Lohr 2009; Arcos et al. 2014). Dual-frame surveys have become widely used by national statistical agencies, particularly for health surveys such as the US National Health Interview Survey (Blumberg and Luke 2007, 2014), and the Canadian Community Health Survey (Béland 2002). They often provide improved access to hard-to-reach populations (Kalton and Anderson 1986; Iachan and Dennis 1993; Flores Cervantes and Kalton 2008) and can reduce sampling costs by tailoring interview mode to respondent needs (Kennedy 2007; Lopez and Gonzalez-Barrera 2013).

Typically, telephone sampling frames will overlap, so that simply taking the union of all the frames will lead to duplication of the individuals in the population who appear in more than one frame. Duplication within combined sampling frames has posed a theoretical

problem, with researchers interested in (1) the best way of combining the disparate information from the different frames, and (2) how to determine the reliability of the derived sample estimates (Fuller and Burmeister 1972; Hartley 1974; Bankier 1986; Skinner 1991). The problem of duplication is almost universal for dual-frame telephone surveys, as many people are likely to have access to both landline and mobile telephones.

2.2. *Estimation of Population Quantities from Combined Sampling Frames*

Figure 2 depicts the general situation when there are two sampling frames in telephone surveys (a landline telephone frame L and a mobile telephone frame M), both with under-coverage of the target population, but when combined leading to improved population coverage. The frames L and M generate three mutually exclusive domains – l (units in L alone), m (units in M alone) and lm (units in both L and M). Following the classical texts of Hartley (1962, 1974) and Skinner (1991), Skinner and Rao (1996), and Lohr and Rao (2000, 2006), we denote the landline and mobile population sizes as N_L and N_M, and the domain sizes as N_l, N_m, and N_{lm}, respectively. It follows that $N_L = N_l + N_{lm}$ and $N_M = N_m + N_{lm}$. Also, the total population size satisfies $N = N_L + N_M - N_{lm} = N_l + N_m + N_{lm}$.

Similarly, let S_L and S_M be samples, of size n_L and n_M, drawn independently from the landline L and mobile M frames, respectively. Denote the overlapping sample as S_{lm}, with sample size n_{lm}. Both the size of the overlapping population N_{lm} and the size of the sample n_{lm} are unknown. However, we do know that $n_L = n_l + n_{lm}^L$ and $n_M = n_m + n_{lm}^M$, where n_{lm}^L is the overlap sample from the landline frame L, and n_{lm}^M is the overlap sample from the mobile telephone frame M.

Finally, let y_i denote the value associated with the observation for individual i, then the population statistic, given by $Y = \sum_{i=1}^{N} y_i$, is simply a sum of the units that appear in the domains l, m, and the overlap lm, that is landline only, mobile telephone only and both landline and mobile telephone users, respectively. Thus,

$$Y = Y_l + Y_m + Y_{lm} = \sum_{i\epsilon l} y_i + \sum_{i\epsilon m} y_i + \sum_{i\epsilon lm} y_i. \tag{1}$$

Suppose also that y_i is observed for each individual in the samples S_L and S_M, then the estimation problem is to use these data to construct a suitable estimator $\hat{Y}$ of Y. It is also of interest to find an estimator of the variance of $\hat{Y}$, denoted $var(\hat{Y})$. Equation (1) shows how Y can be computed using population data. Since the population universe, $\mathcal{U}$, can be decomposed into l, m and lm, then, when we have a dual-frame sample, we can similarly

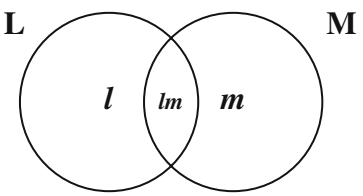


Fig. 2. Landline phone sampling frame L *and mobile phone sampling frame* M *with overlap creating domains l, m and lm.*

decompose the sample into the landline-only sample, the mobile-only sample and the dual sample, represented by S_l, S_m and S_{lm}, respectively. It follows that the estimator for Y using the sample information is now given by (2),

$$\hat{Y} = \sum_{i \epsilon S_l} w_i y_i + \sum_{i \epsilon S_m} w_i y_i + \sum_{i \epsilon S_{lm}} w_i y_i = \hat{Y}_l + \hat{Y}_m + \hat{Y}_{lm} \tag{2}$$

where w_i is the probability weight associated with unit i, and $\hat{Y}_l$, $\hat{Y}_m$, and $\hat{Y}_{lm}$ are sample statistics computed using information from the landline-only, mobile-only and dual samples.

The component $\hat{Y}_{lm}$ in Equation (2) can be estimated using either the data from individuals who reported having both landline and mobile phones from the landline sample, that is, S_{lm}^L, or the data from individuals who reported having both landline and mobile phones from the mobile sample, that is, S_{lm}^M. Although S_{lm}^L and S_{lm}^M are theoretically two independent samples from S_{lm}, it is convenient to think of both samples as duplicates from the same domain, and hence, the dual-frame estimators of the population quantity Y can be calculated using a weighted estimate of the overlap according to frame L and frame M,

$$\hat{Y}_{lm} = \theta \hat{Y}_{lm}^L + (1 - \theta) \hat{Y}_{lm}^M. \tag{3}$$

Combining (2) and (3) gives

$$\hat{Y} = \hat{Y}_l + \hat{Y}_m + \theta \hat{Y}_{lm}^L + (1 - \theta) \hat{Y}_{lm}^M \tag{4}$$

where θ is referred to as the composite weight, $\hat{Y}_{lm}^L$ denotes the estimated statistic for individuals who use both landline and mobile phones derived from the landline sample, while $\hat{Y}_{lm}^M$ denotes the estimated statistic for sampled units who own both landline and mobile telephones derived from the mobile sample.

In general, methods to estimate $\hat{Y}$ from a dual-frame survey differ according to how the information from the individuals in the overlapping samples is used. Following Hartley (1962, 1974), Fuller and Burmeister (1972), Skinner and Rao (1996), and Lohr and Rao (2000), the choice of the composite weight θ is selected subject to some optimisation criteria which aim to minimise a loss function with respect to bias, variance or cost. Typically, θ is unknown and is replaced by $\hat{\theta}$, which is estimated from sample data (Skinner and Rao 1996), and the objective in dual-frame estimation is to find $\hat{\theta}$ such that estimators for the population quantity Y are reliable and unbiased. Common approaches to combining the data from dual frames are summarised below.

2.2.1. Screening Estimator

The two extreme types of composite estimators for dual-frame surveys correspond to screening out sampled dual telephone users from either the landline or mobile telephone sample. In particular, a landline-only screener shown in (5) screens out those dual telephone users who were found in the landline sample. The mobile-only screener (6), on the other hand, screens out the mobile telephone sample dual users. This is accomplished by setting θ to be equal to 0 or 1 in (3), such that

$$\text{when} \quad \theta = 0, \quad \hat{Y} = \hat{Y}_l + \hat{Y}_m + \hat{Y}_{lm}^M, \tag{5}$$

$$\text{and when} \quad \theta = 1, \quad \hat{Y} = \hat{Y}_l + \hat{Y}_m + \hat{Y}_{lm}^L. \tag{6}$$

The screening approach is conceptually simple and estimation is straightforward. Note that (5) and (6) provide lower and upper bounds for other composite estimators.

2.2.2. Average Estimator

The average estimator, also known as the multiplicity estimator (Mecatti 2007) or fixed-weight estimator (Hartley 1962), is the most commonly used estimator among the class of composite survey estimators proposed in the dual-frame sampling literature. Here, dual telephone users from the landline telephone sample contribute the same amount as the dual users from the mobile telephone sample in estimating the quantity of interest for all dual users. Thus, in this case $\theta = \frac{1}{2}$, and

$$\hat{Y}_{ave} = \hat{Y}_l + \hat{Y}_m + \frac{1}{2}\hat{Y}_{lm}^L + \frac{1}{2}\hat{Y}_{lm}^M. \tag{7}$$

There are several advantages to using this estimator. First, it is straightforward to compute and implement because the value of θ does not depend on the quantity of interest (Mecatti 2007). In the absence of nonresponse error, it can be shown that the multiplicity estimator is a consistent estimator of Y (Bankier 1986). It is also straightforward to estimate the variance of $\hat{Y}_{ave}$, since θ is fixed across all individuals. However, the average estimator is not necessarily efficient and other estimators are often more statistically reliable because they use more information about the different frames.

2.2.3. Hartley (Minimum-Variance) Estimator

Although both the screening and average estimators are simple to compute, they are not necessarily as efficient as other estimators that incorporate information about how the data were collected under the different sampling frames. For instance, if the estimator $\hat{Y}_{lm}^L$ is more reliable than $\hat{Y}_{lm}^M$ in regards to estimating the overlapping domain quantity Y_{lm}, then it would make sense to place more weight on $\hat{Y}_{lm}^L$ than on $\hat{Y}_{lm}^M$. One way to achieve this is to minimise the variance of the target parameter $\hat{Y}$. Hartley (1962, 1974) showed that the variance of (4) is minimised when $\theta = \theta_H$,

$$\theta_H = \frac{Var\left(\hat{Y}_{lm}^M\right) + Cov\left(\hat{Y}_m, \hat{Y}_{lm}^M\right) - Cov\left(\hat{Y}_l, \hat{Y}_{lm}^L\right)}{Var\left(\hat{Y}_{lm}^L\right) + Var\left(\hat{Y}_{lm}^M\right)}. \tag{8}$$

Thus, the Hartley estimator takes the form

$$\hat{Y}_H = \hat{Y}_l + \hat{Y}_m + \hat{\theta}_H\hat{Y}_{lm}^L + \left(1 - \hat{\theta}_H\right)\hat{Y}_{lm}^M. \tag{9}$$

The advantage of the Hartley estimator is that it is asymptotically optimal among all fixed-weight composite estimators. However, since the variance and covariance terms in (8) are unknown, the optimal value θ_H must be estimated from the data and different response variables will generate different values for $\hat{\theta}_H$, leading to internal inconsistency. For complex surveys, this inconsistency may be large (Lohr and Rao 2006). In addition, as the estimated variance and covariance terms depend on the quantity being estimated, the randomness of $\hat{\theta}_H$ should be taken into account when computing the variance of (9).

2.2.4. Fuller and Burmeister Estimator

Fuller and Burmeister (1972) proposed modifying the Hartley estimator by making use of additional information about the overlapping units. The Fuller-Burmeister estimator is given by

$$\hat{Y}_{FB} = \left(N_L - \hat{N}_{lm}\right)\hat{y}_l + \left(N_M - \hat{N}_{lm}\right)\hat{y}_m + \hat{N}_{lm}\left(\hat{\theta}_{FB}\hat{y}^L_{lm} + \left(1 - \hat{\theta}_{FB}\right)\hat{y}^M_{lm}\right) \tag{10}$$

where $\hat{y}_l$, $\hat{y}_m$, $\hat{y}^L_{lm}$, and $\hat{y}^M_{lm}$ are the sample estimates from the landline-only, mobile-only, dual users (from landline), and dual users (from mobile) samples. Finally, $\hat{N}_{lm}$ is the smallest root of the quadratic equation

$$(n_L + n_M)x^2 - \left(n_L N_M + n_M N_L + n^L_{lm} N_L + n^M_{lm} N_M\right)x + \left(n^L_{lm} + n^M_{lm}\right)N_L N_M = 0. \tag{11}$$

The Fuller-Burmeister weights differ for different response variables and are therefore not internally consistent. Nevertheless, it can be shown that the Fuller-Burmeister estimator is more efficient than the Hartley estimator (Skinner and Rao 1996). Skinner and Rao (1996) proposed modifying this estimator by accounting for the complex sampling design through pseudomaximum-likelihood estimation, and showed that the weight adjustments do not depend on the covariances of the particular response being studied.

2.2.5. Pseudomaximum-Likelihood Estimator (Skinner and Rao Estimator)

In complex surveys, maximum-likelihood estimators do not usually have closed analytic forms (Gong and Samaniego 1981). To provide an internally consistent composite estimator for dual-frame surveys, Skinner and Rao (1996) proposed a pseudomaximum-likelihood estimator that uses a fixed value θ_{PML} for any population characteristic of interest. The proposed approach draws strongly on the ideas of Fuller and Burmeister (1972) but aims to find a consistent dual-frame estimator through pseudolikelihood maximisation. Therefore, in order to obtain the pseudomaximum-likelihood estimator of the (unknown) population quantity, that is population total, average or proportion, first define $\hat{N}^L_{lm} = \frac{N_L}{n_L} n^L_{lm}$ and $\hat{N}^M_{lm} = \frac{N_M}{n_M} n^M_{lm}$. The pseudomaximum-likelihood estimator is given by

$$\begin{aligned}\hat{Y}_{PML} = &\left(N_L - \hat{N}^{PML}_{lm}(\theta)\right)\hat{y}_l + \left(N_M - \hat{N}^{PML}_{lm}(\theta)\right)\hat{y}_m \\ &+ \hat{N}^{PML}_{lm}(\theta)\left\{\theta\hat{y}^L_{lm} + (1-\theta)\hat{y}^M_{lm}\right\}\end{aligned} \tag{12}$$

where $\hat{N}^{PML}_{lm}(\theta)$ is a function of $\hat{N}^L_{lm}, \hat{N}^M_{lm}$ and $\theta = \theta_{PML}$ which is the smaller of the roots of the quadratic equation

$$\left[\frac{\theta}{N_M} + \frac{1-\theta}{N_L}\right]x^2 - \left[1 + \theta\frac{\hat{N}^L_{lm}}{N_M} + (1-\theta)\frac{\hat{N}^M_{lm}}{N_L}\right]x + \theta\hat{N}^L_{lm} + (1-\theta)\hat{N}^M_{lm} = 0. \tag{13}$$

Skinner and Rao (1996) showed that the asymptotic variance of $\hat{N}^{PML}_{lm}(\theta)$ is minimised when

$$\theta_P = \frac{N_l N_M Var\left(\hat{N}^L_{lm}\right)}{N_l N_M Var\left(\hat{N}^L_{lm}\right) + N_m N_L Var\left(\hat{N}^M_{lm}\right)}. \tag{14}$$

In practice, N_l, N_m and the variances are not known and so are estimated from the data, resulting in

$$\hat{\theta}_P = \frac{\hat{N}_l N_M \widehat{Var}\left(\hat{N}_{lm}^L\right)}{\hat{N}_l N_M \widehat{Var}\left(\hat{N}_{lm}^L\right) + \hat{N}_m N_L \widehat{Var}\left(\hat{N}_{lm}^M\right)}, \tag{15}$$

where $\hat{N}_{lm}^{PML}(\theta)$ is the Fuller-Burmeister estimate of the overlapping population, and $\hat{N}_l = N_L - \hat{N}_{lm}^{PML}(\theta) \approx N_L - \hat{N}_{lm}^L$ and $\hat{N}_m = N_M - \hat{N}_{lm}^{PML}(\theta) \approx N_L - \hat{N}_{lm}^M$.

The pseudomaximum-likelihood estimator avoids the internal consistency problems present in the Hartley and Fuller-Burmeister estimators and has smaller mean squared error even in the presence of domain misclassification (Lohr and Rao 2000, 2006).

2.2.6. Single-Frame Estimator

The estimators discussed in Subsubsections 2.2.1 – 2.2.5 are based on choosing the composite weight subject to an optimisation criterion, for example minimising the variance in the class of linear unbiased estimators. The single-frame estimator assumes just one frame that encompasses information about both mobile telephone and landline use for each individual. Unlike the previous composite estimators that entail two-stage estimation of the population quantity, the single-frame estimator has an implicit adjustment in the estimation of the survey weights (i.e., inverse of selection probabilities) for sampled units from different frames. Assuming that landline and mobile telephone samples were drawn independently, the inclusion probability for the ith sampled individual is given by $\pi_i^L + \pi_i^M - \pi_i^L \pi_i^M$; thus, $w_i = \frac{1}{\pi_i^L + \pi_i^M - \pi_i^L \pi_i^M}$.

The single-frame estimator uses the same set of weights for all response variables, and is therefore internally consistent (for details, see Bankier 1986). However, calculating the weights in the overlapping domain requires knowledge of the inclusion probability of each unit for both frames. The single-frame estimator is always less efficient than the Hartley estimator (Skinner and Rao 1996).

2.2.7. Adjusting for Differential Nonresponse from Dual Frames

Accounting for the multiple-frame coverage is one stage of the weighting adjustment, but this does not result in unbiased population estimates in most cases (Brick et al. 2011). Despite the best efforts, in every survey nonresponse occurs and is differential by social, economic, demographic, and geographical characteristics. Therefore, a second weighting adjustment is required to ensure that the weighted sample is representative of the population. This is achieved through calibrating the estimates to be consistent with known population benchmarks through poststratification (Holt and Smith 1979; Little 1993). While there are a number of different methods of poststratification, we will be using poststratification raking to repeatedly adjust the sample margins to the corresponding population control marginal totals. By using auxiliary information, we can ensure that the sample aligns to the population benchmarks for a set of characteristics. In Australia, raking is used in most official surveys to adjust for the effect of nonresponse (ABS 2014).

The poststrata (or population benchmarks) used for the raking are location (State capital, Rest of State), age group (18–24, 25–39, 40–49, 50–64, 65+), sex (Male, Female),

educational attainment (University Graduate vs Non-University Graduate), birth place (Australian, Non-Australian), and telephone status (Mobile only, Dual user, Landline only). Aggregate data available on these population characteristics were available from the 2011 national census (Australian Bureau of Statistics 2012). The raking was then accomplished through adjusting the sample counts in each of these poststrata to the known population marginal control totals through an iterative proportional fitting procedure. This essentially combines the data from the survey with the aggregate information on the population from sources that have greater precision and unbiasedness (for instance the census) to adjust for nonresponse bias.

An important caveat is that these adjustments for nonresponse are based on associations of social, demographic, and selected characteristics with a known model of the probability to respond to surveys (Brick 2013). Moreover, reliable external information is required to benchmark the sample characteristics to these known population characteristics. Evidence from the USA shows that it is necessary to adjust for differential nonresponse bias by telephone type and usage (Brick et al. 2006; Brick et al. 2011).

It has been demonstrated that the different sampling frames have different response profiles, and hence population-level information about the profiles of coverage on the landline and mobile frames will be important for the poststratification. Unfortunately, in Australia, as in most other countries, detailed demographic information by telephone-usage status is not available. However, we can estimate the proportion of the population by telephone status broadly into landline only, dual landline and mobile, and mobile only. This information on telephone-usage status from the national media regulatory body, in addition to population figures from the Census of Population and Housing, will be used in the nonresponse adjustment.

3. Dual-Frame Telephone Sampling in Australia

3.1. The Omnibus Surveys

3.1.1. Overview

Compared to the USA, the Australian dual-frame experience is relatively new, with the earliest survey conducted in 2010. That survey showed that 72% of respondents from the landline frame and 78% from the mobile frame used both types of telephone. The analysis of the survey data also found significant differences between mobile and landline telephone users. In particular, mobile users were more likely to be younger, reside in a capital city, be born outside Australia, and to be studying and living in group households (Pennay 2010). A larger omnibus survey using random digit dialling (RDD) that helped generate landline and mobile telephone sampling frames was administered in January 2012. This (first dual-frame omnibus survey) was designed to provide nationally representative statistics. The landline sample was proportionally stratified by geographical location across Australia. Selection probabilities for the landline sampling frame were derived through size quotas for capital city and noncapital city regions of the Australian states/territories. As there were no geographic identifiers available to stratify the mobile telephone sampling frame in Australia, a simple national random sampling frame was

devised. Data collection was via computer-assisted telephone interviewing (CATI) and the in-scope population for the dual-frame survey was Australian residents aged 18 years and over, who are contactable by either a landline or mobile telephone (see Appendix A for questionnaire items on phone status). A subsequent omnibus survey with a very similar design and execution was conducted in March 2013 to further explore emerging trends describing mobile-telephone-only individuals in Australia (see Pennay and Vickers 2012, 2013).

3.1.2. Survey Procedures, Call Results, and Analysis of Response

The analysis presented in this article is based on the data from the 2012 survey. This survey was chosen due to its temporal proximity to the 2011 Australian Census of Population and Housing – the source of many of our external benchmarks. The 2012 survey comprised 1,012 interviews completed via the landline sample frame with a response rate (AAPOR Response Rate 3, as defined in AAPOR 2011) of 22.2%, while the mobile sample frame yielded 1,002 completed interviews with a response rate of 12.7%. The average interview length was 19.8 minutes for both samples. A total of 76,342 calls were placed to achieve the total 2,014 completed interviews, which equates to an interview every 37.9 calls, but this average number differed significantly ($p < 0.001$) by frame with 27.7 for the landline frame and 48.2 for the mobile frame. The main reason for this was that roughly a third (32.5%) of all calls to mobile telephones resulted in voicemail outcomes, compared with 14.3% for landlines.

A number of strategies were used to maximise response and participation, including repeated callbacks to establish contact, the operation of a 1800 (free-to-call) number by the survey organisation, and leaving messages on answering machines/voicemail. Additionally, refusal conversion interviewing was used to identify the reasons for refusal and discretionary calls made to those identified as 'soft refusals', and the survey offered interviewing in languages other than English. Finally, an unlimited call cycle was used for the survey. This had the advantage of enabling interviews to be achieved with hard-to-reach individuals (a six-call cycle is typical), and ten percent (197 interviews) were achieved from the seventh or subsequent call attempt.

In terms of final call outcomes, there was a much higher proportion of telephone answering devices (answering machines/voicemail) for the mobile frame (20.0%) compared to the landline frame (7.4%). There was also a higher proportion of 'no answer' outcomes among the mobile frame (24.3%) compared with the landline frame (12.9%). There was a higher proportion of nonworking or disconnected numbers in the mobile frame (17.9%) compared with the landline frame (6.6%). The relatively high number of uncontactable numbers in the mobile frame is reflected in the much higher ratio of records used per interview in the mobile frame (11.6:1), compared with the landline frame (6.1:1).

The results, in Table 1, show differences in the age profiles and country of birth profiles for respondents in the two samples (with higher proportions of younger and overseas-born people included in the mobile sample). Since the mobile telephone interviews were conducted with little control over the geographical distribution since location information is not available, more interviews were conducted in the more populated regions. The mobile telephone sample contained a larger proportion of males, a larger proportion of those residing in rented group households, and were predominantly in the capital cities. Furthermore, the mobile sample had a younger age profile and was more likely to be

Table 1. Sample profile of the Dual-Frame Omnibus Survey by frame status, along with selected population estimates from the Australian 2011 Census.

		Landline Frame			Mobile Frame		
	ABS Population 2011 census	Total ($n = 1012$)	Landline only ($n = 174$)	Dual user ($n = 838$)	Total ($n = 1,002$)	Mobile only ($n = 295$)	Dual user ($n = 707$)
Selected Characteristics	%	%	%	%	%	%	%
Gender							
Male	49.3	36.6	35.1	36.9	52.3	56.9	50.4
Female	50.7	63.4	64.9	63.1	47.7	43.1	49.6
Age group (years)							
18–24	12.8	3.5	2.3	3.7	20.2	23.1	19.0
25–39	28.3	15.3	6.9	17.1	33.1	48.1	26.9
40–49	18.0	18.4	6.3	20.9	15.9	8.8	18.8
50–64	24.0	31.3	24.7	32.7	23.8	15.6	27.2
65+	16.9	31.5	59.8	25.7	7.1	4.4	8.2
Region							
Capital city	62.8	63.9	63.8	64.4	71.6	70.2	72.1
Other	37.2	36.1	36.2	35.6	28.4	29.8	27.9
Indigenous status							
A&TSI	2.5	1.3	1.4	0.6	1.4	3.7	0.4
Country of Birth							
Australia	69.8	74.1	75.3	73.9	63.0	61.0	63.8
Overseas	30.2	25.9	24.7	26.1	37.0	39.0	36.2

Table 1. Continued.

Selected Characteristics	*ABS Population 2011 census* %	Landline Frame			Mobile Frame		
		Total ($n = 1012$) %	Landline only ($n = 174$) %	Dual user ($n = 838$) %	Total ($n = 1,002$) %	Mobile only ($n = 295$) %	Dual user ($n = 707$) %
Home ownership							
Own home	32.1	48.6	60.9	46.1	21.1	8.1	26.4
Paying mortgage	34.9	31.1	15.5	34.4	29.5	19.7	33.7
Paying rent	29.6	13.9	16.1	13.5	40.3	67.5	29.0
Other	3.4	6.4	7.5	6.0	9.1	4.7	10.9
Living Arrangement							
Group household	4.1	3.1	3.1	2.9	14.6	26.8	9.5
Time in neighbourhood							
5 years or less	–	22.3	17.2	23.4	51.1	73.2	41.9
Employment status							
Employed	59.7	58.0	31.6	63.5	75.3	68.8	77.7
Educational attainment							
Bachelor's degree or higher	19.7	27.9	15.5	30.4	35.1	30.8	36.9

university educated and to be in employment. Additionally, the mobile telephone sample had a larger proportion of people born overseas in comparison with the landline sample. There were also differences in the characteristics of the dual telephone users depending on the sampling frame, with mobile-frame dual users more likely to be living in capital cities, group households and living in the neighbourhood for shorter periods, whereas landline-frame dual users were more likely to be Australian born and home owners. These differences have nontrivial implications for combining estimates from the landline and mobile frames.

3.2. Variables

A primary purpose of these surveys was to produce unbiased estimates for specific health characteristics and behaviours in the Australian population. The omnibus surveys collected data on health, attitudes, and behaviours, such as tobacco and alcohol consumption, experiences of discrimination, self-assessed medical health, and attitudes to the environment. In our evaluation of the performance of the various dual-frame estimators and weighting strategies, we considered a selection of the variables measuring these outcomes. Analyses of the data focused on:

1. investigating the link between transiency and telephone-usage status,
2. exploring the differences in social and health behaviour outcomes with telephone-usage status, and
3. examining the association between sedentary behaviour and telephone sampling.

To investigate the effect of transiency on responses from the landline and mobile telephone samples, we examined two variables measuring (i) group households and (ii) length of time in the neighbourhood (stayed less than five years or longer). We examined differences in social and health behaviour outcomes measured by smoking status, reports of being anxious or depressed, and belief in climate change. There is evidence to suggest a relationship between telephone-usage status and health-related behaviour. For example, it has been shown that the mobile-only population is likely to experience greater adverse health and behavioural outcomes (Blumberg and Luke 2014). Specifically, we considered smoking behaviour and incidence of depression and anxiety to investigate the hypothesis that the mobile-only population had poorer health outcomes (Lee et al. 2010; Thomée et al. 2011). In addition, we investigated attitudes to climate change, which are hypothesised to be related to age (Akter and Bennett 2011), with younger people more likely to believe that climate change is occurring and that humans are responsible for this. As mobile telephone usage is also associated with age, an estimate of the proportion of those who believed in climate change will be influenced by the choice of sampling frame and could be improved by combining data from both landline and mobile telephone sampling frames.

To investigate the association of telephone usage with sedentary behaviour, the recorded number of hours of television watched per day was analysed. According to the Australian Bureau of Statistics, the average amount of time spent watching television was just under three hours in 2008 (Australian Bureau of Statistics 2008). People who watch television for long periods of time are more likely to be those who are at home and this is related to accessibility by telephone. Long periods of television watching are also related

to a sedentary lifestyle and poorer health outcomes (Hu et al. 2003), and combining data from both landline and mobile telephone sampling frames is likely to improve the accuracy of population estimates for time spent watching television.

Finally, to compensate for nonresponse in the sample and differences between telephone-usage status, we compute poststratification weights from a selection of variables that are often associated with nonresponse and survey quality: demographic (age and gender), socioeconomic (educational attainment), country of birth (comparing Australia born to overseas born), and geographic location (based on state of residence).

4. Comparison of Dual-Frame Estimation Approaches

In Section 2 we described different approaches to estimating a population quantity using combined data from dual overlapping sampling frames. In this section we assess the performance of these estimators empirically when applied to the 2012 Australian dual-frame omnibus survey using information from the landline and mobile telephone samples. Three questions are addressed: Are there biases in the survey estimates if the mobile-only population is excluded? Are the biases in population estimates reduced when data from the two samples are combined? Finally, is there a preferred approach to weighting the combined samples?

4.1. Weighting the Dual-Frame Omnibus Surveys for Multiple Coverage and Nonresponse

To examine the biases in population estimates of the selected outcome variables, we firstly explored the differences in the response patterns of individuals by telephone sampling frame. In the presence of these differences, we then compared the population estimates using the various estimators described in Section 2 to combine sample data from the mobile telephone and landline frames. Lastly, as we have available information on age, sex, tenure status, and part-time employment from the most recent national census of Australia in 2011 (Australian Bureau of Statistics 2012), we compare these census estimates to estimates from the dual-frame survey to assess biases.

The five dual-frame estimators as described in Subsection 3.2 are computed for the analyses of the selected variables:

a. the screening estimators
b. the average estimator
c. the minimum-variance (Hartley) estimator
d. the pseudolikelihood (Skinner and Rao) estimator
e. the single-frame (Bankier) estimator.

The Skinner and Rao estimator (12) is similar to the Fuller and Burmeister estimator (10). But the Skinner and Rao estimator has the practical advantage of using the same weights for all variables and is approximately unbiased relative to the Fuller and Burmeister estimator. In addition, it is not possible to estimate the unknown population size N using the Fuller-Burmeister estimator because the estimation process involves the inversion of a singular matrix (Skinner and Rao 1996; Lohr and Rao 2006). Thus, we do not consider it further. For the poststratified estimators, the weights are calibrated through

a raking procedure so that the sample estimates correspond to known total population benchmarks across age, gender, educational attainment, country of birth, location, and telephone status. The survey estimates are examined before and after poststratification for the selected outcomes discussed in Subsection 3.3.

4.2. Estimation of the Composite Weights Using Omnibus Survey Data

The estimators identified above in (a) and (b) are straightforward to compute. For instance, the average estimator takes the simple average of the overlap between the two samples so that $\theta = 1/2$. Additionally, the screening estimators are also easy to compute because θ is fixed to be either zero or one. For the other composite estimators, some algebra is required. As discussed in the previous section, estimation of θ is based on optimising some functions. In the case of Hartley's estimator, the optimal value of θ_H is computed by minimising the variance of the estimator. As shown in Equation (8), the optimal choice of θ_H is a function of the variances and covariances of the estimated domain totals, and the consequence of this is that they differ for each response variable. The optimal choice of the compositing weight was found to be 0.50 for having a bachelor's degree, 0.51 for being anxious or depressed, 0.46 for TV watching, 0.55 for daily smoking status, 0.34 for belief in climate change, 0.70 for group-household living arrangement, and 0.38 for short lengths of neighbourhood residence. This shows that there are differences in the choice of compositing value depending on the measure of interest ranging from 0.34 to 0.70 due to the associated variability in the different outcome measures.

The remaining estimators in (c), (d), and (e) are more complex. The pseudomaximum-likelihood estimator (15) depends on knowing the number of landline and mobile telephones in Australia, and then finding the maximum-likelihood estimator of the overlapping population, $\hat{N}_{lm}^{PML}$ as the smallest root of (12). This is challenging as telephone-usage data is submitted by telecommunication service providers to the Australian Communications and Media Authority (ACMA), which is an independent statutory authority responsible for media regulation. In the USA, external information on telephone status is available through national surveys, such as the National Health Interview Survey (NHIS). This information allows for the adjustment of potential nonresponse biases associated with the different frames. Notably, there is area-level information available on mobile telephones in the USA. The situation is different in Australia. However, these figures from ACMA are the best available estimates of telephone coverage. We attempt to adjust for this lack of areal identifiers and any bias associated by including geographical information in the poststratification. Based on these figures, our computed estimate of the pseudomaximum-likelihood estimator is $\hat{\theta}_P = 0.59$ (see Appendix B for details of how this was calculated). For the single-frame estimator, it is assumed that the landline and mobile samples are two stratified samples from the 'same' frame and are used to compute individual weights that have been adjusted for overlap in the two samples.

4.3. Results

Table 2 shows the population estimates for each measure, separately for the landline and mobile telephone samples and both with and without poststratification. Standard errors

Table 2. Survey estimates for selected characteristics by sampling frame (with estimates of the standard errors): before and after poststratification raking.

	Landline sample	Landline sample	Mobile sample	Mobile sample
	Before poststratification	After poststratification	Before poststratification	After poststratification
Selected Characteristics	proportion	proportion	proportion	proportion
Bachelor's degree	0.30	0.20	0.36	0.20
	(0.012)	(0.016)	(0.015)	(0.012)
Anxiety or depression	0.15	0.15	0.21	0.22
	(0.012)	(0.015)	(0.013)	(0.016)
Hours of TV watched	0.10	0.10	0.08	0.11
	(0.010)	(0.012)	(0.009)	(0.014)
Daily smoking	0.12	0.14	0.18	0.19
	(0.011)	(0.015)	(0.012)	(0.015)
Belief in climate change	0.74	0.74	0.84	0.81
	(0.016)	(0.021)	(0.012)	(0.017)
Group household	0.04	0.05	0.15	0.10
	(0.008)	(0.013)	(0.011)	(0.010)
Short time in neighbourhood	0.44	0.27	0.51	0.42
	(0.014)	(0.021)	(0.016)	(0.019)

The standard errors appear in parentheses.

provide estimates of the precision, and these were computed using Taylor series approximation. However, qualitatively similar findings were observed when using the jackknife and bootstrap procedures (Lohr and Rao 2006; Lohr 2011). As previously shown, the estimates vary when using information from only the landline telephone frame or the mobile telephone frame. Since the estimators are estimated on independent samples, formal tests can be used to determine statistically significant differences. In particular, there are large and significant differences in the estimates for most of the variables analysed, including holding a bachelor's degree ($p = 0.04$), short length of time in neighbourhood ($p = 0.002$), having anxiety or depression ($p < 0.001$), smoking status ($p < 0.001$), living in a group household ($p < 0.001$), and long hours of TV watching ($p = 0.085$), before poststratification. This may be a consequence of differences in the individual profiles of mobile and landline phone users as well as differential nonresponse.

In general, the nonresponse-adjusted estimates should provide unbiased estimates of the population totals if we can construct a model that can completely explain the nonresponse mechanism. However, doing this usually changes the structure of the survey weights and can subsequently produce contrasting results for variables that are not included in the controls if these variables are related to the nonresponse mechanism (Deville and Särndal 1992). To demonstrate this, Table 2 shows the sample proportions for educational attainment measured by the qualification of having a bachelor's degree or not. The sample proportion of people with a bachelor's degree is higher before poststratification raking in both the landline telephone (29.7%) and mobile telephone (35.7%) samples. Approximately one third of respondents in the sample are educated to degree level and above, as compared to 20% in the Australian population, based on the 2011 Census. Thus raking has the effect of weighting the sample proportion to the national average of approximately 20%. It is well known that people who are highly educated are more likely to respond to telephone surveys, and as such this population subgroup is overrepresented in both the landline and mobile telephone samples. Similarly, raking adjusts the estimates for under- and overrepresentation of the sample in characteristics measured by the other control variables.

Tables 3(a) and 3(b) provide a comparison of the different compositing approaches that integrate the information from the mobile and landline telephone samples, specifically addressing frame overlap. These compositing approaches apply weights to the dual users in the mobile and landline samples. We compare the two screening approaches (for landline and mobile telephones), the average, minimum-variance (Hartley), pseudolikelihood (Skinner and Rao), and single-frame estimators. In order to examine the effect of nonresponse we apply raking (to population control totals), and compare how the estimates from each of the estimators differ with and without raking. Table 3(a) shows the estimates without poststratification raking, while Table 3(b) shows the results with poststratification raking.

The screening estimators provide a lower and upper bound for the average, Hartley, pseudolikelihood and single-frame estimators. This is not surprising, because the screening estimators use $\theta = 0$ or $\theta = 1$ while the other composite estimators use θ values that fall within the [0, 1] range. In general, the choice of the compositing factor leads to differences in estimates of the population characteristics of interest. However, these differences are not statistically significant. The average estimator produced almost the same results to the

Table 3(a). Estimates of selected characteristics (as a proportion) based on the different weighting methods (before poststratification raking).

Selected Characteristics	Average Estimator	Landline Screener	Mobile Screener	Single-Frame Estimator	Hartley Estimator	Skinner and Rao Estimator
Bachelor's degree	0.32	0.30	0.34	0.32	0.32	0.32
	(0.011)	(0.014)	(0.014)	(0.011)	(0.011)	(0.011)
Anxiety or depression	0.19	0.17	0.21	0.19	0.19	0.19
	(0.010)	(0.011)	(0.012)	(0.010)	(0.009)	(0.009)
Hours of TV watched	0.09	0.09	0.09	0.09	0.09	0.09
	(0.007)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)
Daily smoking	0.16	0.15	0.17	0.16	0.16	0.16
	(0.009)	(0.017)	(0.011)	(0.009)	(0.009)	(0.009)
Belief in climate change	0.79	0.76	0.82	0.79	0.79	0.79
	(0.010)	(0.013)	(0.012)	(0.010)	(0.011)	(0.010)
Group household	0.12	0.10	0.13	0.12	0.11	0.110
	(0.009)	(0.009)	(0.010)	(0.009)	(0.008)	(0.008)
Short time in neighbourhood	0.31	0.35	0.47	0.408	0.40	0.40
	(0.011)	(0.014)	(0.015)	(0.012)	(0.012)	(0.012)

The standard errors appear in parentheses.

Table 3(b). Estimates of selected characteristics (as a proportion) based on the different weighting methods (after poststratification raking).

Selected Characteristics	Average Estimator	Landline Screener	Mobile Screener	Single-Frame Estimator	Hartley Estimator	Skinner and Rao Estimator
Bachelor's degree	0.20	0.20	0.20	0.18	0.20	0.20
	(0.008)	(0.010)	(0.009)	(0.007)	(0.008)	(0.008)
Anxiety or depression	0.20	0.18	0.22	0.20	0.20	0.20
	(0.010)	(0.011)	(0.013)	(0.010)	(0.010)	(0.010)
Hours of TV watched	0.10	0.10	0.11	0.10	0.10	0.10
	(0.008)	(0.009)	(0.010)	(0.008)	(0.008)	(0.008)
Daily smoking	0.16	0.16	0.18	0.17	0.17	0.17
	(0.010)	(0.011)	(0.012)	(0.010)	(0.010)	(0.010)
Belief in climate change	0.78	0.76	0.80	0.77	0.78	0.78
	(0.011)	(0.014)	(0.013)	(0.011)	(0.011)	(0.011)
Group household	0.10	0.10	0.10	0.10	0.10	0.10
	(0.008)	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)
Short time in neighbourhood	0.38	0.36	0.41	0.38	0.38	0.38
	(0.012)	(0.014)	(0.015)	(0.012)	(0.012)	(0.012)

The standard errors appear in parentheses.

more complex compositing approaches, such as the pseudolikelihood and minimum-variance estimators. The results after the poststratification raking adjustment show that the average estimator, the Hartley estimator and the pseudomaximum-likelihood estimator all produce an equivalent estimate. The average estimator produced similar results to the more complex estimators. It is also worth noting that – with the exception of the variable time in neighbourhood – there is still little difference between the estimators, even before the raking to the benchmark control totals. Since the average estimator is theoretically inferior to the pseudomaximum likelihood, the consequence of this is that we may be lacking some control information for the benchmarking. In fact, the suggestion is that if population information on telephone status and the size of the mobile- and landline-frame populations was available (as it is in the USA and Europe), we would expect the average estimator to be radically inferior to the others.

These results are similar to the findings by Skinner and Rao (1996) and Lohr and Rao (2000). These authors have suggested that when there is little empirical evidence to select a preferred estimator, the decision to choose one estimator over the others should be made on theoretical and practical considerations, such as internal consistency and the availability of good-quality poststratification raking for benchmarking (Brick 2013; Arcos et al. 2014).

4.4. *Comparison of Composite Estimators With Census Figures*

To compare the performance of the dual-frame estimators with population quantities, we selected two variables for which existing census data was available, restricting the choice to those variables that were not used as benchmark controls (such as age, sex, country of birth, and region). Population estimates for part-time employment and housing-tenure status from the 2011 census were compared to estimates from the 2012 dual-frame omnibus survey. To ensure comparability to the survey, we restricted the census population to those aged 18 years and over.

We therefore examined how the estimates of part-time employment and tenure status varied first by doing the composite estimation (using the different estimators discussed) and then poststratifying to the control totals. We compared the precision and biases of the various weighting procedures, and the results are shown in Figure 3. If we have independent population information about the landline and mobile frames, such as demographic and geographical distributions, then this can be used to poststratify the mobile and landline samples to the control population controls. This is what is suggested by Brick et al. (2006) and Brick (2011). They suggest that by doing this, the fixed-weight average estimator is sufficient to reduce bias due to nonresponse, but they are dependent on knowing the response rates in the different domains. In Australia, currently, we do not have access to this level of telephone-usage population information, so it is of interest to see which estimator performs best.

Figure 3 shows estimates for part-time employment and tenure status (i.e., proportion of renters and home owners) using population census data, sample proportions and the dual-frame estimators. To provide an indication of how the dual-frame estimators (average, single-frame, Hartley, and Skinner-Rao estimators) perform, panels (c) and (d) present the results before and after poststratification for these estimators. It may be due to these figures

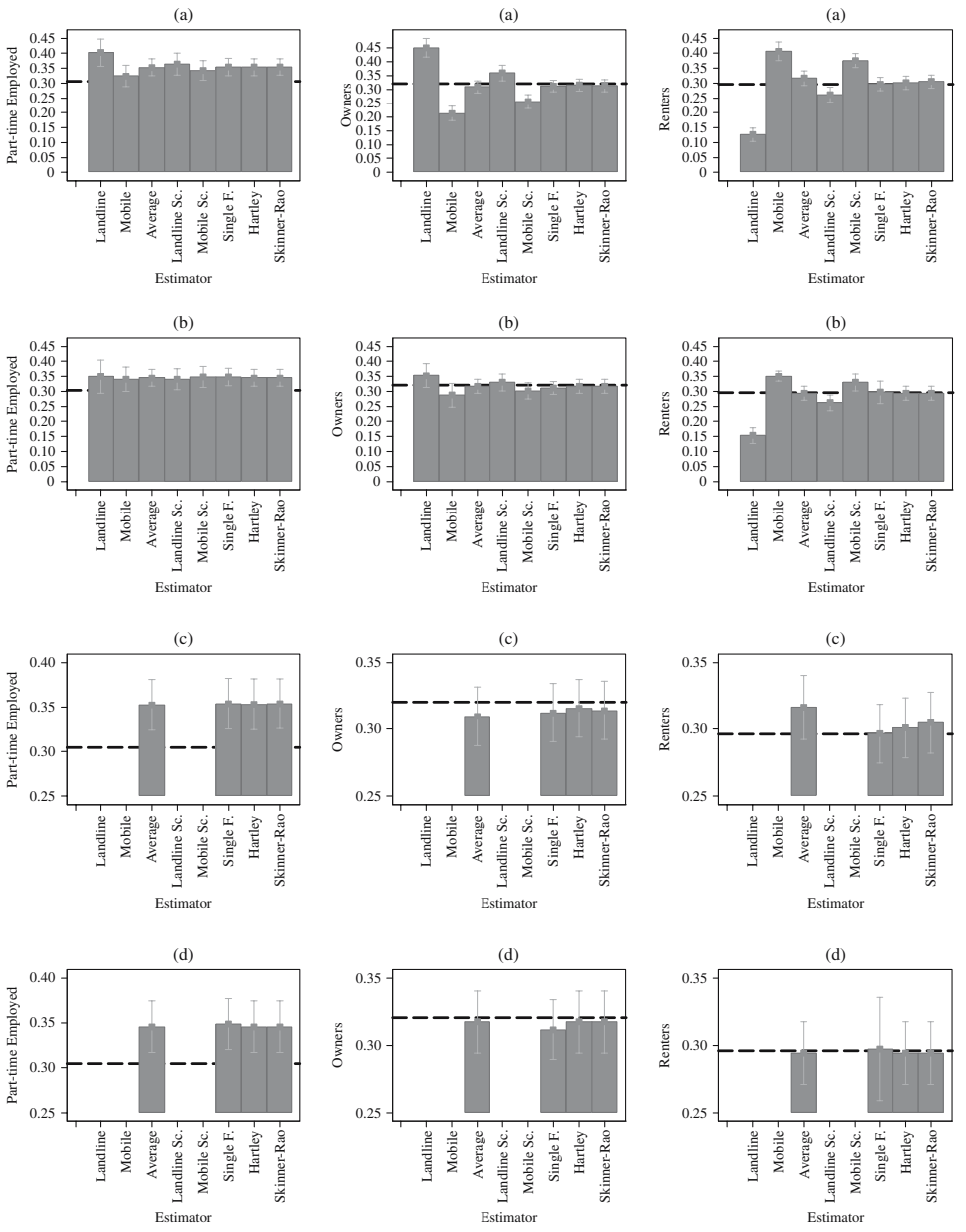


Fig. 3. Plots showing the estimates and the 95% confidence interval of selected characteristics for different dual-frame estimators; the left panel is for part-time employed, the middle panel is for owners, and the right panel is for renters. The census estimates are superimposed as the dotted line. For these plots (a) is the results ***before*** *poststratification raking, (b) is the results* ***after*** *poststratification raking; (c) gives the results of the four dual-frame estimators (average, single-frame, Hartley and Skinner-Rao)* ***before*** *poststratification raking; (d) gives the results of the four dual-frame estimators (average, single-frame, Hartley and Skinner-Rao)* ***after*** *poststratification raking.*

that information from only the mobile or the landline telephone samples leads to biased estimates of the population quantities. The landline sample overestimates the proportion of home owners and underestimates the proportion of renters. It also underestimates those who are in part-time employment. The mobile sample produces a reasonable estimate of the proportion of part-time employees, but the estimates of home owners and renters are both biased by more than ten percent. These biases are still present, although to a lesser degree, after poststratification raking to adjust for noncoverage and nonresponse. The compositing weighting approaches overestimate the part-time employed; however, they are less biased in regards to the proportion of home owners or renters. This is true for all composite estimators apart from the screening approaches. This result is intuitive because, as shown above, the screeners tend to overestimate (or underestimate) and provide upper and lower bounds of the estimates.

After poststratification raking, the average, Hartley, and pseudolikelihood estimators give almost exactly the same results, and provide estimates that are closer to the population figures overall. Although the single frame does better than the screeners, it fares worse in comparison to the others. It appears to do worse if the poststratification raking is carried out before the combining. This supports Lohr (2011), who stated that the single-frame estimator may not be as efficient as the Skinner and Rao or the Hartley estimators, although there may be efficiency gains in raking to population totals. In the main, the choice of the estimator depends on the differential response patterns on the sampling frames, as well as the availability of auxiliary, good-quality data on the frame nonresponse patterns that can be used for benchmarking purposes (Kennedy 2007; Brick et al. 2011). Raking, in general, has the effect of reducing the bias since it calibrates the estimates to population totals. Raking after combining has the effect of preserving the structure of the estimates so as to be closer to the population quantities, even in the cases when the variables are not used directly in the poststratification raking.

Population benchmarks should be used to adjust the sample so that the weighted sample aligns to the population and produces unbiased estimates. But it is not entirely evident which population characteristics should be used as benchmarks. The general recommendation is to calibrate to age, sex, education, geography, race/ethnicity, marital status, home tenure, and population density, as well as telephone status (AAPOR 2010). In Australia, regular official statistics about the telephone status of the population are not available from the census or other government sources, meaning that the benchmark data available may not be of the desired accuracy. Owing to the uncertainty surrounding the available information on telephone status, we also carried out investigations as to how the population estimates changed when the raking was carried out with and without telephone status as control. The results did not show any differences (not shown here).

While it is preferable to adjust for both nonresponse and multiple coverage, and to seek to compensate for any biases associated with the differential patterns of coverage and response in the two sampling frames, the literature is not clear about how to do this. In our empirical study, we have found that the raking adjustment does appear to have the most influence. Similar results are found in US dual-frame studies (e.g., Brick et al. 2006; Brick et al. 2011). The main difference, however, is that the USA has good telephone-status information from external data. Nonetheless, in the absence of 'accurate' nationally representative data on telephone usage and availability, it is possible to rely on the age,

sex, and other demographic and geographical characteristics that are related to telephone status. In our application to Australian dual-frame surveys, we have demonstrated that the uncertainty in the telephone-status information can be ameliorated by adjusting the samples to population characteristics from the census for differences in response profiles. As the mobile-only population increases, the situation will perhaps be different because of the potential nonsampling biases introduced.

5. Conclusion

Dual-frame telephone surveys will provide better coverage of the population than single-frame landline telephone surveys in most circumstances, due the absence of landline telephones in an increasing number of households and the exponential growth of households that are contactable only via mobile. We have shown that there are biases in survey estimates if the mobile-only population is excluded. These biases in the population estimates are significantly reduced when data from the mobile and landline sample are combined. However, although there are biases inherent in relying solely on a single frame, there are a number of issues that need to be addressed when proceeding to take a dual-frame survey approach. In the first instance, the combination of the information is not straightforward, as demonstrated by the various techniques available in the multiple-frame literature. Another aspect of estimation is the need to adjust for nonresponse through poststratification raking. Nonetheless, decisions about how to apply the poststratification raking are contingent on the availability of good-quality population-level information on sample characteristics that affect nonresponse and telephone status. As well as the current population benchmarks, the calculation of the nonresponse adjustments is contingent on phone-use benchmark information, which is routinely available in other countries, through nationally representative surveys that collect information on telephone status and usage. In Australia, there is no comparable survey that collects this type of information, and the influence this has on the computation of the compositing and nonresponse weights is unclear (Barr et al. 2014). This highlights the need for better population information on telephone usage in Australia, especially as the mobile-only population reaches the levels of the US.

Our results have demonstrated that the choice of an optimal dual-frame estimation approach depends on a number of factors. The first is the availability of good-quality information on telephone status. The second is the availability of raking information to account for the differential patterns of nonresponse. Essentially, our empirical results show that there is an interplay between the choice of dual-frame estimator and how to apply poststratification raking. In the absence of good-quality information on nonresponse, there is no preferable approach to weighting the combined samples. This supports Arcos et al. (2014), who showed that for the situation when accurate information on the mobile and landline populations is present, the single-frame and pseudomaximum-likelihood estimators give internally consistent results and are preferable. Although this was previously investigated by Lohr and Rao (2000, 2006), Brick et al. (2011) and Lohr (2011), we have now demonstrated this for the Australian dual-frame context. In particular, we have shown that the average estimator performs similarly to the more complicated estimators. We have shown that there are different conclusions as regards the best choice of dual-frame estimator because of the lack of independent information

available on the landline- and mobile-frame population totals for compensating for nonresponse.

The need to properly understand the behaviour of the different compositing approaches in the presence of uncertain poststratification benchmarking totals is an area for future research, possibly through a rigorous Monte Carlo simulation study in which the dual-frame estimators are compared across different features of the data. Due to the differential nonresponse that exists in the mobile and landline samples and the fact that – despite the best intentions of survey practitioners – there will be noncoverage in the samples, dual-frame surveys need to adjust for nonresponse through poststratification. For this to work effectively, not only is information on the mobile and landline population totals needed, but detailed information on the differential nonresponse profiles of mobile and landline frames is also required.

Appendix A – Questionnaire Items

Introductory Questions for Mobile Sample

Intro1: Good morning/afternoon/evening. My name is < SAY NAME >. I am calling from the Social Research Centre. The reason I'm calling is to see if you can help out with an important academic survey about health and wellbeing issues. To be eligible you need to be aged 18 years or over. The interview will take around 15 minutes depending on your answers. Would you be willing to do the survey at this time?

Intro2: May I just check whether or not it is safe to take this call at the moment. If not, I am happy to call you back when it is more convenient for you.

Introductory Questions for Landline Sample

Intro1: Good morning/afternoon/evening. My name is < SAY NAME >. I am calling from the Social Research Centre. The reason I'm calling is to see if you can help out with an important academic survey about health and wellbeing issues. The results will be used to improve the quality of population research in Australia. The interview will take around 15 minutes depending on your answers. Would you be willing to do the survey at this time?

Intro2: For this research we'd like to speak to the person in the household aged 18 years and over who had the most recent birthday – will that be you? IF NECESSARY – This is just a way of randomising who we talk to in the household.

Telephone Status Questions

SMP1: To start with I have a question or two about your use of telephone services. Is there at least one working fixed line telephone inside your home that is used for making and receiving calls?

1. Yes
2. No
3. (Don't know)
4. (Refused)

*(IF LANDLINE SAMPLE OR SMP1 = 1 (MOBILE SAMPLE WITH LANDLINE), CONTINUE, ELSE GO TO SMP3)

SMP2: How many residential phone numbers do you have in your household not including lines dedicated to faxes, modems or business phone numbers? Do not include mobile phones.

INTEREVIEWER NOTE: If needed explain as how many individual landline numbers are there at your house that you use to make and receive calls?

1. Number of lines given (Specify _____) RECORD WHOLE NUMBER (ALLOWABLE RANGE 1 TO 15)
2. No
3. (Don't know/Not stated)

SMP3: Do you have a working mobile phone?

1. Yes
2. No
3. (Don't know)
4. (Refused)

Appendix B – Derivation of the Pseudomaximum-likelihood Estimator

Recall that n_L and n_M are 1012 and 1002, and there are 838 landline dual users $\left(n_{lm}^L\right)$, and 707 mobile dual users $\left(n_{lm}^M\right)$. N_L and N_M are the population-level number of landlines and mobile phones in Australia (which is not known with accuracy). We therefore undertook a procedure to estimate these based on information provided by the Australian Bureau of Statistics (ABS) and the Australian Communications and Media Authority (ACMA). According to the ABS, there are 8,498,668 private dwellings in Australia, and ACMA estimates that 81% of people aged 18 years and over live in households with a landline connection (ACMA 2011, p. 8). As households can have more than one landline connection we apply an adjustment factor of 1.05. This gives the estimated number of residential telephone numbers in Australia as 7228,117 ($= 8{,}498{,}668 \times 0.81 \times 1.05$).

Similarly, for the estimated number of people with mobile telephones, we first start with the proportion of the adult population with a mobile telephone in Australia as 89% (ACMA 2011, p. 13). There are 17,229,344 people aged 18 years and over according to the census. This gives the estimated number of mobile population as 15,334,107 (($= 0.89 \times 17{,}229{,}344$).

Finally, landlines are household devices whereas mobiles are personal (individual) level, so to make it comparable we use the fact that there are 2.2 adults per household in Australia on average (Australian Bureau of Statistics 2012), and adjust the landline population from household level to individual level.

Since the population of dual users is not known, we use a result from Fuller and Burmeister (1972) to estimate dual users, such that $\hat{N}_{lm}^L = \frac{N_L}{n_L} n_{lm}^L = 13{,}167{,}743.58$ for the landline dual–user population, and $\hat{N}_{lm}^M = \frac{N_M}{n_M} n_{lm}^M = 10{,}819{,}574.50$ for the mobile dual–user population.

The pseudomaximum-likelihood estimator of the dual-user population, $\hat{N}_{lm}^{PML}$, is the smallest root of

$$(n_L + n_M)x^2 - \left(n_L N_M + n_M N_L + n_{lm}^L N_L + n_{lm}^M N_M\right)x + \left(n_{lm}^L + n_{lm}^M\right)N_L N_M = 0.$$

After some algebra, this is found to be 11,900,766.15.

This estimate of the overlapping population seems reasonable and roughly equal to the average of the two estimates from the landline and mobile samples.

Finally to obtain the pseudomaximum-likelihood estimator of the compositing weight, we use the expression in Skinner and Rao (1996)

$$\hat{\theta}_P = \frac{\hat{N}_l N_M \widehat{Var}\left(\hat{N}_{lm}^L\right)}{\hat{N}_l N_M \widehat{Var}\left(\hat{N}_{lm}^L\right) + \hat{N}_m N_L \widehat{Var}\left(\hat{N}_{lm}^M\right)}$$

where $\hat{N}_{lm}^{PML} = 11{,}900{,}766.15$, $\hat{N}_l \approx N_L - \hat{N}_{lm}^{PML}$ and $\hat{N}_m \approx N_L - \hat{N}_{lm}^{PML}$.

The pseudolikelihood-compositing weight θ_P is dependent on the variance terms $Var\left(\hat{N}_{lm}^L\right)$ and $Var\left(\hat{N}_{lm}^M\right)$, which are difficult to compute. However, we can use the fact that $Var\left(\hat{N}_{lm}^L\right) = \left(\frac{N_L}{n_L}\right)^2 Var\left(n_{lm}^L\right)$ and $Var\left(\hat{N}_{lm}^M\right) = \left(\frac{N_M}{n_M}\right)^2 Var\left(n_{lm}^M\right)$. However, it still remains to estimate $Var\left(n_{lm}^L\right)$ and $Var\left(n_{lm}^M\right)$, which are complex functions of the inclusion probabilities. Therefore, following on from Lohr and Rao (2000), $\widehat{Var}\left(n_{lm}^L\right)$ and $\widehat{Var}\left(n_{lm}^M\right)$ are estimated from the data. Substituting these values, the pseudomaximum-likelihood compositing weight is $\hat{\theta}_P = 0.59$.

6. References

Akter, S. and J. Bennett. 2011. "Household Perceptions of Climate Change and Preferences for Mitigation Action: The Case of the Carbon Pollution Reduction Scheme in Australia." *Climatic Change* 109: 417–436. Doi: http://dx.doi.org/10.1007/s10584-011-0034-8.

Arcos, A., M.M. Rueda, M. Trujillo, and D. Molina. 2014. "Review of Estimation Methods for Landline and Cell Phone Surveys." *Sociological Methods and Research* 44: 458–485. Doi: http://dx.doi.org/10.1177/0049124114546904.

American Association for Public Opinion Research (AAPOR). 2010. "Cell Phone Task Force Report. New Considerations for Survey Researchers When Planning and Conducting RDD Telephone Surveys in the US With Respondents Reached via Cell Phone Numbers." AAPOR. Available at: http://www.aapor.org/AAPORKentico/Education-Resources/Reports/Cell-Phone-Task-Force-Report.aspx. (accessed 25 November 2015).

American Association for Public Opinion Research (AAPOR). 2011. "Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys 7th Edition." AAPOR. Available at: http://www.esomar.org/knowledge-and-standards/research-resources/aapor-standard-definitions.php. (accessed 25 November 2015).

Australian Bureau of Statistics (ABS). 2008. "How Australians Use Their Time: Time spent on cultural activities. ABS cat.no.4153.0." Canberra: Australian Bureau of

Statistics. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/mf/4153.0 (accessed 25 November 2015).

Australian Bureau of Statistics (ABS). 2012. "Basic Community Profiles, Census 2011. Canberra: Australian Bureau of Statistics." Available at: http://www.abs.gov.au/websitedbs/censushome.nsf/home/communityprofiles. (accessed 25 November 2015).

Australian Bureau of Statistics (ABS). 2014. "Australian Health Survey: Users' Guide, 2011-13." ABS cat.no. 4363.0.55.001. Canberra: Australian Bureau of Statistics. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/mf/4363.0.55.001 (accessed 25 November 2015).

Australian Communications and Media Authority (ACMA). 2006, 2011, 2014, 2015. Communications Report, 2005–2006, 2010–2011, 2013–2014 & 2014–2015 Series "Converging Communications Channels: Preferences and Behaviour of Australian Communications Users." ACMA: Canberra. Available at: www.acma.gov.au/Comms-Report. (accessed 25 November 2015).

Bankier, M.D. 1986. "Estimators Based on Several Stratified Samples With Applications to Multiple Finite Surveys." *Journal of the American Statistical Association* 81: 1074–1079. Doi: http://dx.doi.org/10.1080/01621459.1986.10478376.

Barnes, G.M., J.H. Hoffman, M.O. Tidwell, and J.W. Welte. 2015. "Gambling and Substance Use: Co-Occurrence Among Adults in a Recent General Population Study in the United States." *International Gambling Studies* 15: 55–71. Doi: http://dx.doi.org/10.1080/14459795.2014.990396.

Barr, M.L., J.J. van Ritten, D.G. Steel, and S.V. Thackway. 2012. "Inclusion of Mobile Phone Numbers Into an Ongoing Population Health Survey in New South Wales, Australia: Design, Methods, Call Outcomes, Costs and Sample Representativeness." *BMC Medical Research Methodology* 12: 177–185. Doi: http://dx.doi.org/10.1186/1471-2288-12-177.

Barr, M.L., R.A. Ferguson, P.J. Hughes, and D.G. Steel. 2014. "Developing a Weighting Strategy to Include Mobile Telephone Numbers Into an Ongoing Population Health Survey Using an Overlapping Dual Frame Design With Limited Benchmark Information." *BMC Medical Research Methodology* 14: 102–112. Doi: http://dx.doi.org/10.1186/1471-2288-14-102.

Béland, Y. 2002. "Canadian Community Health Survey – Methodological Overview." *Health Reports - Statistics Canada* 13: 9–14.

Blumberg, S.J. and J.V. Luke. 2007. "Coverage Bias in Traditional Telephone Surveys of Low-Income and Young Adults." *Public Opinion Quarterly* 71: 734–749. Doi: http://dx.doi.org/10.1093/poq/nfm047.

Blumberg, S.J. and J.V. Luke. 2014. "Wireless Substitution: Early Release Estimates from the National Health Interview Survey, January-June 2014." National Center for Health Statistics. Available at: http://www.cdc.gov/nchs/nhis/releases.htm (accessed 25 November 2015).

Blumberg, S.J. and J.V. Luke. 2015. "Wireless Substitution: Early Release Estimates from the National Health Interview Survey, July-December 2014." National Center for Health Statistics. Available at: http://www.cdc.gov/nchs/nhis/releases.htm (accessed 25 November 2015).

Brick, J.M. 2011. "The future of survey sampling." *Public Opinion Quarterly* 75: 872–888. Doi: http://dx.doi.org/10.1093/poq/nfr045.

Brick, J.M. 2013. "Unit Nonresponse and Weighting Adjustments: a Critical Review." *Journal of Official Statistics* 29: 329–353. Doi: http://dx.doi.org/10.2478/jos-2013-0026.

Brick, J.M., I.F. Cervantes, S. Lee, and G. Norman. 2011. "Nonsampling Errors in Dual Frame Telephone Surveys." *Survey Methodology* 37: 1–12.

Brick, J.M., S. Dipko, S. Presser, and C. Tucker. 2006. "Nonresponse Bias in a Dual Frame Sample of Cell and Landline Numbers." *Public Opinion Quarterly* 70: 780–793. Doi: http://dx.doi.org/10.1093/poq/nfl031.

Busse, B. and M. Fuchs. 2012. "The Components of Landline Telephone Survey Coverage Bias. The Relative Importance of No-Phone and Mobile-Only Populations." *Quality & Quantity* 46: 1209–1225. Doi: http://dx.doi.org/10.1007/s11135-011-9431-3.

Callegaro, M. and T. Possio. 2004. "Mobile Telephone Growth and Coverage Error in Telephone Surveys." *Polis: Research and Studies in Italian Society and Politics* 18: 477–506.

Deville, J. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382.

Flores Cervantes, I. and G. Kalton. 2008. "Methods for Sampling Rare Populations in Telephone Surveys." In *Advances in Telephone Survey Methodology*, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P. Lavrakas, M.W. Link, and R.L. Sangster, 113–132. Hoboken, NJ: Wiley.

Fuller, W.A. and L.F. Burmeister. 1972. "Estimators for Samples Selected From Two Overlapping Frames." In *Proceedings of the Social Statistics Section of American Statistical Association*, August 14–17, 245–249. Alexandria, VA: American Statistical Association.

Gong, G. and F.J. Samaniego. 1981. "Pseudo Maximum Likelihood Estimation: Theory and Applications." *The Annals of Statistics* 9: 861–869. Doi: http://dx.doi.org/10.1214/aos/1176345526.

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Non-Response Rates on Non-Response Bias." *Public Opinion Quarterly* 72: 167–189. Doi: http://dx.doi.org/10.1093/poq/nfn011.

Hartley, H.O. 1962. "Multiple Frame Surveys." In *Proceedings of the Social Statistics Section of American Statistical Association, date of conference*, 203–206. Alexandria, VA: American Statistical Association.

Hartley, H.O. 1974. "Multiple Frame Methodology and Selected Applications." *Sankhya C* 36: 99–118.

Holt, D. and T.M.F. Smith. 1979. "Post-Stratification." *Journal of the Royal Statistical Society, Series A* 142: 33–46.

Hu, S.S., L. Balluz, M.P. Battaglia, and M.R. Frankel. 2011. "Improving Public Health Surveillance Using a Dual Frame Survey of Landline and Cell Phone Numbers." *American Journal of Epidemiology* 173: 703–711. Doi: http://dx.doi.org/10.1093/aje/kwq442.

Hu, F.B., T.Y. Li, G.A. Colditz, W.C. Willett, and J.E. Manson. 2003. "Television Watching and Other Sedentary Behaviours in Relation to Risk of Obesity and Type 2

Diabetes Mellitus in Women." *Journal of the American Medical Association (JAMA)* 289: 1785–1791. Doi: http://dx.doi.org/10.1001/jama.289.14.1785.

Iachan, R. and M.L. Dennis. 1993. "A Multiple Frame Approach to Sampling the Homeless and Transient Population." *Journal of Official Statistics* 9: 747–764.

Kalton, G. and D.W. Anderson. 1986. "Sampling Rare Populations." *Journal of the Royal Statistical Society, Series A* 149: 65–82. Doi: http://dx.doi.org/10.2307/2981886.

Keeter, S., C. Miller, A. Kohut, R.M. Groves, and S. Presser. 2000. "Consequences of Reducing Non-Responses in a National Telephone Survey." *Public Opinion Quarterly* 64: 125–148. Doi: http://dx.doi.org/10.1086/317759.

Keeter, S., C. Kennedy, A. Clark, T. Tompson, and M. Mokrzycki. 2007. "What's Missing From National Landline RDD surveys? The Impact of the Growing Cell-Only Population." *Public Opinion Quarterly* 71: 772–792. Doi: http://dx.doi.org/10.1093/poq/nfm053.

Keeter, S., S.C. Kennedy, M. Dimock, J. Best, and P. Craighill. 2006. "Gauging the Impact of Growing Non-Response on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70: 759–779. Doi: http://dx.doi.org/10.1093/poq/nfl035.

Kennedy, C. 2007. "Evaluating the Effects of Screening for Telephone Service in Dual Frame RDD surveys." *Public Opinion Quarterly* 71: 750–771. Doi: http://dx.doi.org/10.1093/poq/nfm050.

Kuusela, V., M. Callegaro, and V. Vehovar. 2008. "The Influence of Mobile Telephones on Telephone Surveys." In *Advances in Telephone Survey Methodology*, edited by J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link, and R.L. Sangster, 87–112. Doi: http://dx.doi.org/10.1002/9780470173404.ch4.

Lee, S., J.M. Brick, E.R. Brown, and D. Grant. 2010. "Growing Cell-Phone Population and Noncoverage Bias in Traditional Random Digit Dial Telephone Health Surveys." *Health Services Research* 45: 1121–1139. Doi: http://dx.doi.org/10.1111/j.1475-6773.2010.01120.x.

Little, R.J.A. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88: 1001–1012.

Livingston, M., P. Dietze, J. Ferris, D. Pennay, L. Hayes, and S. Lenton. 2013. "Surveying Alcohol and Other Drug Use Through Telephone Sampling: A Comparison of Landline and Mobile Phone Samples." *BMC Medical Research Methodology* 13: 41. Doi: http://dx.doi.org/10.1186/1471-2288-13-41.

Lohr, S. 2009. "Multiple Frame Surveys." In *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, 71–88. Amsterdam: North Holland.

Lohr, S.L. 2011. "Alternative Survey Sample Designs: Sampling With Multiple Overlapping Frames." *Survey Methodology* 37: 197–213.

Lohr, S. and J. Rao. 2000. "Inference From Dual Frame Surveys." *Journal of the American Statistical Association* 95: 271–280. Doi: http://dx.doi.org/10.1080/01621459.2000.10473920.

Lohr, S. and J. Rao. 2006. "Estimation in Multiple-Frame Surveys." *Journal of the American Statistical Association* 101: 1019–1030. Doi: http://dx.doi.org10.1198/016214506000000195.

Lopez, M.H. and A. Gonzalez-Barrera. 2013. "Inside the 2012 Latino Electorate." Washington, D.C.: Pew Research Center's Hispanic Trends Project. Available at:

http://www.pewhispanic.org/2013/06/03/inside-the-2012-latino-electorate. (accessed 25 November 2015).

Mecatti, F. 2007. "A Single Frame Multiplicity Estimator for Multiple Frame Surveys." *Survey Methodology* 33: 151–157.

Mohorko, A., E. de Leeuw, and J. Hox. 2013. "Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage Across Countries and Over Time." *Survey Methods: Insights from the Field (SMIF) 2013*, 1–13. Doi: http://dx.doi.org/10.13094/SMIF-2013-00002.

Pennay, D.W. 2010. "Profiling the 'Mobile Only' Population. Results From a Dual-Frame Telephone Survey Using a Landline and Mobile Phone Sample Frame." In *Proceedings of the Australian Consortium for Social and Political Research Incorporated (ACSPRI) Social Science Methodology Conference,* 1–3 December, Sydney, Australia. Available at: http://www.srcentre.com.au/docs/publications/dual_frame-survey_acspri-conference-paper_finalv2.pdf?sfvrsn=0 (accessed 25 November 2015).

Pennay, D.W. and N. Vickers. 2012. "Dual-Frame Omnibus Survey." Technical and Methodological Summary Report. Social Research Centre Pty Ltd, Melbourne, Australia. Available at: http://www.srcentre.com.au/docs/event-workshop-july-2012/dual-frame-omnibus-technical-report-(pennay).pdf?sfvrsn=2 (accessed 25 November 2015).

Pennay, D.W. and N. Vickers. 2013. "Second Dual-Frame Omnibus Survey." Technical and Methodological Summary Report. Social Research Centre Pty Ltd, Melbourne, Australia. Available at: http://www.srcentre.com.au/docs/publications/full-report-here.pdf?sfvrsn=0 (accessed 27 November 2015).

Pew Research Centre. 2012. "Assessing the Representativeness of Public Opinion Surveys." Available at: http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/ (accessed 25 November 2015).

Skinner, C.J. 1991. "On Efficiency of Raking Ratio Estimation for Multiple Frame Surveys." *Journal of the American Statistical Association* 86: 779–784.

Skinner, C.J. and J.N.K. Rao. 1996. "Estimation in Dual Frame Surveys with Complex Designs." *Journal of the American Statistical Association* 91: 349–356.

Steeh, C. 2008. "Telephone Surveys." In *International Handbook of Survey Methodology*, edited by E.D. de Leuuw, J.J. Hox, and D.A. Dillman, 221–238. New York: Routledge.

Thomée, S., A. Härenstam, and M. Hagberg. 2011. "Mobile Phone Use and Stress, Sleep Disturbances, and Symptoms of Depression Among Young Adults - A Prospective Cohort Study." *BMC Public Health* 11: 66. Doi: http://dx.doi.org/10.1186/1471-2458-11-66.

Tucker, C. and J.M. Brick. 2007. "Household Telephone Service and Usage Patterns in the United States in 2004: Implications for Telephone Samples." *Public Opinion Quarterly* 71: 3–22. Doi: http://dx.doi.org/10.1093/poq/nfl047.

Vicente, P. and E. Reis, E. 2009. The "Mobile Only Population in Portugal and its Impact in a Dual Frame Telephone Survey." *Survey Research Methods* 3: 105–111.

Received January 2015
Revised December 2015
Accepted January 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 579–600, http://dx.doi.org/10.1515/JOS-2016-0030

Using Data Mining to Predict the Occurrence of Respondent Retrieval Strategies in Calendar Interviewing: The Quality of Retrospective Reports

Robert F. Belli[1], *L. Dee Miller*[2], *Tarek Al Baghal*[3], *and Leen-Kiat Soh*[4]

Determining which verbal behaviors of interviewers and respondents are dependent on one another is a complex problem that can be facilitated via data-mining approaches. Data are derived from the interviews of 153 respondents of the Panel Study of Income Dynamics (PSID) who were interviewed about their life-course histories. Behavioral sequences of interviewer-respondent interactions that were most predictive of respondents spontaneously using parallel, timing, duration, and sequential retrieval strategies in their generation of answers were examined. We also examined which behavioral sequences were predictive of retrospective reporting data quality as shown by correspondence between calendar responses with responses collected in prior waves of the PSID. The verbal behaviors of immediately preceding interviewer and respondent turns of speech were assessed in terms of their co-occurrence with each respondent retrieval strategy. Interviewers' use of parallel probes is associated with poorer data quality, whereas interviewers' use of timing and duration probes, especially in tandem, is associated with better data quality. Respondents' use of timing and duration strategies is also associated with better data quality and both strategies are facilitated by interviewer timing probes. Data mining alongside regression techniques is valuable to examine which interviewer-respondent interactions will benefit data quality.

Key words: Calendar interviewing; data mining; interviewing; memory aids.

1. Introduction

In the collection of retrospective reports, calendar interviewing methods have reliably led to better data quality in comparison to conventional standardized methods, at times with only limited costs in which increases in interviewing and programming time are negligible or minimal at most (for reviews see Belli 2014; Belli and Callegaro 2009; Glasner and van der Vaart 2009). In calendar interviews, instead of having questions written in advance as in conventional standardized interviewing, interviewers develop queries to satisfy questionnaire objectives that are largely visually displayed by timelines within various domains (see, for example, Balán et al. 1969; Freedman et al. 1988). Each timeline is constructed with

[1] University of Nebraska, Department of Psychology, Lincoln, NE 68588-0308, U.S.A. Email: bbelli2@unl.edu
[2] University of Nebraska, 2343 Stone Creek Loop South, Lincoln, NE 68512, U.S.A. Email: ldeemiller@gmail.com
[3] University of Essex, ISER, Colchester, UK CO4 3SQ. Email: talbaghal@gmail.com
[4] University of Nebraska, 122E Avery Hall, Lincoln, NE 68588-0115, U.S.A. Email: lksoh@cse.unl.edu
Acknowledgments: This article is based upon work supported by the National Science Foundation under Grant No. 1132015. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

a specified unit of analysis (e.g., week, month, or year) and reference period (e.g., one year, ten years, or from birth to the present), and they are aligned with calendar time depending on their unit of analysis. Within each timeline, queries by interviewers will seek to get respondents to report periods of stability and points of transition, such as being employed with one employer for a period of time and then transitioning to another employer at another period of time. A domain represents a topic of interest, such as information on residential, partnering, parenting, labor, and health histories, and each domain may consist of one to several timelines. For example, when collecting labor histories, separate timelines may be devoted to employment and unemployment, respectively.

The improvements in data quality with calendar interviewing methods have been examined both theoretically and empirically within the context of the structure of autobiographical memory (Belli 1998; Belli et al. 2007; Bilgen and Belli 2010). Specifically, calendar methods have been shown to encourage the use of verbal retrieval behaviors in both interviewers and respondents that, in comparison to conventional questionnaires, are associated with better data quality for retrospective reports of life-course labor histories (Belli et al. 2007), especially for respondents who have experienced complicated pasts (Belli et al. 2013). Further, these behaviors align with the structure of autobiographical memory (Belli et al. 2004; Bilgen and Belli 2010).

Although calendar methods have produced encouraging results, as noted by Belli et al. (2013), these results are limited because they do not examine the communicative interactions between interviewers and respondents directly. In this article, to overcome this limitation, we examine those series of communicative interactions that are most likely to lead to respondents' use of retrieval strategies. We focus on respondent retrieval strategies as the outcome of interviewer-respondent interactions because we believe that their use is tied most directly to the successful remembering of past events. As for interviewer retrieval probing, our expectation is that the use of these probes will promote the use of retrieval strategies on the part of respondents, a result that we expect to confirm via our interactional analyses.

In terms of the structure of autobiographical memory, we examine those retrieval strategies *consisting of parallel and sequential cues* (Belli 1998; Belli and Callegaro 2009; Belli et al. 2013). With *parallel* retrieval strategies, respondents cue themselves by remembering a contemporaneous event from a different life domain as an apparent attempt to more fully reconstruct the past. An example of parallel cuing would occur if a respondent is asked about when a job ended, and they spontaneously remember the birth of a child when answering this query. With *sequential* retrieval strategies, respondents seek to order what happened earlier and later in time within the same domain by seeking to remember the time location of the beginning and ending of events, the duration of events, and/or what event occurred earlier or later. An example of sequential retrieval would occur if the respondent remembered that working as a librarian at a university immediately followed working as an office worker for a private company.

Hence, we are concerned with two main issues. First, we seek to determine which series of verbal exchanges between interviewers and respondents in calendar interviews are more likely to lead to respondent retrieval strategies. Earlier research examining interviewer and respondent verbal exchanges with conventional questionnaires has demonstrated the challenges of these approaches. Brenner (1982) was interested in determining via tree

structures those combinations of behaviors that followed earlier specific behaviors, such as what follows from interviewers asking questions as written versus when interviewers altered questions. Although such modeling could be applied in reverse, so that the tree structure from a later behavior could be propagated forward in time, the hand calculation of these tree structures is cumbersome. In addition, Brenner (1982) concentrated only on the occurrence of behaviors; we are also interested in determining whether the nonoccurrence of behaviors is similarly predictive of a final respondent retrieval. Adding nonoccurrence leads to further computational challenges. The only study we found that sought to examine which behaviors occurred earlier focused only on single behaviors (Dijkstra and Ongena 2006), and not on different combinations of earlier behaviors. In order to identify those series of behaviors that are predictive of the occurrence of respondent retrievals, while accounting for the occurrence and nonoccurrence of those behaviors contained in these series, we use data-mining techniques that have been developed in the field of computer science.

Second, having identified different behavioral series that are predictive of the presence of respondent retrieval behaviors, we conduct analyses to determine which, if any, of these series are predictive of retrospective reports of better data quality. Although, as noted above, Belli et al. (2013) have demonstrated that respondent retrieval strategies are associated with better data quality, their research used a confirmatory factor-analysis approach to create a single latent measure of respondent retrieval from several behaviors. This work extends that in Belli et al. (2013) by providing more focused interactional analyses. Specifically, by identifying which behavior series are predictive of better data quality and which are not, we show that behavioral interactions between interviewers and respondents lead to respondent retrieval strategies that vary in their effectiveness.

2. Data-Collection Method

Response data were collected from 313 Panel Study of Income Dynamics respondents of 45 years of age and older in 2002 (93% cooperation rate, AAPOR standard definition 1). Respondents were interviewed with a computer-assisted telephone interviewing (CATI) calendar instrument that asked for reports on residence, relationship, labor (employment and unemployment), and health lifetime histories. 297 interviews were audio recorded with respondent permission, with 291 audible tapes transcribed. Greater detail on the calendar CATI data-collection methods can be found in Belli et al. (2007) and Belli et al. (2013).

A random sample of 165 interviews was behavior coded with a scheme that comprised 30 interviewer and 29 respondent verbal behaviors. Behaviors were identified within turns of speech, a turn being defined as a transcribed uninterrupted utterance by either the interviewer or respondent. Greater detail on the behavior-coding methods, the reliability among coders, and the verbal behaviors that were identified can be found in Bilgen and Belli (2010).

3. Data Analyses and Results

3.1. Data-mining Algorithm

Our overall aim is to implement a data-mining algorithm able to isolate different series of verbal behaviors immediately preceding three turns of speech – an interviewer turn,

a respondent turn, and another interviewer turn – to those respondent turns that contained one of four respondent retrieval strategies. We selected three preceding turns of speech as an attempt to come to some compromise in which either too few or too many turns of speech would be subjected to analysis. We did not want to select only the single turn of speech that immediately preceded the targeted turn as we understood that especially in calendar interviews, the behaviors of turns that had occurred earlier could have a lasting influence for a number of subsequent turns. However, we also did not want to extend our analyses too far backward, as impact would diminish as the number of intervening turns increased. With these constraints in mind, we fully understand that isolating three preceding turns is based on more subjective than empirical criteria, and that future work may wish to examine more turns.

For partly empirical (e.g., Belli et al. 2013) and partly theoretical reasons (e.g., Belli 1998), the four respondent retrieval strategies we examined were parallel, timing, duration, and sequential behaviors. These four behaviors exhaust what our empirical work in behavior coding has discovered as comprising both parallel and sequential retrieval strategies, which are those respondent retrieval strategies that have been theoretically hypothesized and empirically shown to be associated with better data quality in calendar interviews. All occurred spontaneously in respondents' verbal behavior, that is, the behaviors were not a direct reflection of a query made by an interviewer. A *parallel* retrieval strategy occurred when a respondent spontaneously remembered a contemporaneous event from a life domain that was different than the one being queried. A *timing* retrieval strategy was present when a respondent spontaneously indicated a beginning or ending location in time of a reported event. A *duration* retrieval strategy consisted of spontaneous reports of the length in time of events. Finally, a *sequential* retrieval strategy occurred when a respondent spontaneously reported an event that occurred earlier or later than one which had already been identified within the same domain.

Of the 35,291 transcribed respondent turns of speech from the 165 interviews, 1,744 were identified as including a respondent parallel retrieval, and 2,821, 1,191, and 765 included timing, duration, and sequential behaviors, respectively. In order to apply the data-mining algorithm, in addition to turns which included respondent retrieval strategies, we had to simultaneously analyze turns in separate tests for each behavior that did not consist of any respondent retrievals in order to find series in the preceding turns that were diagnostic of each of the behaviors in the targeted turns. As nonretrieval turns are considerably more numerous than those turns that contain a respondent retrieval, problems associated with imbalance arise. Data-mining algorithms tend to assume relatively balanced distributions (He and Garcia 2009) as imbalanced data sets reduce the predictive power of these algorithms (Weiss and Provost 2001).

To achieve balanced distributions, we kept all turns that did include a respondent retrieval behavior and randomly sampled, for each behavior separately, an equal number of respondent turns that did not include the targeted behavior. Hence, we conducted four separate analyses in which a total of 3,488 turns of speech were examined for parallel retrieval behaviors in one analysis, and 5,642, 2,382, and 1,530 turns were examined for timing, duration, and sequential retrieval behaviors in each of three analyses, respectively. We adopted a decision-tree data-mining algorithm called C4.5 (Witten et al. 2011) and

separately applied it to each of the four respondent retrieval behaviors in their respective analyses. This algorithm was used to discover what behavioral series in the prior three turns are most predictive of the state for the respondent retrievals in the targeted fourth turn.

The decision-tree algorithm "grows" multiple-behavior series using a top-down approach with two heuristics. First, the algorithm applies a heuristic to choose the behavior that most improves the predictive power for the series (*behavior-select heuristic*) from all of the 30 interviewer and 29 respondent verbal behaviors that could potentially appear in the selected preceding three turns. The behavior chosen by this heuristic is added to the tree as an internal node. More on this behavior-select heuristic (based on information gain) will be discussed later in this section.

After the application of the behavior-select heuristic, the algorithm divides the turns into two groups based on the occurrence or nonoccurrence of the chosen behavior. The first group contains the turns with occurrence of that behavior, while the second group contains turns with nonoccurrence. The occurrence and nonoccurrence are added as edges under the node in the tree.

Now, the algorithm uses another heuristic to decide whether growing the series further would improve its predictive power (*continue-growing heuristic*). The heuristic makes this decision *separately* for each group by evaluating (1) the distribution of the respondent behavior for turns in the group and (2) the size of the group. In general terms, there are two cases where the continue-growing heuristic should decide to stop:

1. The heuristic should stop when the group has a sufficiently homogenous state for the respondent behavior (present or absent) – the series has already mastered the group and growing the series further will not improve predictive power.
2. The heuristic should stop when the group is too small, since growing the series on outliers could actually hurt overall predictive power; that is, it could lead to overfitting.

In both cases, the series is ready to make a final prediction (on the group) since predictive power is unlikely to be improved. The final prediction for the respondent behavior is set to the majority respondent-behavior state for the turns in the group. This final prediction is then added as a node to the decision tree.

Alternatively, the heuristic should continue to grow the series to try and improve its predictive power. To this end, the decision-tree algorithm restarts, running the above process again using only the turns in the group. The new behaviors are added as additional internal nodes connected to the previous behavior occurrence and nonoccurrence edges.

Figure 1 provides a graphical illustration for a possible decision tree. The behaviors chosen are listed inside boxes with the occurrence or nonoccurrence of these behaviors given as edges. The round nodes are the final prediction for the respondent behavior on a group. As alluded to earlier, the decision tree allows for multiple series of behaviors to be grown. These series can be discovered by tracing a path through the decision tree from the behavior at the top to a circle. As an example, the occurrence of Behavior A at any of the preceding three turns combined with occurrence of Behavior B at any of the preceding three turns leads to a targeted respondent parallel being present at the fourth turn, as indicated by a yes circle.

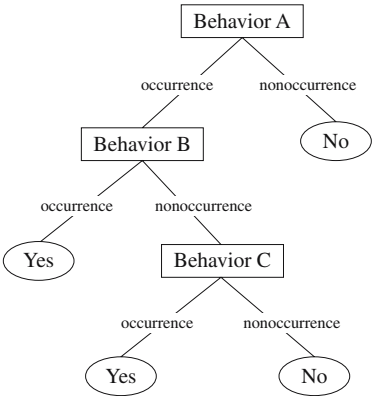


Fig. 1. Graphical example of decision tree. Behaviors are given in boxes with occurrence or nonoccurrence on the edges. The circle indicates whether or not respondent behavior is present (Yes or No). Sequences are discovered by tracing from the behavior at the top (Behavior A) to a circle. Example sequence on the above tree: occurrence of Behavior A combined with occurrence of Behavior B leads to respondent-parallel being present.

Behavior-Select Heuristic

Our behavior-select heuristic uses the same information gain as the C4.5 algorithm. We first test all of the behaviors that occur in the previous turns separately and calculate the extent to which the presence and absence of each behavior affects the distribution for the respondent retrieval in the targeted fourth turn. A completely homogenous distribution is one that only contains retrieval-present turns, or retrieval-absent turns. Hence, and using respondent parallel retrievals as an example, the greatest degree of homogeneity would result if a behavior was identified that (1) when it occurred in the previous turns, only respondent parallel-*present* targeted turns were observed, *and* that (2) when this behavior did not occur in the previous turns, only parallel-*absent* targeted turns were observed, or vice versa.

Our heuristic measures the degree of homogeneity using the entropy index. This index is measured *solely based on the homogeneity of the* ***retrieval behavior***. This index is calculated to equal 1 in conditions in which there is a lack of homogeneity, and it is equal to 0 when there is complete homogeneity. The entropy index for a single state is calculated according to the formula:

$$Entropy(S) = -\left(\frac{|S_{r=p}|}{|S|} * \log_2\left[\frac{|S_{r=p}|}{|S|}\right) - \left(\frac{|S_{r=a}|}{|S|} * \log_2\left[\frac{|S_{r=a}|}{|S|}\right)\right]\right), \qquad (1)$$

where S is the set of turns that we are interested in (e.g., the initial state of 3,488 turns), $S_{r=p}$ is the subset of these turns with the targeted retrieval behavior being present, and $S_{r=a}$ is the subset with the retrieval behavior being absent. Hence, and again using parallel retrievals as an example, the initial state of 1,744 parallel-present target turns and 1,744 parallel-absent target turns as a whole is calculated as having an $Entropy = 1$. On the other hand, a state with only parallel-present turns would have $Entropy = 0$.

As alluded to earlier, our behavior-select heuristic calculates the extent to which the presence and absence of each behavior affects the distribution for the respondent retrieval. This calculation measures the information gain *for each preceding occurring behavior*.

The information gain measures how applying the behavior, by splitting the turns into subsets where the behavior is present and absent, affects the entropy measured on the targeted final retrieval behavior. For each final retrieval behavior, two entropy indices are calculated, one index for the behavior-present state in which the behavior occurred in the previous three turns, and one index for the behavior-absent state:

$$Gain = Entropy(S) - \frac{|S_{b=p}|}{|S|} Entropy(S_{b=p}) - \frac{|S_{b=a}|}{|S|} Entropy(S_{b=a}), \qquad (2)$$

where S is the set of turns that we are interested in, $S_{b=p}$ is the subset of these turns with retrieval behavior-present, and $S_{b=a}$ is the subset with retrieval behavior-absent turns.

Based on the above equation, the preceding behavior with the highest information gain is the one that, when it is applied, provides the highest increase in homogeneity for the final retrieval behavior. Using the previous example, the highest information gain occurs when the presence of a preceding behavior always results in a final respondent parallel behavior being present, and the absence of a preceding behavior always results in the absence of a final respondent parallel behavior (or vice versa). Such a result would "zero out" terms 2 and 3 in the equations leading to $Gain = 1 - 0 - 0 = 1$.

Continue-Growing Heuristic

Once an initial behavior is identified that produces the largest information gain, we need to decide whether to continue growing the series adding additional behaviors. To address the homogenous-state stop case, our heuristic considers both (1) the branch in which a previously identified behavior was present in the preceding three turns, and (2) the branch in which a previously identified behavior was *not* present in the preceding three turns. The goal is to allow the decision-tree algorithm to proceed down any branch until the group achieves an $Entropy = 0$ when all the turns share a homogenous state. Any verbal behavior that is identified at any step as maximizing information gain is contingent on the presence or absence of behaviors identified during previous steps within that branch. As alluded to earlier, this allows the algorithm to produce a hierarchical network of series that each consist of steps with the isolation of specific behaviors whose presence or absence is required for each of the additional subsequent steps.

To address the small-group stop case, our heuristic uses a criterion rule, for each of the four analyses separately, which requires that a minimum of five percent (or 80) targeted turns need to contain a respondent retrieval for any specific behavior to be retained (either in the behavior-present state or behavior-absent state). Implementing the criterion rule resulted in a network of four steps and five different behavior series for parallel and timing retrievals (see Tables 1 and 2), and two steps and three series for both duration and sequential retrievals (see Tables 3 and 4).

Multiple-Behavior Series Results

For parallel retrievals (see Table 1), in the three turns (two interviewer and one respondent) that preceded the targeted respondent turns, four behaviors were found to discriminate between the occurrence and nonoccurrence of these retrieval strategies in the targeted turns: respondent timing in the second turn, and interviewer parallel, timing, and duration probes in the first or third turns. *Parallel probes* are defined as verbal behaviors in

Table 1. Behavior series and their statistics: parallel retrieval.

		Series				
Step	Behavior	I	II	III	IV	V
1	R Timing (2nd turn)	P	A	A	A	A
2	I Parallel		P	A	A	A
3	I Timing			P	A	A
4	I Duration				P	A
		Series				
	Statistics	I	II	III	IV	V
	N turns	340	202	594	124	484
	Proportion Turns	.195	.116	.341	.071	.278
	Series Ratio	.744	.795	.537	.582	.332

Notes: P = Behavior-Present; A = Behavior-Absent.

which interviewers use a contemporaneous event in another life domain as an anchor to cue respondents in the remembering of a domain-relevant event, *timing probes* asked when an event started or stopped and *duration probes* consist of interviewers asking how long an event occurred. Series I in Table 1 consists of at least one respondent timing retrieval in the second preceding turn, and this behavior is not at all constrained by the presence or absence of any other behaviors within the preceding three turns. Series II requires that there is no respondent timing retrieval behavior in the second turn, but that there is an interviewer parallel probe in either the first and third preceding turns. Series III is marked by the absence of respondent timing and interviewer parallel behaviors, and the presence of an interviewer timing probe. In Series IV, there must be an absence in the three preceding turns of respondent timing and interviewer parallel and timing behaviors, but the presence of an interviewer duration probe. Series V requires the absence of all four of these behaviors in the preceding three turns.

Table 2. Behavior series and their statistics: timing retrieval.

		Series				
Step	Behavior	I	II	III	IV	V
1	I Duration	P	P	A	A	A
2a	I Timing	A	P			
2b	R Timing (2nd turn)			P	A	A
3	I Data Elements				P	A
		Series				
	Statistics	I	II	III	IV	V
	N turns	541	160	287	350	1483
	Proportion Turns	.192	.057	.102	.124	.526
	Series Ratio	.780	.608	.651	.327	.467

Notes: P = Behavior-Present; A = Behavior-Absent.

Table 3. Behavior series and their statistics: duration retrieval.

		Series		
Step	Behavior	I	II	III
1	R Timing (2nd turn)	P	A	A
2	I Timing		P	A
		Series		
	Statistics	I	II	III
	N turns	200	442	549
	Proportion Turns	.168	.371	.461
	Series Ratio	.694	.562	.420

Notes: P = Behavior-Present; A = Behavior-Absent.

Table 1 also includes statistics that are associated with each series. The number of turns out of the total of 1,744 that include a respondent parallel retrieval are provided, as is the proportion (Series *N*/1744). As can be seen, Series III, which included the presence of interviewer timing, accounted for the most turns, and Series IV the least. In addition, the series ratio indicates the extent to which each series discriminated between turns with and without respondent parallel retrievals, with the higher values indicating greater discriminability. The ratio is calculated as the number of turns with respondent parallel retrieval in the targeted turn divided by the total number of turns that fit the series in the three turns that preceded the targeted turn. This ratio does not reflect the actual discriminability among all of the turns in the interview, as the ratio accounts only for the 1,744 nonrespondent retrieval turns that were randomly sampled.

Tables 2–4, which depict the data-mining results for respondent timing, duration, and sequential behaviors respectively, can be interpreted in the same way. For the most part these behaviors include interviewer or respondent timing and duration behaviors. Table 2 reveals in Series IV that interviewer data-element probes, in which interviewers ask for

Table 4. Behavior series and their statistics: sequential retrieval.

		Series		
Step	Behavior	I	II	III
1	R Timing (2nd turn)	P	A	A
2	R Parallel (2nd turn)		P	A
		Series		
	Statistics	I	II	III
	N turns	163	90	512
	Proportion Turns	.213	.118	.669
	Series Ratio	.751	.732	.310

Notes: P = Behavior-Present; A = Behavior-Absent.

detailed information such as the names of persons or employers, are predictive of respondent timing retrievals in the fourth turn.

All of the data-mining results (see Tables 1–4) consistently reveal that those behaviors that best predict the occurrence of respondent retrievals are either interviewer retrieval probes or other types of respondent retrievals. Importantly, retrieval behaviors are a subset of possible behavior types that were included in the analysis; they included interviewer and respondent conversational and rapport behaviors (e.g., clarifications, digressions), interviewer feedback behaviors that followed responses (e.g., "thank-you"), and respondent cognitive-difficulty behaviors (e.g., requests for question repeats). These data-mining results confirm the general finding from factor-analytic approaches that the same types of verbal behaviors tend to cluster together (Belli et al. 2001; Belli et al. 2013; Belli et al. 2004), but our data-mining results provide greater detail on the actual sequencing of behaviors as they occur in the exchanges between interviewers and respondents.

3.2. *Validation Regression Models*

The calendar retrospective reports were validated against these same respondents' reports, which had been provided in annual PSID interviews. Twelve cases suffered from processing errors, making the comparison of calendar retrospective reports to panel reports unfeasible, resulting in 153 validated cases. In this validation, we tested four domains in the reporting of residential, relationship, and labor histories; one was associated with residence, a second with marriage. With labor histories, we tested both employment and unemployment. These domains were selected as they each were designed to ask respondents to provide retrospective reports of objective facts. For each of these domains, we calculated respective measures of discrepancy; these were calculated as the proportion of years in which there was no match in status between the calendar and panel reports.

We tested logistic regression models to determine relationships between discrepancy and behaviors. In order to demonstrate that parallel, duration, timing, and sequential retrievals were associated with discrepancy, each of the respondent retrieval behaviors were tested. Models were examined for each of the domains separately and for each behavior separately, leading to sixteen analyses. In each model, per case, the discrepancy measure was regressed on the number of retrieval behaviors that had occurred up to and including the point at which each respective domain had been finalized during the interview, an experiential complexity measure based on the number of status changes for each respective domain in the panel data, a term for the interaction of the number of retrievals with experiential complexity, and control variables that included interviewer age, gender, and years of interviewing experience, and respondent age, gender, race, and years of education. The interaction term was included to determine whether the association of behavior with discrepancy hinged on those respondents who have more complicated histories; it may be the case that advantages of retrieval behaviors, if any, would only exist when retrieval is more difficult because the respondent has a complicated past (see Belli et al. 2013). If the model revealed a significant respondent retrieval behavior by experiential complexity interaction, a regions-of-significance (ROS) analysis was performed to determine at what level of experiential complexity the retrieval behavior

was significantly associated with discrepancy. If the model did not reveal a significant interaction, the same core regression model without the interaction term was tested to determine whether the number of retrieval behaviors revealed a significant main effect on discrepancy.

To account for clustering of respondents within interviewers, covariance matrices were inflated using the estimated interviewer design effect for residence ($deff = 1.64$), marriage ($deff = 2.00$), and employment ($deff = 1.45$) discrepancy measures. Due to this clustering, it is appropriate to estimate the degrees of freedom used in significance tests as the number of interviewers – 1 = 12 (see Belli et al. 2013). The estimated design effect for the unemployment discrepancy measure was slightly less than 1 (0.97), indicating there was no interviewer clustering effect; hence, there was no need to inflate the covariance matrices.

Table 5 presents the results of the regression models, testing for interaction effects and their accompanying ROS results when significant. A greater number of respondent parallel retrieval behaviors is associated with less discrepancy in reports of being employed when respondents experienced greater experiential complexity. However, follow-up main-effects analyses revealed that a higher number of parallel retrievals is associated with greater discrepancy in reports of unemployment ($\beta = 0.014$, $SE = 0.006$, $p = .04$). As for timing retrieval behaviors, their greater propensity is associated with greater discrepancy in reports of residence and unemployment when respondents have less experiential complexity, but less discrepancy in reports of employment and unemployment when respondents have higher experiential complexity. Follow-up main-effects analyses also reveal that the greater number of timing retrievals is associated with less discrepancy in reports of marriage ($\beta = -0.061$, $SE = 0.015$, $p < .001$). With duration retrieval behaviors, their greater number is associated with less discrepancy in the reporting of employment when there is greater experiential complexity, and they also reveal less discrepancy in reports of marriage ($\beta = -0.091$, $SE = 0.032$, $p = .02$). Finally, a greater prevalence of sequential retrieval behaviors is associated with greater discrepancy in the reports of a) residence and unemployment with lower experiential complexity, and b) marriage with higher experiential complexity. They are also associated with less discrepancy in the reports of marriage with less experiential complexity and unemployment with higher experiential complexity. Overall, results indicate that respondents' engagement in timing and duration retrieval behaviors is beneficial to the accuracy of retrospective reports, especially when experiential complexity is high, but that engagement in parallel and sequential retrieval behaviors is mixed and nuanced in terms of data quality.

Having demonstrated the associations between each of the retrieval behaviors and discrepancy across the four domains, we tested logistic regression models to examine each of the series (see Tables 1–4) for each of the domains. These models included the same control variables and inflation of covariance matrices that were included in the models testing for respondent retrievals. In each model, discrepancy was regressed on the number of fourth-turn retrieval behaviors per case that met each interactional series as identified in Tables 1–4 (i.e., the five series for parallel, the five for timing, the three for duration, and the three for sequential retrievals). Only the retrieval behaviors that occurred up to and including the point in which each domain was being interviewed were included in the analysis.

Table 5. Logistic regression coefficients for the interaction of respondent retrieval behaviors and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy.

	Interaction Parameters			Percentiles of experiential complexity in which a greater number of respondent retrieval behaviors (in comparison to a fewer number) is associated with significantly	
Domain	Beta	SE	p	Less discrepancy	Greater discrepancy
Parallel					
Residence	−0.090	0.061	ns		
Marriage	−0.345	0.576	ns		
Employment	−0.180	0.075	0.013	80.41–100	
Unemployment	−0.140	0.069	ns		
Timing					
Residence	−0.151	0.053	0.047		0–51.57
Marriage	0.846	0.580	ns		
Employment	−0.178	0.065	0.006	30.22–100	
Unemployment	−0.241	0.047	<0.001	79.77–100	0–60.36
Duration					
Residence	−0.177	0.134	ns		
Marriage	−1.827	1.177	ns		
Employment	−0.543	0.152	0.001	31.68–100	
Unemployment	−0.142	0.122	ns		
Sequential					
Residence	−0.244	0.092	0.021		0–21.61
Marriage	5.920	1.359	<0.001	0–91.34	99.03–100
Employment	−0.066	0.112	ns		
Unemployment	−0.716	0.145	<0.001	72.08–100	0–52.81

Table 6a presents the interaction parameters of the regression models with parallel retrievals in the fourth turn and their accompanying ROS results when significant. There are no significant interaction effects for unemployment. Results demonstrate that for Series II in which an interviewer parallel probe occurs in the first or third turn, reports of employment are at greater discrepancy at higher levels of experiential complexity, while for Series III in which a respondent parallel is preceded by an interviewer timing probe, reports of employment are at greater discrepancy at lower levels of experiential complexity, but at less discrepancy at higher levels of complexity. For Series V, which is marked by a lack of preceding behaviors, reports of residence demonstrate greater discrepancy with lower complexity, and less discrepancy with higher complexity. There are also significant main effects: Series I, which includes a preceding respondent timing retrieval, Series IV, which has a preceding interviewer duration probe, and Series V are all

Table 6a. Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: parallel.

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	−0.212 (0.291)			1.010 (2.286)			−0.184 (0.230)			0.380 (0.256)		
II	−0.133 (0.202)			1.484 (2.283)			0.543 (0.238)*		91.21–100	0.139 (0.227)		
III	−0.166 (0.159)			−0.714 (1.794)			−0.852 (0.170)**	85.88–100	0–61.62	0.110 (0.158)		
IV	−0.100 (0.212)			−0.303 (1.716)			−0.604 (0.507)			0.478 (0.226)		
V	−1.189 (0.454)*	98.71–100	0–35.02	−7.620 (5.466)			−0.341 (0.303)			0.525 (0.388)		

*$p < .05$; **$p < .01$; ***$p < .001$.

associated with less discrepancy in reports of marriage ($\beta = -0.138$, $SE = 0.059$, $p = .04$; $\beta = -0.120$, $SE = 0.047$, $p = .03$; $\beta = -0.395$, $SE = 0.124$, $p = .01$, respectively), and Series I is also associated with less discrepancy in reports of employment ($\beta = -0.091$, $SE = 0.028$, $p = .01$). Results demonstrate that the effectiveness of parallel retrieval behaviors is dependent on the preceding behavioral context. When preceded by a respondent timing behavior (Series I) or interviewer duration probe, data quality is improved; when preceded by an interviewer timing probe (Series III) or when there is no preceding behavior (Series V), data quality is improved when there is a more demanding retrieval task (higher experiential complexity); and when preceded by an interviewer parallel probe (Series II), data quality is worsened, especially with a more demanding retrieval task.

Results are also nuanced by which behaviors precede timing retrieval behaviors (see Table 6b). Series I, in which an interviewer duration probe precedes a timing retrieval behavior, reveals greater discrepancy at lower experiential complexity in reports of residence, but greater discrepancy at higher experiential complexity in reports of employment; in reports of employment, Series I also reveals less discrepancy with lower experiential complexity. Series II, in which a fourth-turn respondent timing behavior is preceded by both interviewer duration and timing probes, reveals less discrepancy with higher experiential complexity in reports of marriage, employment, and unemployment. With Series V, in which there are no preceding behaviors, there is less discrepancy with lower experiential complexity. There are also significant main effects: Series I and IV with marriage ($\beta = -0.157$, $SE = 0.037$, $p = .001$; $\beta = -0.093$, $SE = 0.029$, $p = .01$, respectively), and Series III, IV, and V with employment ($\beta = -0.208$, $SE = 0.043$, $p < .001$; $\beta = -0.065$, $SE = 0.012$, $p < .001$; $\beta = -0.221$, $SE = 0.049$, $p = .001$, respectively) all reveal less discrepancy. Taken together, results reveal that whereas the preceding occurrence of only duration probes (Series I) leads to mixed results with data quality, preceding duration probes in combination with timing probes (Series II) improve data quality when the retrieval task is difficult (higher experiential complexity). Moreover, preceding respondent timing behaviors (Series III) and preceding interviewer data-element probes (Series IV) improve data quality, whereas no preceding behaviors (Series V) leads to mixed results concerning data quality.

Duration retrieval behaviors at the fourth turn (see Table 6c) are noted for generally being associated with better data quality, especially at higher experiential complexity, regardless of the preceding behaviors. Both Series I, in which there is a preceding respondent timing behavior, and Series III, in which there are no preceding behaviors, are associated with less discrepancy at higher experiential complexity for employment reports; Series III also reveals less discrepancy at higher complexity and greater discrepancy at lower complexity with reports of marriage. As for Series II, in which there is a preceding interviewer timing probe, there are significant main effects in which there is less discrepancy with reports of both marriage ($\beta = -0.197$, $SE = 0.064$, $p = .01$) and employment ($\beta = -0.109$, $SE = 0.030$, $p < .01$). Table 6d reveals results for respondent sequential behaviors at the fourth turn. With Series II, in which there is a preceding respondent parallel behavior, whereas reports of residence reveal less discrepancy at higher complexity and greater discrepancy at lower complexity, the opposite pattern is seen with reports of marriage. There is also a main effect with Series I in which there is a

Table 6b. Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: timing.

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	−0.407 (0.136)**		0–82.40	0.199 (1.588)			−1.114 (0.220)***	0–47.13	83.13–100	−0.274 (0.129)		
II	1.046 (0.509)			−12.364 (5.226)*	91.11–100		−1.605 (0.492)**	56.06–100		−1.068 (0.349)**	82.93–100	
III	−0.280 (0.379)			−1.040 (2.928)			0.015 (0.394)			0.166 (0.258)		
IV	−0.144 (0.138)			1.836 (1.088)			−0.309 (0.450)			−0.034 (0.076)		
V	−0.263 (0.377)			8.292 (3.799)*	0–86.99		−0.209 (0.118)			−0.115 (0.219)		

*$p < .05$; **$p < .01$; ***$p < .001$.

Table 6c. Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: duration.

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	0.193 (0.482)			2.157 (4.263)			−2.397 (0.697)**	40.01–100		0.037 (0.321)		
II	−0.307 (0.273)			1.278 (2.613)			−0.548 (0.332)			−0.119 (0.204)		
III	−0.451 (0.267)			−12.095 (2.531)***	90.97–100	0–74.03	−0.837 (0.254)**	34.39–100		−0.105 (0.195)		

*$p < .05$; **$p < .01$; ***$p < .001$.

Table 6d. Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: sequential.

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	−0.743 (0.382)			−2.669 (3.827)			−0.172 (0.396)			0.236 (0.351)		
II	−0.310 (0.120)*	98.52–100	0–26.90	12.095 (1.937)***	0–91.11	94.49–100	−0.016 (0.130)			−0.060 (0.126)		
III	−0.204 (0.428)			2.683 (4.568)			−1.021 (0.496)			0.364 (0.508)		

*$p < .05$; **$p < .01$; ***$p < .001$.

preceding respondent timing behavior such that there is less discrepancy in reports of marriage ($\beta = -0.284$, $SE = 0.111$, $p = .03$).

4. Implications for Interviewing Research and Practice

Results present a nuanced and complicated pattern of interactions among interviewer and respondent verbal behaviors in impacting the quality of retrospective reports in calendar interviews. Of the retrieval behaviors examined, both respondent and interviewer timing behaviors and respondent duration behaviors demonstrated the most consistent association with higher data quality, especially when the retrieval task was more difficult as measured by experiential complexity. In terms of respondent timing behaviors, overall prevalence was associated with better data quality; data quality was improved when preceded by interviewer timing and data-elements probes and respondent timing retrievals, and data quality was also improved when respondent timing behaviors preceded respondent parallel, timing, duration, and sequential retrieval behaviors. As for interviewer timing behaviors, their occurrence facilitated better data quality and preceding respondent parallel, timing, and duration behaviors. Respondent duration behaviors also revealed better data quality with overall prevalence with heightened retrieval difficulty, and data quality was improved when preceded by respondent and interviewer timing, or by the absence of behaviors.

Mixed results in terms of data quality were found for respondent and interviewer parallel behaviors, and respondent sequential and interviewer duration behaviors. The overall prevalence of both respondent parallel and sequential behaviors was not consistently associated with improved data quality, even when examining only situations in which the retrieval task was difficult, nor did the presence of these behaviors when preceded by other behaviors, or when preceding retrieval behaviors, produce consistent results in terms of data quality. The preceding presence of interviewer parallel probes led to poor data quality outcomes when followed by respondent parallel retrieval behaviors. Interviewer duration probes, when alone in preceding respondent timing retrieval behaviors, led to mixed results with data quality.

These results have implications for interviewer behaviors in calendar interviews. Whereas interviewer timing probes are to be encouraged, interviewer parallel probes are to be discouraged. As for interviewer duration probes, they appear to be effective only when used in combination with interviewer timing probes. Respondent timing strategies are also to be encouraged, and their occurrence is facilitated by interviewer timing, duration, and data-elements probes, although, as noted above, interviewer duration probes should not be administered alone. The encouragement of effective respondent duration strategies is also facilitated by interviewer timing probes.

One troubling aspect of providing advice in encouraging interviewer and respondent timing behaviors is that although heightened prevalence is associated with better data quality when the retrieval task is difficult, in some situations there is also poorer data quality when the retrieval task, as measured by lower levels of experiential complexity, is relatively easy. Such a pattern has also been observed by Belli et al. (2013), who speculate that some respondents experience general difficulty in remembering their pasts, and that interviewers are more prone to unsuccessfully use retrieval probes as an attempt to

improve these respondents' memories. Accordingly, it may be advisable to attempt introducing screener questions to assess how much status changes have occurred in respondents' pasts, and to encourage interviewer and respondent timing behaviors for those respondents whose pasts are more complicated.

Another caveat in terms of attempts at implementing more successful interviewer training regimens is that some level of better data quality is due to respondents engaging in retrieval strategies on their own. For example, although respondent timing retrieval behaviors appear to be encouraged by interviewer timing, duration, and data-element probes, they also may occur spontaneously, and hence their benefits to data quality may be present only among a certain subset of respondents or circumstances in which interviewer probing has no impact. As other examples, both respondent parallel and duration behaviors, when preceded by no behaviors, are associated with better data quality especially with respondents who have complicated pasts, and hence, appear to be driven by respondents on their own.

5. Theoretical Considerations

A major surprise in our results is the lack of solid evidence that interviewer and respondent parallel behaviors improve the quality of retrospective reports. Much of the theoretical rationale of calendar interviewing has hinged on the notion that its implementation encourages the occurrence of effective cuing mechanisms, especially parallel behaviors (for examples see Balán et al. 1969; Belli 1998; Belli and Callegaro, 2009; Yoshihama et al. 2002). The finding that interviewer parallel probes lead to poorer retrospective data quality is particularly contrary to theoretical expectations. Gaining a better understanding of the role of parallel associations in human autobiographical memory may assist in determining how such associations impact accuracy.

Psychologists who have theorized about the structure of autobiographical knowledge have differed in opinion on whether parallel associations exist across contemporaneous events from different autobiographical domains or themes. On the one hand, theories of autobiographical memory that incorporate associations among different life domains are supported by the existence of respondent parallel cuing, as observed by Bilgen and Belli (2010) and as evident from the results reported in this article. Specifically, Barsalou (1988) observed that persons will follow parallel tracks of events that associate contemporaneous events across themes, such as events that encompass a project such as school being associated with contemporaneous events of being with one's family. He projected that such associations could exist between different events from different life domains. Similarly, Means, Loftus, and colleagues (Means and Loftus 1991; Means et al. 1989) observed that individuals would jump between work and health events when answering questions about their memory for health visits.

On the other hand, the presence of these parallel associations is not emphasized in some theories of autobiographical memory that highlight hierarchical associations among events within the same life domain (Conway and Bekerian 1987; Conway 1996). Specifically, Conway and Bekerian propose the existence of Autobiographical Memory Organization Packets (A-MOPs) that hierarchically organize more specific episodic events within abstract lifetime periods. As these A-MOPs are thematic with respect to encapsulating

hierarchies consisting of different life domains, such as one's relationships versus one's career, an autobiographical memory structure consisting only of A-MOPS would not predict events belonging within one life domain to be cued by contemporaneous events that had occurred in a different life domain.

Overall, our results point to a compromise between these views. It may be the case that direct parallel associations are somewhat uncommon, and hence, that benefits from respondents' use of parallel retrievals may arise, but only when these direct parallel associations exist. However, directing respondents to engage in parallel retrieval through the use of parallel probes may divert respondents from more beneficial within-domain associations and increase task difficulty, leading to poorer autobiographical remembering.

6. Limitations and Other Considerations

Although results suggest that interviewer parallel probing does more harm than good and hence ought to be discouraged in interviewer training, the observational nature of this research means that these causal inferences are tentative. It may be possible, for example, that interviewers are more likely to engage in parallel probing with respondents who exhibit poor memory, and that the association of parallel probing with poorer data quality is the outcome of respondents who are not able to remember their pasts very well. It may also be the case that our data are limited in that they arose from telephone interviewing in which the calendar was not observed by respondents, and the use of other data-collection modes in which respondents can view the calendar may lead to overall benefits from parallel associations.

In addition, as our research is not experimental, more definitive answers regarding the impact of parallel probing could be gained through experimental work in which interviewers are either encouraged or discouraged to use parallel and other types of probes. Of course, the concern with making causal inferences also applies to timing probes, which were found to be associated with better quality data.

Results are also limited in that they have only examined the value of data-mining techniques with retrieval behaviors in calendar interviews, and with a relatively small sample. Extensions of data mining should also be applied to other behaviors in both calendar and conventional questionnaire interviewing instruments, especially those of a more direct motivational flavor. As for calendar interviews, the various series of behaviors that will lead to respondent rapport behaviors, such as digressions and laughter, may be of particular interest, as the use of rapport has been shown to be associated with better retrospective reporting in some domains but not others (Belli et al. 2013). As for conventional questionnaires, question-answer series that lead to behaviors signifying that respondents are having cognitive problems with the question may be especially informative, given that these behaviors have often been associated with poorer data quality (Belli and Lepkowski 1996; Draisma and Dijkstra 2004; Dijkstra and Ongena 2006; Dykema et al. 1997).

These are but examples, of course. The key message to take away is that data-mining techniques can be used in behavior-coding analyses to uncover those series of behaviors that produce key data-quality relevant behaviors. In combination with regression techniques, it can also be determined which of these series are associated with better or poorer data quality. The results from these investigations are important theoretically in the

understanding of cognitive and communicative processes, and they have implications for interviewer training and in the development of best interviewing practices.

7. References

Balán, J., H.L. Browning, E. Jelin, and L. Litzler. 1969. "A Computerized Approach to the Processing and Analysis of Life Histories Obtained in Sample Surveys." *Behavioral Science* 14: 105–114.

Barsalou, L.W. 1988. "The Content and Organization of Autobiographical Memories." In *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*, edited by U. Niesser and E. Winograd, 193–243. New York: Cambridge University Press.

Belli, R.F. 1998. "The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys." *Memory* 6: 383–406. Doi: http://dx.doi.org/10.1080/741942610.

Belli, R.F. 2014. "Autobiographical Memory Dynamics in Survey Research." In *SAGE Handbook of Applied Memory*, edited by T.J. Perfect and D.S. Lindsay, 366–384. Los Angeles: Sage.

Belli, R.F., I. Bilgen, and T. Al Baghal. 2013. "Memory, Communication, and Data Quality in Calendar Interviews." *Public Opinion Quarterly* 77: 194–219. Doi: http://dx.doi.org/10.1093/poq/nfs099.

Belli, R.F. and M. Callegaro. 2009. "The Emergence of Calendar Interviewing: A Theoretical and Empirical Rationale." In *Calendar and Time Diary Methods in Life Course Research*, edited by R.F. Belli, F.P. Stafford, and D.F. Alwin, 31–52. Thousand Oaks, CA: Sage.

Belli, R.F., E.H. Lee, F.P. Stafford, and C.-H. Chou. 2004. "Calendar and Question-List Survey Methods: Association Between Interviewer Behaviors and Data Quality." *Journal of Official Statistics* 20: 185–218.

Belli, R.F. and J.M. Lepkowski. 1996. "Behavior of Survey Actors and the Accuracy of Response." In *Proceedings of the Conference on Health Survey Research Methods*, June, 1995, Breckenridge, CO, 69–74. DHHS Publication No. (PHS) 96-1013.

Belli, R.F., J.M. Lepkowski, and M.U. Kabeto. 2001. "The Respective Roles of Cognitive Processing Difficulty and Conversational Rapport on the Accuracy of Retrospective Reports of Doctor's Office Visits." In *Seventh Conference on Health Survey Research Methods*, edited by M.L. Cynamon and R.A. Kulka, 197–203. DHHS Publication No. (PHS) 01-1013. Hyattsville, MD: U.S. Government Printing Office.

Belli, R.F., L. Smith, P. Andreski, and S. Agrawal. 2007. "Methodological Comparisons between CATI Event History Calendar and Conventional Questionnaire Instruments." *Public Opinion Quarterly* 71: 603–622. Doi: http://dx.doi.org/10.1093/poq/nfm045.

Bilgen, I. and R.F. Belli. 2010. "Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires." *Journal of Official Statistics* 26: 481–505.

Brenner, M. 1982. "Response-Effects of 'Role Restricted' Characteristics of the Interviewer." In *Response Behavior in the Survey Interview*, edited by W. Dijkstra and J. van der Zouwen, 131–165. London: Academic Press.

Conway, M.A. 1996. "Autobiographical Knowledge and Autobiographical Memories." In *Remembering Our Past: Studies in Autobiographical Memory*, edited by D.C. Rubin, 67–93. New York: Cambridge University Press.

Conway, M.A. and D.A. Bekerian. 1987. "Organization in Autobiographical Memory." *Memory and Cognition* 15: 119–132. Doi: http://dx.doi.org/10.3758/BF03197023.

Draisma, S. and W. Dijkstra. 2004. "Response Latency and (Para)Linguistic Expressions as Indicators of Response Error." In *Methods for Testing and Evaluation of Survey Questionnaires*, edited by S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer, 131–147. Hoboken, NJ: Wiley.

Dijkstra, W. and W. Ongena. 2006. "Question-Answer Sequences in Survey Interviews." *Quality and Quantity* 40: 983–1011. Doi: http://dx.doi.org/10.1007/s11135-005-5076-4.

Dykema, J., J.M. Lepkowski, and S. Blixt. 1997. "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 287–310. New York: J.W. Wiley and Sons.

Freedman, D., A. Thornton, D. Camburn, D. Alwin, and L. Young-DeMarco. 1988. "The Life History Calendar: A Technique for Collecting Retrospective Data." In *Vol. 18 of Sociological Methodology*, edited by C.C. Clogg, 37–68. San Francisco: Jossey-Bass.

Glasner, T. and W. van der Vaart. 2009. "Applications of Calendar Instruments in Social Surveys: A Review." *Quality and Quantity* 43: 333–349. Doi: http://dx.doi.org/10.1007/s11135-007-9129-8.

He, H. and E. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21: 1263–1284. Doi: http://dx.doi.org/10.1109/TKDE.2008.239.

Means, B. and E.F. Loftus. 1991. "When Personal History Repeats Itself: Decomposing Memories for Recurring Events." *Applied Cognitive Psychology* 5: 297–318. Doi: http://dx.doi.org/10.1002/acp.2350050402.

Means, B., A. Nigam, M. Zarrow, E.F. Loftus, and M.W. Donaldson. 1989. "Autobiographical Memory for Health-Related Events." *Vital and Health Statistics*. DHHS Publication No. PHS 89-1077, Series 6, Number 2. Washington, DC: US Government Printing Office.

Weiss, G. and F. Provost. 2001. "The Effect of Class Distribution on Classifier Learning: An Empirical Study." Rutgers University Technical Report ML-TR-44.

Witten, I., E. Frank, and M. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam: Elsevier.

Yoshihama, M., K. Clum, A. Crampton, and B. Gillespie. 2002. "Measuring the Lifetime Experience of Domestic Violence: Application of the Life History Calendar Method." *Violence and Victims* 17: 297–317. Doi: http://dx.doi.org/10.1891/vivi.17.3.297.33663.

Received April 2015
Revised December 2015
Accepted May 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 601–618, http://dx.doi.org/10.1515/JOS-2016-0031

Is the Short Version of the Big Five Inventory (BFI-S) Applicable for Use in Telephone Surveys?

Oliver A. Brust[1], *Sabine Häder*[2], *and Michael Häder*[1]

The inclusion of psychological indicators in survey research has become more common because they offer the possibility of explaining much of the variance in sociological variables. The Big Five personality dimensions in particular are often used to explain opinions, attitudes, and behavior. However, the short versions of the Big Five Inventory (BFI-S) were developed for face-to-face surveys. Studies have shown distortions in the identification of the Big Five factor structure in subsamples of older respondents in landline telephone surveys. We applied the same BFI-S but with a shorter rating scale in a telephone survey with two subsamples (landline and mobile phone). Using exploratory structural equation modeling (ESEM), we identified the Big Five structure in the subsamples and the age groups. This finding leads us to conclude that the BFI-S is a powerful means of including personality characteristics in telephone surveys.

Key words: Exploratory structural equation modeling; telephone surveys; Big Five personality dimensions.

1. Introduction

Psychological indicators are being used increasingly in survey research as determining factors for the explanation of behavior and attitudes. Rammstedt (2007a), for example, showed that a significant proportion of the variance of sociological variables of interest could be explained using personality dimensions. This holds true for very different social phenomena, such as political attitudes (Heaven and Bucci 2001; Saucier 2000; Van Hiel et al. 2004), educational careers, career choices, the interaction of personality development and social relations, health history, and life course trajectories (Caspi et al. 2005; Goldberg et al. 1998). One empirical question that needs to be addressed in this context is: with which instruments and modes or devices of data collection can personality structure be assessed efficiently – that is, as briefly, reliably, and validly as possible?

In situations where personality is the primary topic of interest, the measurement of these psychological variables requires long inventories. The original German-language version of the Big Five Inventory (BFI), for example, comprises 42 items (Lang et al. 2001; Rammstedt and John 2007). However, in survey situations, where brevity is a high

[1] Technische Universität Dresden, Faculty of Philosophy, Institute of Sociology, D-01062 Dresden, Germany. Emails: oliver.brust@tu-dresden.de and michael.haeder@tu-dresden.de
[2] Leibniz Institute for the Social Sciences, PBox 122155, D-68072 Mannheim, Germany. Email: sabine.haeder@gesis.org
Acknowledgment: We would like to thank the three anonymous reviewers and the associate editor for their helpful suggestions.

priority, an inventory of this length is too time consuming. Efforts have therefore been made to develop short inventories that can be easily applied in surveys (Gosling et al. 2003; Gerlitz and Schupp 2005, 204).

Short versions of the BFI (with 10 or 15 items) have been used in large studies such as the International Social Survey Programme (ISSP), the German Socio-Economic Panel (SOEP), the British Household Panel Survey, the Household, Income and Labour Dynamics in Australia Survey (HILDA), and the German National Cohort. The Big Five personality dimensions represent a powerful means of analyzing interindividual differences in personality dimensions (Lang et al. 2011). These five dimensions can be described as follows:

- Neuroticism refers to individual differences in the susceptibility to distress and the experience of negative emotions such as anxiety, anger, and depression.
- Extraversion refers to individual differences in sociability, gregariousness, level of activity, and the experience of positive emotions.
- Openness to experience refers to individual differences in the propensity for originality, creativity, and the acceptance of new ideas.
- Agreeableness refers to individual differences in altruistic behavior, trust, warmth, and kindness.
- Conscientiousness refers to individual differences in self-control, task orientation, and rule abiding (Taylor et al. 2010, a3-21–a3-22).

As the use of psychological variables in surveys becomes more common, mixed-device surveys are becoming more frequent as well. The utilization of several devices is aimed at taking advantage of the rapid technological progress for survey research (Toepoel and Lugtig 2015). However, existing short scales for the measurement of psychological constructs have been developed and tested only for particular modes and/or devices of data collection – mostly for face-to-face interviewing with CAPI or PAPI. Hence, the question that arises is whether these short scales are also suitable for use in other modes and with other devices – for example mobile phone and landline phone surveys – and are therefore applicable in mixed-device surveys.

Because surveys will be increasingly conducted using both landline and mobile phones, it is very important that the short version of the BFI applied is suitable for users of both devices. To determine whether this is the case, we incorporated a BFI-S into the CELLA2 study (acronym for CELl Phone and LAndline Phone Survey 2), in which telephone numbers for both samples were drawn from different frames, and interviews were conducted by mobile and landline phone. CELLA2 was conducted by GESIS–Leibniz Institute for the Social Sciences, Mannheim, Germany, and Dresden University of Technology, Germany, and was funded by a research grant from the German Research Foundation (DFG).

1.1. Hypotheses

The following hypotheses will be tested:

H1: The Big Five structure of the personality dimensions Neuroticism, Extraversion, Openness to experience, Agreeableness, and Conscientiousness is clearly represented in both the landline and the mobile phone subsamples.

In an evaluation of possible mode effects among landline and mobile phone respondents in the CELLA surveys, Häder and Kühne (2010) showed that differences between the two subsamples in terms of response quality were not significant. Because the device used by respondents (landline vs. mobile phone) does not appear to have an influence on their answers, we expect to find the Big Five personality dimensions in both subsamples.

H2: The Big Five structure of personality dimensions is reproduced in all age groups (age groups: 16–39, $n = 1{,}244$; 40–59, $n = 1{,}133$; 60 and older, $n = 514$).

In 2005, Lang et al. (2011) used a short version of the BFI when conducting a landline-only computer-assisted telephone interview (CATI) study with 1,200 respondents in which they applied a 15-item BFI-S with a seven-point rating scale. However, they observed distortions in the CATI assessment of the Big Five personality dimensions among older adults. One possible explanation is the assumption "that the mental workload of the telephone interviewing context would preclude valid self-report responses, since it requires listening to interviewers while reflecting responses on a 7-point rating scale" (Lang et al. 2011, 559).

In CELLA2, we shortened the seven-point scale to a five-point scale in order to lighten the workload of answering the Big Five items. However, although the coarser scale reduces the cognitive effort required to find a satisfactory answer, it could lead to a loss of information.

2. Method

In CELLA2, 3,051 participants (aged 16–93 years, $M = 43.43$, $SD = 16.52$, 48% female) were interviewed about their telephone usage behavior. The questionnaire also included several items aimed at measuring data quality (e.g., question order effects, social desirability, and response stability; see Häder 2012; Kühne et al. 2009). In order to compare the two subsamples, 1,516 interviews were conducted via landline phone and 1,535 via mobile phone. Participants were randomly selected. The same instrument was used nationwide for both subsamples. The landline sample was drawn from the universe of possible landline numbers in Germany using simple random sampling (Gabler and Häder 2002). The sampling frame comprised 139,366,300 numbers, from which 31,358 numbers were selected. For the selection of participants for the mobile phone survey, a modified RDD method was used (Gabler et al. 2012, 147ff.). The sampling frame comprised 197,490,000 mobile phone numbers, from which 44,330 numbers were selected. Both samples were drawn by GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany. The fieldwork was carried out in the summer of 2010 by a commercial survey research firm and lasted six weeks. A maximum of 15 contact attempts was made for the gross sample ($M = 2.4$, $SD = 2.42$). Of the interviews, 42.3% were conducted on the first contact attempt and 23.3% on the second attempt. The mean duration of the interviews was 12.33 minutes. The following response rates according to AAPOR standards were realized: $RR3_l = 0.148$ for the landline phone sample and $RR3_m = 0.117$ for the mobile phone sample (Schneiderat and Schlinzig 2012, 124).

As is also the case with telephone surveys in other countries, the mobile phone subsample of CELLA2 had a higher percentage of men, was younger than the landline subsample, and more mobile phone respondents were single. Furthermore, Schneiderat and Schlinzig (2012, 129) observed an education bias in the complete sample compared to the 2009 German microcensus, the official reference statistic (complete sample/microcensus: lower secondary educational level: 20.5%/38.8%, intermediate secondary levels: 32.8%/28.5%, higher levels: 36.3%/25.7%, other: 10.4%/7.0%, for further details see Schneiderat and Schlinzig 2012, 129), in such a way that participants with a lower level of education were underrepresented. This is likely to be due only partly to the very low response rate in CELLA2 (approx. 10%), because in the 1990s, when response rates of up to 80% were reached in German social surveys, this bias was also observed (Koch 1998).

In sum, the CELLA2 sample performed well in representing the subgroups of the survey's target population, as a comparison to official reference statistics shows (Schneiderat and Schlinzig 2012, 131). Therefore, Schneiderat and Schlinzig (2012, 131) conclude: "The integration of a mobile sample by applying a dual-frame approach nearly always leads to better sample quality."

To apply the dual-frame approach in our study, the parameters described in Figure 1 are needed.

Following Gabler and Ayhan (2007), the inclusion probability of the target person i is

$$\pi_i \approx k_i^F \frac{m^F}{M^F} \cdot \frac{1}{z_i} + k_i^C \frac{m^C}{M^C}.$$

These inclusion probabilities are used to construct both the Horvitz-Thompson estimator (design weighting) and the GREG estimator (design and adjustment weighting, adjusted for sex, age, and education; see Gabler et al. 2012, 163).

In order to maximize statistical power within age group comparisons, three age groups were defined: young adulthood from 16 to 39 years (total: $n = 1{,}244$, $M = 27.80$, $SD = 6.57$; landline: $n = 516$, $M = 28.73$, $SD = 6.69$; mobile: $n = 728$, $M = 27.14$, $SD = 6.41$), middle adulthood from 40 to 59 years (total: $n = 1133$, $M = 48.80$, $SD = 5.61$; landline: $n = 583$, $M = 48.75$, $SD = 5.71$; mobile: $n = 550$, $M = 48.81$, $SD = 5.50$), and late adulthood comprising people aged 60 years and older (total: $n = 504$, $M = 69.03$, $SD = 6.66$; landline: $n = 343$, $M = 69.81$, $SD = 6.72$; mobile: $n = 171$, $M = 67.45$, $SD = 6.29$). The CELLA2 data can be found in the Data Archive for the Social Sciences at GESIS – Leibniz Institute for the Social Sciences, Germany.

Landline phone		Mobile phone	
M^F	Number of numbers in sampling frame	M^C	Number of numbers in sampling frame
m^F	Number of numbers in sample	m^C	Number of numbers in sample
k_i^F	Number of landline numbers at which target person i can be reached	k_i^C	Number of mobile phone numbers at which target person i can be reached
z_i	Size of household to which target person i belongs		

Fig. 1. Parameters for the dual-frame approach.

To capture the Big Five personality dimensions it was important to choose the instrument most suitable for use in a telephone survey. One possibility was the very short and efficient ten-item BFI-S developed by Rammstedt (2007b). Within the framework of the MOBILEPANEL project (see Häder et al. 2010), we conducted a pretest of this instrument with a panel of $n = 203$ persons who were interviewed via mobile phone. The results of the pretest showed that a more extensive instrument was needed for the telephone-based measurement of the Big Five because, even with ipsative data, exploratory factor analysis (EFA) failed to identify the five-factor structure. The personality dimensions were therefore measured using another short version of the BFI – namely, a 15-item instrument that was constructed for the SOEP and was used in this context for the first time in 2005 (Dehne and Schupp 2007). Whereas the SOEP presented the instrument visually (using a template) within the framework of a face-to-face interview, in our study CELLA2 it was administered by landline and mobile phone. To avoid overburdening the respondents, the response scale was reduced from seven to five scale points (1 – strongly disagree, 5 – strongly agree; Lang et al. 2011; Dehne and Schupp 2007, 8). Similar to Lang et al. (2011, 554), Cronbach's alpha values for the BFI-S scales were low, reflecting the brevity of the three-item scale and the width of these broad constructs (Neuroticism $\alpha = .52$, Extraversion $\alpha = .60$, Openness $\alpha = .55$, Agreeableness $\alpha = .45$, Conscientiousness $\alpha = .52$). As Gosling et al. (2003) demonstrate by comparing Cronbach's alpha values to test-retest reliability values in a ten-item Big Five measure, Cronbach's alpha might not be the right indicator to evaluate reliability for very brief scales due to an underestimation of the true reliability. However, an evaluation of the test-retest reliability was not possible in CELLA2 due to the design of the study. To avoid possible sequence effects, the items of the BFI-S were presented in random order. This randomization was implemented by the CATI software and changed with every participant. Within the questionnaire, the BFI-S was situated after a set of items concerning telephone usage behavior and was followed by other personality measures.

In the 2005 SOEP study, over 20,000 people were interviewed with this instrument. Therefore it was used as a reference for the CELLA2 results. The response rate of the SOEP is about 50% (see Goebel et al. 2008), which is significantly higher than that of CELLA2 (approx. 10%). Table 1 shows the results of the comparisons of the item means of the BFI-S between the two studies.

Overall, it can be seen that the differences between the SOEP and CELLA2 can be regarded as small. They do not exceed less than half a scale point on the seven-point scale. This indicates a satisfactory quality of the realized CELLA2 sample, despite the low response rate. A further comparison of the sample means for the BFI-S items in CELLA2 using the Horvitz-Thompson estimator and the GREG estimator did not reveal a general tendency that could be interpreted as an improvement or deterioration of the estimators (see Table 2). Therefore, in our next model-based analyses we forgo the use of weights.

2.1. Statistical Analysis for Testing Measurement Invariance

To compare the landline and mobile phone samples within the total sample and across the three different age groups (young adulthood, middle adulthood, late adulthood), we conducted exploratory structural equation modeling (ESEM) analyses using Mplus

Table 1. Means of the BFI-S Items in the SOEP and CELLA2.

Item: I see myself as someone who. . .	CELLA2	SOEP	Difference: CELLA2–SOEP
N: Worries a lot	4.77	4.76	0.01
N: Gets nervous easily	3.58	3.77	−0.19
N: Is relaxed, handles stress well	3.44	3.47	−0.03
E: Is talkative	5.33	5.49	0.16
E: Is outgoing, sociable	5.29	5.07	0.22
E: Is reserved	4.21	3.86	0.35
O: Is original, comes up with new ideas	4.71	4.54	0.17
O: Values artistic, aesthetic experiences	4.22	4.09	0.13
O: Has an active imagination	4.78	4.83	−0.05
A: Is sometimes rude to others	5.35	5.06	0.29
A: Has a forgiving nature	5.59	5.52	0.07
A: Is considerate and kind to almost everyone	5.93	5.78	0.15
C: Does a thorough job	6.08	6.15	−0.07
C: Tends to be lazy	5.54	5.71	−0.17
C: Does things efficiently	5.75	5.75	0.00

Note. N = Neuroticism, E = Extraversion, O = Openness to experience, A = Agreeableness, C = Conscientiousness, CELLA2 items adapted by multiplying by 7/5; Sources: data set GREG weighted.

(Version 7, Muthén and Muthén 2012) to test for measurement invariance of the short version of the Big Five Inventory (BFI-S). In the ESEM procedure, a model is estimated on the basis of an a priori postulated number of factors, thereby combining the advantages of EFA and confirmatory factor analysis (CFA). Within the procedure all factor loadings, item intercepts, and item uniquenesses are estimated. It is also possible to evaluate the fit

Table 2. Comparison of the sample means, the Horvitz-Thompson estimator and the GREG estimator for the BFI-S items in CELLA2.

Item: I see myself as someone who. . .	Sample Mean	HT-Estimator	GREG Estimator
Worries a lot	3.27	3.28	3.41
Gets nervous easily	2.49	2.51	2.55
Is relaxed, handles stress well	2.43	2.45	2.46
Is talkative	3.84	3.80	3.81
Is outgoing, sociable	3.78	3.80	3.78
Is reserved	3.15	3.12	3.01
Is original, comes up with new ideas	3.44	3.42	3.37
Values artistic, aesthetic experiences	3.09	3.10	3.02
Has an active imagination	3.52	3.49	3.42
Is sometimes rude to others	3.83	3.83	3.82
Has a forgiving nature	3.94	3.96	3.99
Is considerate and kind to almost everyone	4.21	4.22	4.23
Does a thorough job	4.33	4.32	4.34
Tends to be lazy	3.86	3.87	3.96
Does things efficiently	4.15	4.13	4.11

of the model to the data and to test for measurement invariance of the estimated parameters across multiple groups (Asparouhov and Muthén 2008). Simultaneously, disadvantages of the individual methods are reduced. Traditional EFA does not offer a method of comparing different factor structures in regard to their equivalence, whereas CFA typically requires indicators to be assigned to single factors. This rules out the possibility of indicators loading on another factor at the same time. Therefore the CFA procedure alone may not be adequate for evaluating the model fit of the Big Five model, because fit indices do not show adequate fit while correlations between the five factors are artificially inflated at the same time (Hopwood and Donnellan 2010).

We used oblique Geomin rotation, following Marsh et al. (2010). Geomin rotation is recommended when indicators have substantial loadings on more than one factor (Browne 2001; Muthén and Muthén 2012), which is often the case with the Big Five model (Hopwood and Donnellan 2010). To evaluate model fit, the maximum-likelihood estimator (ML) with conventional standard errors and chi-square test statistic was used. Compared to other estimators (e.g., maximum-likelihood estimator with robust standard errors, MLR), the ML chi-square test statistic can be used easily for chi-square difference testing and therefore for multiple-group comparisons. However, ML requires a large sample size and multivariate normal distribution. Considering that the response format was ordered categorical, normal distribution could not be ensured for all variables. However, in large samples ML has proved to be relatively robust even when slight deviations from the normal distribution occur (West et al. 1995; Ximénez 2006). Nevertheless, all models were calculated with the maximum-likelihood estimator with robust standard errors (MLR), as well. The results did not differ substantially between ML and MLR. In the following section, we report the results of the calculations using Geomin rotation and the maximum-likelihood estimator. Listwise deletion was used to handle missing data because the percentage of missing data was very small and missing data were missing completely at random (Little's MCAR test: $p = .992$; chi-square = 13,153, $df = .28$; Little 1988). Comparably to SEM, within the ESEM procedure we can test whether an a priori postulated model fits the data.

Chi-square tests were used to test for model fit and a nonsignificant result was regarded as the indicator of a fitting model. However, chi-square test results are influenced by sample size (Tucker and Lewis 1973). Therefore, we used goodness-of-fit indices, which are considered to be relatively robust even in the case of sample size differences. To evaluate model fit, the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the Standard Root Mean square Residual (SRMR) were used. CFI and TLI values greater than .90 and .95 indicate acceptable and excellent data fits, RMSEA values less than .05 and .08 indicate close and reasonable fits, and SRMR values less than .06 and .10 indicate close and reasonable fits, following common guidelines for the evaluation of model fits to the data (see Marsh et al. 1988; Hu and Bentler 1999; Marsh et al. 2004).

To test for differences between the subsamples (landline vs. mobile phone), ESEM multiple-group analyses were conducted. Five different models of measurement invariance were compared within the total sample and within the three different age groups following Lang et al. (2011; see also Marsh et al. 2013):

(1) Configural invariance is the least demanding model; it imposes no invariance constraints at all. It is used to establish a baseline condition according to which the five-factor structure exists in the two different model groups (landline vs. mobile).
(2) Weak invariance constrains the factor loadings to be invariant across the two different model groups.
(3) Strong invariance constrains the factor loadings and the item intercepts to be invariant across the two groups. A rejection of this model implies different item functioning (i.e., different item means between the two groups cannot be explained merely by differences at the factorial mean levels).
(4) Strict invariance constrains the factor loadings, the item intercepts, and item uniquenesses to be invariant across both groups. A rejection of this model indicates that differences exist in the measurement errors across both models.
(5) In the last model, factor loadings, item intercepts, item uniquenesses and factor means are constrained to be equal.

In order to determine the more parsimonious model, Bentler (1990) suggested the testing of nested models using chi-square difference testing. However, this procedure is also dependent on sample size (Brannick 1995). Therefore, the examination of changes in fit indices is also used as an alternative to this procedure (Cheung and Rensvold 1999; Chen 2007). According to Chen (2007), a more parsimonious model is supported if the CFI change is less than .01 or the RMSEA change is less than .015. According to Marsh et al. (2009), equally good or better TLI and RMSEA values compared to the less restrictive model are a more conservative criterion for the more parsimonious model.

3. Results

Table 3 shows the Geomin rotated loadings for the mobile and the landline phone samples. The Big Five factor structure is clearly identified by the 15 items of the BFI-S in both samples. The solutions show close fit (landline: chi-square/df = 80,884/40, p = .001, CFI/TLI = .987/.967, RMSEA/SRMR = .026/.014; mobile: chi-square/df = 139,827/40, p = .001, CFI/TLI = .969/.918, RMSEA/SRMR = .040/.019) and are almost textbook-like.

The fit indices of the ESEM multiple-group analysis (landline vs. mobile) for the whole sample are reported in Table 4.

In the total sample, fit indices of the ESEM multiple-group analysis (landline vs. mobile) showed close fit for the configural-invariance model and for the weak measurement-invariance model. A comparison of the fit indices of the two models indicated model improvement favoring the weak invariance model. Chi-square difference testing revealed no differences between the two models (chi-square/df = 41,863/50, p = .787, n.s.). The fit indices of the strong-invariance model also showed a close model fit. Compared to the weak invariance model, fit indices remained essentially stable. Chi-square difference testing revealed a significant disparity between the two models (chi-square/df = 25,830/10, p = .004). However, the differences between the CFI values of the strong measurement and the weak measurement-invariance model were less than .01 and a chi-square difference test between the strong- and the configural-invariance model revealed no differences (chi-square/df = 67,693/60, p = .231, n.s.). The strict measurement-invariance model also proved satisfactory. Compared to the

Table 3. Geomin rotated loadings of the BFI-S in the landline and mobile phone samples.

	Landline Sample					Mobile Phone Sample				
Item: I see myself as someone who. . .	N	E	O	A	C	N	E	O	A	C
Worries a lot	**.578**	.052	−.005	.124	.122	**.465**	.016	.009	.118	.168
Gets nervous easily	**.628**	−.023	.031	−.037	−.039	**.627**	−.010	.010	−.023	−.019
Is relaxed, handles stress well[a]	**.440**	−.028	−.048	−.148	−.129	**.424**	−.013	−.151	−.115	.134
Is talkative	.061	**.655**	.062	.069	.036	.058	**.643**	.071	.084	.027
Is outgoing, sociable	.010	**.733**	.008	.008	.036	.023	**.568**	.148	.044	.038
Is reserved[a]	−.111	**.475**	−.016	−.265	−.131	−.073	**.533**	−.069	−.191	−.130
Is original, comes up with new ideas	−.178	.070	**.520**	−.026	.108	−.106	.178	**.502**	−.072	.093
Values artistic, aesthetic experiences	.036	−.107	**.529**	.042	−.019	.022	−.066	**.507**	.076	−.063
Has an active imagination	.004	.049	**.557**	.020	−.057	.030	.047	**.580**	.008	−.054
Is sometimes rude to others[a]	−.141	−.026	−.137	**.485**	−.077	−.159	−.027	−.110	**.474**	−.075
Has a forgiving nature	−.015	.031	.141	**.273**	.038	−.015	.028	.169	**.337**	.037
Is considerate and kind to almost everyone	.018	.046	.032	**.707**	.022	.009	.094	.008	**.576**	.091
Does a thorough job	.038	−.005	−.022	−.031	**.706**	.049	−.023	−.022	.031	**.705**
Tends to be lazy[a]	−.036	.043	−.160	.077	**.286**	−.095	.088	−.116	.044	**.295**
Does things efficiently	−.046	.005	.054	.028	**.707**	−.057	.020	.037	−.021	**.693**

Note. Landline sample: $n = 1{,}516$, chi-square/df = 80,884/40, $p = .001$, CFI/TLI = .987/.967, RMSEA/SRMR = .026/.014; Mobile phone sample: $n = 1{,}535$, chi-square/df = 139,827/40, $p = .001$, CFI/TLI = .969/.918, RMSEA/SRMR = .040/.019; N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness.

[a]Item was recoded (inversed).

Table 4. Model fit of the ESEM multiple-group analysis (landline vs. mobile) for the whole sample.

Model	ML/*df*	p_{fit} p_{diff}	CFI	TLI	RMSEA	SRMR
(1) Configural invariance	221/80	.001	.978	.942	.034	.017
Differences: (2) vs. (1)	42/50	.787	.001	.025	.008	.004
(2) Weak measurement invariance	263/130	.001	.979	.967	.026	.021
(3) vs. (2)	26/10	.004	.002	.002	.000	.000
(3) vs. (1)	68/60	.231				
(3) Strong invariance	288/140	.001	.977	.965	.026	.021
(4) vs. (3)	30/14	.007	.003	.000	.000	.010
(4) vs. (1)	98/74	.033				
(4) Strict invariance	319/154	.001	.974	.965	.026	.031
(5) vs. (4)	11/5	.045	.001	.000	.001	.001
(5) vs. (1)	109/79	.014				
(5) Strict invariance and fixed factor means	330/159	.001	.973	.965	.027	.032

Note. Total sample: $n = 3{,}051$; ML/*df* = maximum-likelihood chi-square/degrees of freedom; p_{fit} = chi-square test to evaluate model fit; p_{diff} = chi-square difference test between two models; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standard Root Mean square Residual.

strong-invariance model and the configural-invariance model, chi-square difference testing revealed significant differences (chi-square/$df = 30{,}204/14$, $p = .007$; chi-square/$df = 97{,}897/74$, $p = .033$). However, compared to the strong-invariance model, there were no CFI changes greater than .01, and TLI and RMSEA values also remained stable. The test of the model for strict invariance and fixed factor means was also satisfactory. Compared to the strict-invariance model and the configural-invariance model, chi-square difference testing revealed significant differences (chi-square/$df = 11{,}321/5$, $p = .045$; chi-square/$df = 109{,}218/79$, $p = .014$). However, there were no CFI changes greater than .01 compared to the strict-invariance model, and TLI and RMSEA values also remained essentially stable.

Regarding the complete sample, the comparison of the five models of measurement invariance supports the assumption of equal factor loadings, equal item intercepts, equal item uniquenesses, and equal factor means for the landline and mobile phone samples.

In the next step we tested for measurement invariance for the landline and mobile phone samples for different age groups. The Geomin rotated loadings representing the Big Five factor structure for the three age groups and fit indices are reported in Table 5. TLI and CFI indices show at least acceptable fits; RMSEA and SRMR indicate close fits. Once again, the solutions are almost textbook-like.

3.1. Measurement Invariance in Young and Middle Adulthood

The fit indices of the ESEM multiple-group analysis (landline vs. mobile) are reported in Table 6 for the young adulthood sample, and in Table 7 for the middle adulthood sample.

With regard to the young and middle adulthood groups, the comparison of the five measurement-invariance models supports the assumption of equal factor loadings, equal item intercepts, equal item uniquenesses, and equal factor means for the landline and mobile

Table 5. Geomin rotated loadings for young, middle-aged and older adults.

Item: I see myself as someone who. . .	Young adults					Middle-aged adults					Older adults				
	N	E	O	A	C	N	E	O	A	C	N	E	O	A	C
Worries a lot	**.558**	.084	.032	.064	.168	**.483**	−.050	.037	.131	.206	**.379**	.030	−.059	.217	.064
Gets nervous easily	**.591**	−.061	.033	−.058	−.016	**.653**	.011	−.014	.003	−.003	**.759**	−.001	.014	−.047	−.028
Is relaxed, handles stress well[a]	**.497**	−.024	−.116	−.151	−.048	**.522**	.020	−.127	−.140	−.109	**.217**	−.100	−.005	.003	−.317
Is talkative	.099	**.656**	.018	.129	.028	.022	**.681**	.087	.037	.058	.031	**.593**	.094	.135	.031
Is outgoing, sociable	−.015	**.660**	.114	.009	.023	.014	**.541**	.189	.040	.073	.021	**.845**	−.078	.011	−.020
Is reserved[a]	−.110	**.535**	−.034	−.098	−.143	−.055	**.545**	−.057	−.261	−.096	−.024	**.276**	.023	−.207	−.127
Is original, comes up with new ideas	−.149	.131	**.540**	−.036	.085	−.155	.020	**.653**	−.007	.063	−.133	.283	**.356**	−.203	.137
Values artistic, aesthetic experiences	.062	−.059	**.480**	.068	−.074	.005	−.013	**.507**	.010	−.135	−.072	−.038	**.715**	.170	−.001
Has an active imagination	.039	.070	**.505**	.021	−.067	.044	.104	**.486**	−.015	−.059	−.024	.162	**.591**	−.008	−.040
Is sometimes rude to others[a]	−.052	−.006	−.242	**.584**	−.035	−.153	−.014	−.095	**.547**	−.159	−.223	−.040	−.037	**.388**	−.10
Has a forgiving nature	−.030	−.084	.189	**.280**	.039	−.010	.055	.186	**.339**	−.006	.022	.184	.060	**.238**	.106
Is considerate and kind to almost everyone	.057	.043	.026	**.573**	.053	.033	.085	.011	**.655**	.069	−.046	.159	.057	**.639**	.064
Does a thorough job	.071	−.010	−.029	−.010	**.689**	.029	.021	−.003	.026	**.622**	.009	−.001	−.049	.031	**.702**
Tends to be lazy[a]	−.009	.065	−.125	.058	**.329**	−.145	.070	−.097	.032	**.272**	−.101	.049	−.071	.166	**.170**
Does things efficiently	−.099	.007	.048	.002	**.704**	−.055	.024	.002	−.006	**.758**	.011	−.007	.106	.008	**.725**

Note. Young adults: $n = 1{,}244$, chi-square/df = 109,276/40, CFI/TLI = .971/.924, $p = .001$, RMSEA/SRMR = .037/.018; Middle-aged adults: $n = 1{,}133$, chi-square/df = 104,811/40, $p = .001$, CFI/TLI = .975/.935, RMSEA/SRMR = .038/.018; Older adults: $n = 504$, chi-square = 44,734/40, $p = .2797$, n.s., CFI/TLI = .996/.989, RMSEA/SRMR = .015/.018, ten respondents with very low education were excluded from the mobile phone sample in this age group; N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness.

[a]Item was recoded (inversed).

Table 6. Model fit of the ESEM multiple-group analysis (landline vs. mobile) for young adulthood.

Model	ML/*df*	p_{fit} p_{diff}	CFI	TLI	RMSEA	SRMR
(1) Configural invariance	146/80	.001	.973	.928	.037	.020
Differences: (2) vs. (1)	61/50	.140	.005	.020	.006	.011
(2) Weak measurement invariance	207/130	.001	.968	.948	.031	.031
(3) vs. (2)	8/10	.653	.001	.006	.002	.000
(3) vs. (1)	69/60	.209				
(3) Strong invariance	215/140	.001	.969	.954	.029	.031
(4) vs. (3)	22/14	.083	.003	.001	.000	.004
(4) vs. (1)	90/74	.095				
(4) Strict invariance	237/154	.001	.966	.953	.029	.035
(5) vs. (4)	5/5	.443	.000	.002	.000	.001
(5) vs. (1)	95/79	.104				
(5) Strict invariance and fixed factor means	241/159	.001	.966	.955	.029	.036

Note. Young adults: $n = 1{,}244$; ML/*df* = maximum-likelihood chi-square/degrees of freedom; p_{fit} = chi-square test to evaluate model fit; p_{diff} = chi-square difference test between two models; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standard Root Mean square Residual.

phone samples, with fit indices showing excellent fits and remaining essentially stable across the different models. Chi-square difference testing also revealed no differences.

3.2. Measurement Invariance in Late Adulthood

Finally, we tested measurement invariance of the BFI-S across the two sampling modes in later adulthood (i.e., 60 years and older). Neither the configural-invariance model nor the

Table 7. Model fit of the ESEM multiple-group analysis (landline vs. mobile) for middle adulthood.

Model	ML/*df*	p_{fit} p_{diff}	CFI	TLI	RMSEA	SRMR
(1) Configural invariance	153/80	.001	.973	.928	.040	.021
Differences: (2) vs. (1)	62/50	.120	.005	.020	.006	.008
(2) Weak measurement invariance	215/130	.001	.968	.948	.034	.029
(3) vs. (2)	10/10	.426	.000	.004	.001	.001
(3) vs. (1)	72/60	.137				
(3) Strong invariance	225/140	.001	.968	.952	.033	.030
(4) vs. (3)	17/14	.247	.001	.003	.001	.009
(4) vs. (1)	89/74	.109				
(4) Strict invariance	242/154	.001	.967	.955	.032	.039
(5) vs. (4)	4/5	.612	.000	.002	.001	.002
(5) vs. (1)	93/79	.137				
(5) Strict invariance and fixed factor means	246/159	.001	.967	.957	.031	.041

Note. Middle-aged adults: $n = 1{,}133$; ML/*df* = maximum-likelihood chi-square/degrees of freedom; p_{fit} = chi-square test to evaluate model fit; p_{diff} = chi-square difference test between two models; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standard Root Mean square Residual.

single-group ESEM model showed any convergence for the mobile phone respondents, due to one item that had a negative residual variance leading to a nonpositive definite covariance matrix. After checking item covariances and correlations, we were able to rule out multicollinearity and linear dependency as alternative explanations.

These results are consistent with those obtained by Lang et al. (2011), who compared measurement invariance of the 15-item BFI-S in early, middle, and late adulthood across different modes of data collection (CATI, face-to-face interviewing, self-administered questionnaires). They too observed distortions in the results of the CATI assessment of the Big Five dimensions in older adults. Lang et al. (2001) suggested two possible explanations for these distortions – namely:

a) that they might be due to the fact that the mental workload of the telephone interviewing context caused by the seven-point rating scale led to invalid self-reports, as it might be difficult for older adults to listen to the interviewer and reflect on possible responses on a seven-point rating scale at the same time, and
b) that the costs of the greater workload would manifest themselves in greater variability in item responses, which might result in a reduced likelihood of identifying the expected five-factor structure.

To simplify the assessment situation and to reduce the mental workload for older adults, we used a five-point rating scale. In contrast to the results of Lang et al. (2011, 558f.), which showed a nonacceptable model fit for the single-group ESEM model in the landline sample (chi-square/df = 83/40, $p < .001$, CFI/TLI = .909/.761, RMSEA = .069), the single-group ESEM model in our results showed excellent fit, as evidenced by a nonsignificant chi-square difference test (chi-square/df = 34,202/40, $p = .728$, n.s.).

Therefore, the simplification of the assessment situation by using a five-point rating scale may have helped older respondents to handle the telephone interviewing situation better – at least in the landline sample. By contrast, the mobile phone setting might be more difficult for older respondents. According to Lang et al. (2011), the distortion of self-report responses due to a higher mental workload might be a problem for less-educated older respondents in particular. They therefore conducted ESEM analyses excluding older adults with only eight or nine years of education from the CATI sample, which allowed them to successfully test for measurement invariance across the three different conditions. Following Lang et al. (2011), we also excluded ten respondents who did not have any school graduation qualifications from the mobile phone sample, and then tested for measurement invariance for the landline and mobile phone samples.

The fit indices of the ESEM multiple-group analysis for the late-adulthood sample are reported in Table 8.

The configural-invariance model, the weak measurement-invariance model, the strong-invariance model, and the strict-invariance model showed excellent fits as evidenced by nonsignificant chi-square difference tests. Multiple-group comparison also revealed no differences between the first three models, as evidenced by nonsignificant chi-square difference tests as well as fit indices which remained essentially stable. However, chi-square difference testing revealed significant differences in the strict-invariance model compared to the strong-invariance model (chi-square/df = 24,963/14, $p = .035$) and the configural model (chi-square/df = 98,757/74, $p = .029$). The CFI

Table 8. Model fit of the ESEM multiple-group analysis (landline vs. mobile) for late adulthood.

Model	ML/*df*	p_{fit} p_{diff}	CFI	TLI	RMSEA	SRMR
(1) Configural invariance	84/80	.370	.997	.992	.013	.023
Differences: (2) vs. (1)	59/50	.180	.008	.010	.007	.021
(2) Weak measurement invariance	143/130	.213	.989	.982	.020	.044
(3) vs. (2)	15/10	.139	.004	.005	.002	.004
(3) vs. (1)	74/60	.109				
(3) Strong invariance	157/140	.149	.985	.977	.022	.048
(4) vs. (3)	25/14	.035	.010	.011	.005	.016
(4) vs. (1)	99/74	.029				
(4) Strict invariance	182/154	.059	.975	.966	.027	.064
(5) vs. (4)	7/5	.207	.002	.001	.001	.001
(5) vs. (1)	106/79	.023				
(5) Strict invariance and fixed factor means	190/159	.049	.973	.965	.028	.065

Note. Older adults: $n = 504$, ten respondents with very low education were excluded from the mobile sample; ML/*df* = maximum-likelihood chi-square/degrees of freedom; p_{fit} = chi-square test to evaluate model fit; p_{diff} = chi-square difference test between two models; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standard Root Mean square Residual.

change was .01. Differences in fit indices between the strict-invariance model and the strong-invariance model also revealed no improved parsimony. The strict-invariance model (with fixed factor means) showed reasonable fit. Chi-square difference testing revealed no significant differences between the strict-invariance model with fixed factor means and the strict-invariance model (chi-square/$df = 7{,}195/5$, $p = .207$, n.s.). However, compared to the configural-invariance model, differences were significant (chi-square/$df = 105{,}952/79$, $p = .023$). Compared to the strict-invariance model fit indices remained essentially stable.

Comparing the five measurement-invariance models supports the assumption of equal factor loadings and equal item intercepts. However, the assumption of equal item uniquenesses and equal factor means is not supported. This indicates the existence of differences in measurement error in the landline sample compared to the mobile phone sample. Moreover, factor means may not be invariant across these two assessment conditions among older respondents.

4. Discussion

The results support our hypotheses H1 and H2. We were able to demonstrate that the Big Five personality dimensions were represented in both the landline and mobile phone samples. We were also able to show that these dimensions were reproduced in all three age groups of the respondents. Hence, our research results are not consistent with those of Lang et al. (2011), who questioned the suitability of the BFI-S for use in telephone surveys that include older adults. The fact that, in contrast to Lang et al.'s study, the BFI-S also yielded satisfactory results in the older adults group may be due to the fact that we used a five-point rather than a seven-point rating scale, which may have considerably reduced the mental workload of answering the questions.

However, we also found that some models for testing measurement invariance did not fit for the entire older adult population (mobile phone sample). This finding is consistent with the results of Rammstedt et al. (2010), who found that in subgroups with no, low, or intermediate secondary education, the Big Five structure could not be identified as expected. However, in samples with higher secondary education, the five-factor structure replicated clearly. According to the authors, these factor structures appear to be highly sensitive to a person's educational level.

A second methodological problem concerning the Big Five personality dimensions is acquiescence response bias (Rammstedt et al. 2013), which increases with age. Interestingly, the problems representing the Big Five factor structure for older and/or less-educated respondents occur not only in telephone samples, but also in the samples of the 2004 and 2006 ISSP who completed a self-administered questionnaire (Rammstedt et al. 2010). To reduce this effect, it might be helpful to use ipsative data.

Another limitation of our study arises from the possibility that older people's skills in participating in telephone surveys might have changed since 2005, when the data were collected on which Lang et al.'s (2011) study was based. Therefore, it is not possible to determine with certainty whether (a) the five-point rating scale actually works better, or (b) older adults have become more adept at taking phone surveys. The ideal way to test this would be to randomly assign people to a five- or seven-point scale.

In sum, we were not able to substantiate concerns expressed by Lang et al. (2011) regarding the use of the BFI-S in age-heterogeneous telephone samples in general. On the contrary, we would encourage survey researchers to make more use of such inventories because in many cases personality traits are important determinants of behavior and attitudes. Our findings lead us to conclude that in both landline as well as mobile phone surveys the application of the 15-item BFI-S works sufficiently.

5. References

Asparouhov, T. and B. Muthen. 2008. "Multilevel Mixture Models." In *Advances in Latent Variable Mixture Models*, edited by G.R. Hancock and K.M. Samuelsen, 27–51. Charlotte, NC: Information Age Publishing.

Bentler, P. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107: 238–246.

Brannick, M.T. 1995. "Critical Comments on Applying Covariance Structure Modeling." *Journal of Organizational Behavior* 16: 201–213. Doi: http://dx.doi.org/10.1002/job.4030160303.

Browne, M.W. 2001. "An Overview of Analytic Rotation in Exploratory Factor Analysis." *Multivariate Behavioral Research* 36: 111–150. Doi: http://dx.doi.org/10.1207/S15327906MBR3601_05.

Caspi, A., B.W. Roberts, and R.L. Shiner. 2005. "Personality Development: Stability and Change." *Annual Review of Psychology* 56: 453–484. Doi: http://dx.doi.org/10.1146/annurev.psych.55.090902.141913.

Chen, F.F. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling* 14: 464–504. Doi: http://dx.doi.org/10.1080/10705510701301834.

Cheung, G.W. and R.B. Rensvold. 1999. "Testing Factorial Invariance Across Groups: A Reconceptualization and Proposed New Method." *Journal of Management* 25: 1–27. Doi: http://dx.doi.org/10.1177/014920639902500101.

Dehne, M. and J. Schupp. 2007. "Persönlichkeitsmerkmale im Sozio-ökonomischen Panel (SOEP) – Konzept, Umsetzung und empirische Eigenschaften". *DIW Research Notes* 26. Berlin: DIW Berlin.

Gabler, S. and Ö. Ayhan. 2007. "Gewichtung bei Erhebungen im Festnetz und über Mobilfunk: Ein Dual-Frame Ansatz". In *Mobilfunktelefonie – Eine Herausforderung für die Umfrageforschung*, edited by S. Gabler and S. Häder, 39–45, ZUMA-Nachrichten Spezial, Vol. 13. Mannheim: GESIS.

Gabler, S., and S. Häder. 2002. "Idiosyncrasies in Telephone Sampling – The Case of Germany." *IJPOR* 14: 339–345. Doi: http://dx.doi.org/10.1093/ijpor/14.3.339.

Gabler, S., S. Haeder, I. Lehnhoff, and E. Mardian. 2012. "Weighting for Unequal Inclusion Probabilities and Nonresponse in Dual Frame Telephone Surveys." In *Telephone Surveys in Europe*, edited by S. Häder, M. Häder, and M. Kühne, 147–168. Heidelberg: Springer Verlag.

Gerlitz, J. and J. Schupp. 2005. "Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP". *DIW Research Notes* 4. Berlin: DIW Berlin.

Goebel, J., M. Grabka, P. Krause, M. Kroh, R. Pischner, I. Sieber, and M. Spiess. 2008. "Mikrodaten, Gewichtung und Datenstruktur der Längsschnittstudie Sozio-oekonomisches Panel (SOEP)." In *Vierteljahreshefte zur Wirtschaftsforschung*, Vol. 3, edited by J. Frick, O. Groh-Samberg, J. Schupp and K. Spiess, 77–109. Berlin: Duncker & Humblot. Doi: http://dx.doi.org/10.3790/vjh.77.3.77.

Goldberg, L.R., D. Sweeney, P.F. Merenda, and J.E. Hughes. 1998. "Demographic Variables and Personality: The Effects of Gender, Age, Education, and Ethnic/Racial Status on Self-Descriptions of Personality Attributes." *Personality and Individual Differences* 24: 393–403. Doi: http://dx.doi.org/10.1016/S0191-8869(97)00110-4.

Gosling, S., P. Rentfrow, and W. Swann. 2003. "A Very Brief Measure of the Big Five Personality Domains." *Journal of Research in Personality* 37: 504–528. Doi: http://dx.doi.org/10.1016/S0092-6566(03)00046-1)00046-1.

Häder, M. 2012. "Data Quality in Telephone Surveys via Mobile and Landline Phone." In *Telephone Surveys in Europe*, edited by S. Häder, M. Häder, and M. Kühne, 247–262. Heidelberg: Springer.

Häder, M. and M. Kühne. 2010. "Mobiltelefonerfahrung und Antwortqualität bei Umfragen." *Methoden, Daten, Analysen* 4: 105–112.

Häder, S., I. Lehnhoff, and E. Mardian. 2010. "Mobile Phone Surveys: Empirical Findings from a Research Project." *ASK. Society. Research. Methods* 19: 3–19.

Heaven, P. and S. Bucci. 2001. "Right-Wing Authoritarianism, Social Dominance Orientation and Personality: An Analysis Using the IPIP Measure." *European Journal of Personality* 15: 49–56. Doi: http://dx.doi.org/10.1002/per.389.

Hopwood, C.J. and M.B. Donnellan. 2010. "How Should the Internal Structure of Personality Inventories be Evaluated?" *Personality and Social Psychology Review* 14: 332–346. Doi: http://dx.doi.org/10.1177/1088868310361240.

Hu, L. and P.M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling* 6: 1–55. Doi: http://dx.doi.org/10.1080/10705519909540118.

Koch, A. 1998. "Wenn 'mehr' nicht gleichbedeutend mit 'besser' ist: Ausschöpfungsquoten und Stichprobenverzerrungen in allgemeinen Bevölkerungsumfragen." *ZUMA-Nachrichten* 42: 66–90.

Kühne, M., M. Häder, and T. Schlinzig. 2009. "Mode-Effekte bei telefonischen Befragungen über das Festnetz und den Mobilfunk: Auswirkungen auf die Datenqualität." In *Umfrageforschung. Grenzen und Herausforderung*, edited by M. Weichbold, C. Wolf, and J. Bacher, 45–62. Wiesbaden: VS Verlag für Sozialwissenschaften.

Lang, F.R., O. Lüdtke, and J.B. Asendorpf. 2001. "Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen." *Diagnostica* 47: 111–121. Doi: http://dx.doi.org/10.1026//0012-1924.47.3.111.

Lang, F.R., D. John, O. Lüdtke, J. Schupp, and G.B. Wagner. 2011. "Short Assessment of the Big Five: Robust Across Survey Methods Except Telephone Interviewing." *Behavior Research Methods* 43: 548–567. Doi: http://dx.doi.org/10.3758/s13428-011-0066-z.

Little, R.J.A. 1988. "A Test of Missing Completely at Random for Multivariate Data With Missing Values." *Journal of the American Statistical Association* 83: 1198–1202. Doi: http://dx.doi.org/10.2307/2290157.

Marsh, H.W., J.R. Balla, and R.P. McDonald. 1988. "Goodness of Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size." *Psychological Bulletin* 103: 391–410. Doi: http://dx.doi.org/10.1037//0033-2909.103.3.391.

Marsh, H.W., K.-T. Hau, and Z. Wen. 2004. "In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers of Overgeneralizing Hu and Bentler's (1999) findings." *Structural Equation Modeling* 11: 320–341. Doi: http://dx.doi.org/10.1207/s15328007sem1103_2.

Marsh, H.W., B. Muthén, A. Asparouhov, O. Lüdtke, A. Robitzsch, A.J.S. Morin, and U. Trautwein. 2009. "Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching." *Structural Equation Modeling* 16: 439–476. Doi: http://dx.doi.org/10.1080/10705510903008220.

Marsh, H.W., O. Lüdtke, B. Muthén, T. Asparouhov, A.J.S. Morin, U. Trautwein, and B. Nagengast. 2010. "A New Look at the Big Five Factor Structure Through Exploratory Structural Equation Modeling." *Psychological Assessment* 22: 471–491. Doi: http://dx.doi.org/10.1037/a0019227.

Marsh, H.W., B. Nagengast, and A.J.S. Morin. 2013. "Measurement Invariance of Big-Five Factors Over the Life Span: ESEM Test of Gender, Age, Plasticity, Maturity, and La Dolce Vita Effects." *Developmental Psychology* 49: 1194–1218. Doi: http://dx.doi.org/10.1037/a0026913.

Muthén, L.K. and B.O. Muthén. 1998–2012. *Mplus User's Guide*, 7th ed. Los Angeles, CA: Muthén and Muthén.

Rammstedt, B. 2007a. "Welche Vorhersagekraft hat die individuelle Persönlichkeit für inhaltliche sozialwissenschaftliche Variablen?" *ZUMA-Arbeitsbericht* 2007/01.

Rammstedt, B. 2007b. "The 10-Item Big Five Inventory (BFI-10): Norm Values and Investigation of Socio-Demographic Effects Based on a German Population Representative Sample." *European Journal of Psychological Assessment* 23: 193–201. Doi: http://dx.doi.org/10.1027/1015-5759.23.3.193.

Rammstedt, B., L.R. Goldberg, and I. Borg. 2010. "The Measurement Equivalence of Big Five Factor Markers for Persons with Different Levels of Education." *Journal of Research in Personality* 44: 53–61. Doi: http://dx.doi.org/10.1016/j.jrp. 2009.10.005.

Rammstedt, B. and O.P. John. 2007. "Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German." *Journal of Research in Personality* 41: 203–212. Doi: http://dx.doi.org/10.1016/j.jrp. 2006.02.001.

Rammstedt, B., C. Kemper, M. Klein, C. Beierlein, and A. Kovaleva. 2013. "A Short Scale for Assessing the Big Five Dimensions of Personality: 10 Item Big Five Inventory (BFI-10)." *Methods, Data, Analyses* 7: 233–249. Doi: http://dx.doi.org/10.12758/mda. 2013.013.

Saucier, G. 2000. "Isms and the Structure of Social Attitudes." *Journal of Personality and Social Psychology* 78: 366–385. Doi: http://dx.doi.org/10.1037//0022-3514.78.2.366.

Schneiderat, G. and T. Schlinzig. 2012. "Mobile- and Landline-Onlys in Dual-Frame-Approaches: Effects on Sample Quality." In *Telephone Surveys in Europe*, edited by S. Häder, M. Häder, and M. Kühne, 122–143. Heidelberg: Springer.

Taylor, M.F., J. Brice, N. Buck, and E. Prentice-Lane. 2010. *British Household Panel Survey User Manual: Volume A. Introduction, Technical Report, and Appendices*. Colchester: University of Essex.

Toepoel, V. and P. Lugtig. 2015. "Online Surveys are Mixed-Device Surveys. Issues Associated With the Use of Different (Mobile) Devices in Web Surveys." *Methods, Data, Analyses* 9: 155–162. Doi: http://dx.doi.org/10.12758/mda.2015.009.

Tucker, L. and C. Lewis. 1973. "A Reliability Coefficient for Maximum Likelihood Factor Analysis." *Psychometrika* 38: 1–10. Doi: http://dx.doi.org/10.1007/BF02291170.

Van Hiel, A., M. Pandelaere, and B. Duriez. 2004. "The Impact of Need for Closure on Conservative Beliefs and Racism: Differential Mediation by Authoritarian Submission and Authoritarian Dominance." *Personality and Social Psychology Bulletin* 30: 824–837. Doi: http://dx.doi.org/10.1177/0146167204264333.

West, S.G., J.F. Finch, and P.J. Curran. 1995. "Structural Equation Modeling with Nonnormal Variables: Problems and Remedies." In *Structural Equation Modeling: Concepts, Issues and Applications*, edited by R.H. Hoyle, 37–55. Thousand Oaks, CA: Sage.

Ximénez, C. 2006. "A Monte Carlo Study of Recovery of Weak Factor Loadings in Confirmatory Factor Analysis." *Structural Equation Modeling* 13: 587–614. Doi: http://dx.doi.org/10.1207/s15328007sem1304_5.

Received January 2015
Revised February 2016
Accepted February 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 619–642, http://dx.doi.org/10.1515/JOS-2016-0032

Accuracy of Mixed-Source Statistics as Affected by Classification Errors

Arnout van Delden[1], *Sander Scholtus*[1], *and Joep Burger*[2]

Publications in official statistics are increasingly based on a combination of sources. Although combining data sources may result in nearly complete coverage of the target population, the outcomes are not error free. Estimating the effect of nonsampling errors on the accuracy of mixed-source statistics is crucial for decision making, but it is not straightforward. Here we simulate the effect of classification errors on the accuracy of turnover-level estimates in car-trade industries. We combine an audit sample, the dynamics in the business register, and expert knowledge to estimate a transition matrix of classification-error probabilities. Bias and variance of the turnover estimates caused by classification errors are estimated by a bootstrap resampling approach. In addition, we study the extent to which manual selective editing at micro level can improve the accuracy. Our analyses reveal which industries do not meet preset quality criteria. Surprisingly, more selective editing can result in less accurate estimates for specific industries, and a fixed allocation of editing effort over industries is more effective than an allocation in proportion with the accuracy and population size of each industry. We discuss how to develop a practical method that can be implemented in production to estimate the accuracy of register-based estimates.

Key words: Accuracy; editing; administrative data; short-term business statistics; bootstrap resampling.

1. Introduction

Publications in official statistics are increasingly based on a combination of sources, for instance, a sample survey combined with an administrative source. The combination of data sources sometimes results in a situation where observations are available for nearly the complete target population, but that does not imply that the outcomes are error free. In fact, numerous error types may occur, as exhibited by the total survey error framework for sample surveys (Biemer and Lyberg 2003; Groves et al. 2009), adapted for administrative data by Zhang (2012a). We believe that it is important for NSIs to quantify the implications of those errors for the accuracy of statistical outcomes based on mixed sources, because NSIs aim to publish information of sufficient quality for users.

[1] Statistics Netherlands, Department of Process Development and Methodology, Henri Faasdreef 312, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Email: a.vandelden@cbs.nl and s.scholtus@cbs.nl.
[2] Statistics Netherlands, Department of Process Development and Methodology, CBS-weg 11, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. Email: j.burger@cbs.nl.
Acknowledgments: We thank Arjen de Boer and Danny van Elswijk for providing the raw data and Ton Bonné, Lei Dirrix, Willem Heijnen, Marian Immerzeel, John Spring in 't Veld, and Ivonne Valent for providing their help with the audit sample and Harm Jan Boonstra and Ton de Waal for useful comments on earlier drafts.

Knowledge on the effect of errors on the accuracy of mixed-source statistics is also useful for operational decisions, for instance in the editing process. Time, costs, and quality constraints all play a role in the decision how many units are edited manually in a statistical process to improve data quality. To this end, 'selective editing' methods have been developed (de Waal et al. 2011). These methods aim to limit manual editing by focussing on units with a high risk of influential errors, where an 'influential error' is defined as one "that has a considerable effect on the publication figures" (de Waal et al. 2011). In addition to the influence of records on the *values* of the publication figures, the effect on the *accuracy* of the figures is also important.

Estimating the effect of nonsampling errors on the accuracy of estimates in practical situations is not very straightforward as yet. Depending on the complexity of the combined data sources and the type of nonsampling error, sometimes analytical approaches are possible (Burger et al. 2015; Zhang 2012b). In cases with complicated error structures or when the effects of different processing and estimation steps are taken into account, this may no longer be possible. Bryant and Graham (2015) estimated the uncertainty caused by nonsampling errors using a Bayesian approach. Burger et al. (2015) treated a simplified situation where they did a sensitivity analysis on classification errors for which they used both an analytical and a parametric bootstrap approach. A bootstrap approach can also be applied in more complex situations where an analytical solution cannot be found. In the current article, we proceed with this work towards a more realistic modelling of the error structure.

To illustrate the method, we look at a case study on the estimation of the quarterly turnover of the 'car trade' based on a combination of survey and administrative data. The figures are classified by (groupings of) economic activity according to NACE rev 2, henceforth referred to as industry codes. Determining the correct activity code of economic units is often rather difficult and prone to errors (e.g., Christensen 2008). Reasons are that the surveyed units often have a mixture of economic activities, that activities change over time but those changes are often not reported to the relevant administrative organisations, and that the distinction between different codes is sometimes fuzzy. Previous work on the same case study by Burger et al. (2015) suggested that the publication figures are rather sensitive to classification errors.

The current article studies classification errors for two purposes: 1) to quantify their effect on the accuracy of statistical figures, and 2) to show if and how we can use this information to improve the accuracy of the estimates by selective manual editing. The current article provides key extensions to Burger et al. (2015) on both topics. Concerning the first topic, we estimate the accuracy (due to classification errors) of published figures under more realistic conditions, rather than providing a sensitivity analysis as was done in Burger et al. (2015). Concerning the second topic, we experiment with selective editing aided by the estimated classification-error model.

The remainder of the article is organised as follows. Section 2 presents a theory to estimate accuracy and model classification errors. Section 3 introduces the case study. Results on the estimated accuracy are given in Section 4. Next, Section 5 estimates the effect of supplementary editing on the estimated accuracy. Finally, Section 6 discusses the results and gives suggestions for further research. The Appendix describes a theory for correcting the bias in the bootstrap estimates of accuracy.

2. Theory to Estimate Accuracy and Model Classification Errors

2.1. Estimating Accuracy for Given Classification Errors

Consider a population of units ($i = 1, \ldots, N$) that is divided into industries based on economic activity as derived in a business register (BR). Denote the total set of industries by $\mathcal{H}_{\text{full}}$. Each unit (enterprise) i has an unknown true industry code $s_i = g$ and an observed industry code $\hat{s}_i = h$, where $g, h \in \mathcal{H}_{\text{full}}$. We suppose that for each unit random classification errors occur, independently across units, according to a known (or previously estimated) transition matrix $\mathbf{P}_i = (p_{ghi})$, with $p_{ghi} = P(\hat{s}_i = h | s_i = g)$. Note that – following, for example, Kuha and Skinner (1997) – we consider the true industry code as fixed and the observed industry code as stochastic.

In this article, we consider the relatively simple case where classification errors are the only errors that affect the publication figures. We are interested in the total turnover per industry: $Y_h = \sum_{i=1}^{N} a_{hi} y_i$, with

$$a_{hi} = I(s_i = h) = \begin{cases} 1 & \text{if } s_i = h, \\ 0 & \text{if } s_i \neq h. \end{cases}$$

In practice, Y_h is estimated by $\hat{Y}_h = \sum_{i=1}^{N} \hat{a}_{hi} y_i$, with $\hat{a}_{hi} = I(\hat{s}_i = h)$. Now we would like to assess the bias and variance of $\hat{Y}_h$ as an estimator for Y_h, that is,

$$B\left(\hat{Y}_h\right) = E\left(\hat{Y}_h - Y_h\right) = \sum_{i=1}^{N} \{E(\hat{a}_{hi}) - a_{hi}\} y_i, \tag{1}$$

$$V\left(\hat{Y}_h\right) = \sum_{i=1}^{N} V(\hat{a}_{hi}) y_i^2, \tag{2}$$

where in (2) we used the assumption of independent classification errors across units.

Given the transition matrix $\mathbf{P}_i$, it is not too difficult to derive analytical expressions for the bias and variance of $\hat{Y}_h$ in the situation considered here (Appendix and Burger et al. 2015). Here, we focus on an alternative approach to estimate the accuracy and use bootstrap resampling. In future applications we would like to assess the bias and variance of estimates due to other nonsampling errors besides classification errors, such as measurement, linkage, and coverage errors, as well as combinations thereof (van Delden et al. 2014). The bootstrap method can be generalised to handle these more complex situations.

In the bootstrap approach, following Burger et al. (2015), we apply the transition matrix $\mathbf{P}_i$ to the observed $\hat{s}_i$, which results in a new industry-assignment variable, denoted by $\hat{s}_i^*$. That is to say, we consider realisations of the alternative classification-error model given by

$$P\left(\hat{s}_i^* = h | \hat{s}_i = g\right) \equiv P\left(\hat{s}_i = h | s_i = g\right) = p_{ghi}. \tag{3}$$

We also define: $\hat{a}_{hi}^* = I\left(\hat{s}_i^* = h\right)$. By repeating this procedure R times (for some large R), we obtain a set of so-called bootstrap replications of the estimated total turnover in

industry h: $\hat{Y}^*_{hr} = \sum_{i=1}^{N} \hat{a}^*_{hir} y_i$ $(r = 1, \ldots, R)$. The bootstrap bias and variance are then estimated as follows (Efron and Tibshirani 1993):

$$\hat{B}^*_R(\hat{Y}_h) = m_R\left(\hat{Y}^*_h\right) - \hat{Y}_h, \tag{4}$$

$$\hat{V}^*_R(\hat{Y}_h) = \frac{1}{R-1} \sum_{r=1}^{R} \left\{ \hat{Y}^*_{hr} - m_R\left(\hat{Y}^*_h\right) \right\}^2. \tag{5}$$

with $m_R\left(\hat{Y}^*_h\right) = \frac{1}{R} \sum_{r=1}^{R} \hat{Y}^*_{hr}$. Details about the assumptions and computations can be found in Burger et al. (2015).

In practice, the total number of industries in $\mathcal{H}_{\text{full}}$ is large – about 300 in the Netherlands – and often one will be interested only in the accuracy of turnover estimates for a limited subset of target industries, rather than for all industries at once. In the remainder of this article we use $\mathcal{H}$ to denote the set of target industries, for which we want to compute (4) and (5), and $\mathcal{H}_{\text{full}} \backslash \mathcal{H}$ to denote the other industries.

2.2. *Modelling Classification Errors*

2.2.1. Introduction to Modelling Classification Errors

To apply the above bootstrap method, we first need to estimate the matrix of classification-error probabilities. For simplicity, Burger et al. (2015) introduced three assumptions for this that we want to relax here. First, they assumed that the subset of target industries forms a 'closed' population, with only misclassifications among this subset. In terms of Burger et al.'s case study of the car trade, they assumed misclassifications only among the nine underlying industries within the car trade but no misclassifications between the car trade and other types of industry. Secondly, they assumed that the probabilities of misclassification are the same for all units in all industries; that is, $\mathbf{P}_i = \mathbf{P}$ and all diagonal elements of $\mathbf{P}$ are equal. Thirdly, they assumed that misclassified units are distributed uniformly over the remaining industries; that is, all off-diagonal elements of $\mathbf{P}$ are also equal. In the current article we use a more realistic approach. We still assume random classification errors, but we now estimate the transition probabilities p_{ghi} by means of an audit sample.

Suppose that each unit in the population has a transition matrix $\mathbf{P}_i$ with elements p_{ghi} as in Table 1, where $g, h \in \{1, \ldots, H\}$ stands for the target set of industries $\mathcal{H}$ for which

Table 1. Transition probabilities (subscript i omitted).

	Observed industry				
True industry	1	2		H	$H+1$
1	p_{11}	p_{12}	$\cdots$	p_{1H}	$p_{1,H+1}$
2	p_{21}	p_{22}	$\cdots$	p_{2H}	$p_{2,H+1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
H	p_{H1}	p_{H2}	$\cdots$	p_{HH}	$p_{H,H+1}$
$H+1$	$p_{H+1,1}$	$p_{H+1,2}$	$\cdots$	$p_{H+1,H}$	$p_{H+1,H+1}$

we want to estimate the accuracy of the totals $\hat{Y}_h$, and industry $H+1$ represents the union of all industries outside that target set, that is, the union of all industries in $\mathcal{H}_{\text{full}} \backslash \mathcal{H}$. In our case (see Section 3), we are interested in estimating totals of $H = 9$ industries in the car trade; the other industries outside the car trade but within the total set of possible NACE codes are summarised as a tenth 'industry'.

To reduce the number of parameters to estimate, we split up the estimation of $\mathbf{P}_i$ into three parts: 1) the diagonal elements $\hat{p}_{ggi}$ with $g \in \{1, \ldots, H\}$, 2) the off-diagonal elements $\hat{p}_{ghi}$ ($g \neq h$ and $g, h \in \{1, \ldots, H\}$), and 3) the elements of row and column $H+1$. To begin with, we ignore the last row and column of the matrix and focus on the submatrix with $g, h \in \{1, \ldots, H\}$. We separate the estimation of the diagonal and nondiagonal elements as follows. Consider the contingency table of s_i and $\hat{s}_i$ in the population and let N_{gh} denote the stochastic number of units in cell (g, h). The corresponding expected value M_{gh} is given by

$$M_{gh} = \sum_{i=1}^{N} P(\hat{s}_i = h | s_i = g) \cdot I(s_i = g). \tag{6}$$

Denote the probability that unit i is classified correctly as $\pi_i = P(\hat{s}_i = g | s_i = g)$. The transition probabilities for $g \neq h$ are then given by:

$$\begin{aligned} P\big(\hat{s}_i = h | s_i = g\big) &= P\big(\hat{s}_i = h, \hat{s}_i \neq g | s_i = g\big) \\ &= P\big(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\big) \cdot P\big(\hat{s}_i \neq g | s_i = g\big) \\ &= P\big(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\big) \cdot (1 - \pi_i) \end{aligned} \tag{7}$$

where $P\big(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\big)$ is the conditional probability that unit i receives the code $\hat{s}_i = h$, given that this is a wrong code ($s_i = g \neq h$). From Equations (6) and (7) it follows that

$$\begin{aligned} M_{gg} &= \sum_{i=1}^{N} \pi_i I(s_i = g), \\ M_{gh} &= \sum_{i=1}^{N} (1 - \pi_i) P\big(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\big) I(s_i = g), \quad (g \neq h). \end{aligned} \tag{8}$$

We now introduce separate models for estimating the diagonal probabilities π_i and the conditional off-diagonal probabilities $P\big(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\big)$.

2.2.2. Modelling the Diagonal Probabilities

To estimate the diagonal elements of the $H \times H$ submatrix, we introduce the assumption that the probabilities π_i can be modelled by a logistic regression (McCullagh and Nelder 1989) on a number of independent variables. We estimate the parameters of the model by taking an audit sample of size $n \ll N$ from the population, for which both $\hat{s}_i$ and s_i are observed.

2.2.3. Modelling the Off-Diagonal Probabilities

Similarly to the diagonal probabilities, the off-diagonal probabilities might in reality also vary with i. However, the off-diagonal probabilities concern a large number of parameters

and it would lead to a lack of degrees of freedom in the audit data if we also modelled those as a function of independent variables. To estimate the off-diagonal elements of the $H \times H$ submatrix, we therefore introduce the additional assumption that, given that a unit is misclassified, the conditional off-diagonal probabilities are independent of i:

$$P\left(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\right) = \frac{P\left(\hat{s}_i = h | s_i = g\right)}{1 - \pi_i} \equiv \psi(g,h), \quad (g \neq h). \tag{9}$$

From (8) it now follows that

$$M_{gh} = \psi(g,h) \sum_{i=1}^{N} (1 - \pi_i) I(s_i = g) = \psi(g,h)\left(M_{g+} - M_{gg}\right), \quad (g \neq h) \tag{10}$$

where $M_{g+} = N_{g+} = \sum_{i=1}^{N} I(s_i = g)$ stands for a fixed but unknown row total. Hence we obtain:

$$\psi(g,h) = \frac{M_{gh}}{M_{g+} - M_{gg}}, \quad (g \neq h). \tag{11}$$

Note that, within each row, we have $\sum_{h \neq g} \psi(g,h) = 1$.

Now suppose that, in our audit sample, we count n_{gh} units in cell (g,h). In principle, we could estimate $\psi(g,h)$ by substituting these observed counts directly into Expression (11). However, this would yield unreliable estimates in practice, unless the audit sample was very large or H was very small. Therefore, we propose reducing the number of parameters further by using a log-linear model.

Denote: $m_{gh} = E(n_{gh})$. The information in the audit sample for the off-diagonal cells can be described completely by the following saturated log-linear model:

$$\log m_{gh} = u + u_{1(g)} + u_{2(h)} + u_{12(gh)}, \quad (g \neq h), \tag{12}$$

with the identifying restrictions $\sum_{g=1}^{H} u_{1(g)} = \sum_{h=1}^{H} u_{2(h)} = \sum_{g=1}^{H} u_{12(gh)} = \sum_{h=1}^{H} u_{12(gh)} = 0$. Log-linear models can be used to describe and test effects in contingency tables (Bishop et al. 1975).

Clerical reviewers know from their practical experience that some specific misclassifications of NACE codes occur more often than others. To reduce the number of parameters to estimate, we have asked experts to appoint each off-diagonal cell to a cluster $q \in \{1, \ldots, Q\}$, where cells within the same cluster are supposed to have a comparable probability of misclassification and Q is small compared to the total number of off-diagonal cells. Denote $\delta_q(g,h) \in \{0,1\}$ as the variable indicating whether cell (g,h) is appointed to cluster q. Note that $\sum_{q=1}^{Q} \delta_q(g,h) = 1$ for all $g,h \in \{1, \ldots, H\}$ with $g \neq h$. Instead of the saturated model, we now use the following log-linear model:

$$\log m_{gh} = u + u_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h) u_{3(q)}, \quad (g \neq h), \tag{13}$$

using the identifying restrictions $\sum_{h=1}^{H} u_{2(h)} = \sum_{q=1}^{Q} u_{3(q)} = 0$. This model can be understood as follows. Firstly, the number of units may differ between industries, leading to different expected values m_{gh}. This is accounted for by the column effect $u_{2(h)}$ in the

model. (We have a practical reason for taking the column effect rather than the row effect; see the end of this subsection.) In addition we account for the effect of the clusters $\delta_q(g,h)$. Similarly to the sparse classification-error model by Zhang (2005), the simplifying assumptions used to derive (9) and (13) aim to provide an adequate description of the effects of the classification errors, rather than the mechanisms by which these errors arise. Note that the diagonal probabilities are close to one in most cases (see Subsection 4.1), so the assumption is therefore adequate.

Model (13) has a slightly unusual form, but it can be rewritten as a standard log-linear model with only main effects by embedding the original contingency table in a three-dimensional table with cells (g,h,q), treating all cells for which $g=h$ or $\delta_q(g,h)=0$ as structural zeros. The parameters of Model (13) may then be estimated by maximum likelihood (Bishop et al. 1975), which gives the estimated values:

$$\hat{m}_{gh} = \exp\left\{\hat{u} + \hat{u}_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h)\hat{u}_{3(q)}\right\}, \quad (g \neq h). \tag{14}$$

By substituting these values into (11), with $\hat{m}_{gg}=0$, we obtain estimates of the conditional probabilities $\psi(g,h) = \hat{m}_{gh}/\sum_{h=1}^{H}\hat{m}_{gh}(g \neq h)$.

In practice, it may be useful to draw the audit sample as a stratified sample by observed NACE code (i.e., stratified by column in the above contingency table). In that case, we need to take the sampling fractions into account when estimating the classification probabilities. Suppose that column h has a sampling fraction of n_{+h}/N_{+h}, with $n_{+h} = \sum_{g=1}^{H} n_{gh}$ and $N_{+h} = \sum_{g=1}^{H} N_{gh}$. We can estimate the population count in the cell (g,h) by $\hat{N}_{gh,model} = \hat{m}_{gh}\left(N_{+h}/n_{+h}\right)$. Multiplying the left- and right-hand sides of (14) by N_{+h}/n_{+h} yields

$$\hat{N}_{gh,model} = \exp\left\{\hat{v} + \hat{v}_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h)\hat{v}_{3(q)}\right\}, \quad (g \neq h), \tag{15}$$

with $\hat{v} = \hat{u}$, $\hat{v}_{3(q)} = \hat{u}_{3(q)}$ and $\hat{v}_{2(h)} = \hat{u}_{2(h)} + \log N_{+h} - \log n_{+h}$. The conditional probabilities $\psi(g,h)$ are now estimated by

$$\hat{\psi}_{model}(g,h) = \frac{\hat{N}_{gh,model}}{\hat{N}_{g+,model}}, \quad (g \neq h), \tag{16}$$

where $\hat{N}_{g+,model} = \sum_{h=1}^{H}\hat{N}_{gh,model}$ and $\hat{N}_{gg,model} = 0$. Under the assumption that the transition probabilities are comparable per cluster, this yields an efficient and robust estimation of $\psi(g,h)$. Note in particular that $\hat{m}_{gh}$ (and thus $\hat{N}_{gh,model}$) can be positive even when $n_{gh}=0$.

2.2.4. Modelling the Probabilities in Industry $H+1$

Recall that the set of target industries $\{1,\ldots,H\}$ is only a small subset of all possible industry types in the BR. Estimating transition probabilities among all possible industry combinations within the BR from an audit sample is not realistic, as this would require an extension of the sample to all (several hundred) industries in the NACE domain. Instead we looked into the yearly updates of the NACE codes within the BR. Denote the observed

industry of unit i in year t as $\hat{s}_i^t$. Some of the units switch between industries in year $t+1$ compared to year t: $\hat{s}_i^t = h$ and $\hat{s}_i^{t+1} = g$. We believe that there is at least some association between the (unknown) classification-error probabilities p_{ghi} and the temporal transition probabilities in the BR. The latter reflect natural changes in economic activity, and we know that administrative delays in implementing these changes are an important cause of classification errors in the BR.

Data on yearly updates showed that the distribution of temporal transitions within the BR varies considerably among the $h \in \{1, \ldots, H\}$ industries. From these data we concluded that it is not realistic to use a two-level model whereby we estimate high granular (say one-digit) NACE code transitions within the whole BR as the first level and transitions within the underlying (more detailed) industries as the second level. Instead, we used an alternative two-level model. In the first level we estimate the overall probabilities $p_{g,H+1}$ and $p_{H+1,h}$ (the last column and row of Table 1), and in the second level we model the transitions to specific industries within industry $H+1$.

For the first level, consider the row in Table 1 with the transition probabilities of units with true industry $H+1$ (outside the target set of industries) that are observed in industry $h \neq H+1$ (inside the target set of industries). Some of these units are observed in the audit sample, so these probabilities can be estimated simply by extending the log-linear model from the previous subsection to the last row. (We assume here that the off-diagonal cells in the last row and column can be appointed to one of the clusters $q \in \{1, \ldots, Q\}$ just like the other off-diagonal cells.) Next, we consider the column in Table 1 with the transition probabilities of units with true industry $h \neq H+1$ that are observed in industry $H+1$. This type of classification error cannot be observed in our audit sample. To obtain a result, we assume here that the total number of "missed units" in the true industries $\{1, \ldots, H\}$ is equal to the number of "wrong units" in the observed industries $\{1, \ldots, H\}$, that is, that $\sum_{g=1}^{H} N_{g,H+1} = \sum_{h=1}^{H} N_{H+1,h}$. Note that if this assumption does not hold, the size of the observed population in the industries $\{1, \ldots, H\}$ is structurally too high or too low.

Under the above assumption it should hold that

$$\hat{N}_{+,H+1,model} \equiv \sum_{g=1}^{H} \hat{N}_{g,H+1,model} = \sum_{h=1}^{H} \hat{N}_{H+1,h,model} \equiv \hat{N}_{H+1,+,model}. \tag{17}$$

Using this assumption, we can extend Expression (15) to $h = H+1$, where the cluster parameters $\hat{v}_{3(q)}$ are estimated on the cells (g,h) where $h \in \{1, \ldots, H\}$. In fact, we cannot estimate the effect $\hat{v}_{2(H+1)}$ in (15) directly from the audit sample. However, taking the sum of (15) with $h = H+1$ over all cells in this column we obtain:

$$\sum_{g=1}^{H} \hat{N}_{g,H+1,model} = \exp\left\{\hat{v}_{2(H+1)}\right\} \sum_{g=1}^{H} \exp\left\{\hat{v} + \sum_{q=1}^{Q} \delta_q(g, H+1)\hat{v}_{3(q)}\right\} \tag{18}$$

According to (17), the left-hand sum should be equal to $\hat{N}_{H+1,+,model}$, which is known after the estimation of the log-linear model, including row $H+1$. In that case $\hat{v} = \hat{u}$ and the cluster effects $\hat{v}_{3(q)} = \hat{u}_{3(q)}$ are also known. Hence, $\hat{v}_{2(H+1)}$ can be solved from Expression (18). Next, the underlying estimates $\hat{N}_{g,H+1,model}$ can be obtained from (15).

Finally, we can use all estimated counts $\hat{N}_{gh,model}$ to obtain estimates of $\hat{\psi}_{model}(g,h)$ as in (16). This completes the first level of the model for industry $H+1$.

2.2.5. Subdividing Units in $H+1$ Into Underlying Industries

The model from the previous subsection allows us to estimate $P\left(\hat{s}_i \in \mathcal{H}_{\text{full}} \backslash \mathcal{H} | s_i = h\right)$ and $P\left(\hat{s}_i = h | s_i \in \mathcal{H}_{\text{full}} \backslash \mathcal{H}\right)$, with $h \in \mathcal{H}$. During bootstrap simulation, these probabilities refer to the events of, respectively, a unit moving from a given target industry to an unspecified industry outside the target set ("outflow of turnover") and vice versa ("inflow of turnover"). For the purpose of quantifying the accuracy of turnover estimates for our target set of industries, it is not necessary to model the "outflow of turnover" in more detail. We do need a more detailed model for the "inflow of turnover". We applied a second-level model in which:

- the transition probabilities $P\left(\hat{s}_i = h | s_i = g\right)$ with $h \in \mathcal{H}$ and $g \in \mathcal{H}_{\text{full}} \backslash \mathcal{H}$ are proportional to the corresponding yearly transitions in the BR, that is, the transitions from $h \in \mathcal{H}$ at $t-1$ to $g \in \mathcal{H}_{\text{full}} \backslash \mathcal{H}$ at time t; and
- the turnover of those units are drawn from a log-normal distribution. For the log-normal distribution we made a distinction between units with size class 0–3 and other units.

The exact procedure for drawing from the second-level model and estimating its parameters is given in van Delden et al. (2015a).

2.3. *Bias Correction*

Burger et al. (2015) explain that $\hat{B}^*_R\left(\hat{Y}_h\right)$ in (4) is a biased estimator of $B\left(\hat{Y}_h\right)$ in (1). This can be understood, since the bootstrap replications start from the observed $\hat{s}_i = h$ rather than the true $s_i = g$ values. In the more simple situation described in Burger et al. (2015) this bias could be corrected easily. In our case it is also possible to compute an unbiased bootstrap estimator of $B\left(\hat{Y}_h\right)$; see the Appendix. In terms of the notation in the Appendix, we denote the original (biased) estimator $\hat{B}^*_R\left(\hat{Y}_h\right)$ as $\hat{B}^*_{0R}\left(\hat{Y}_h\right)$ and the corrected (unbiased) estimator by $\hat{B}^*_{1R}\left(\hat{Y}_h\right)$. A disadvantage of $\hat{B}^*_{1R}\left(\hat{Y}_h\right)$ is that it may have a large variance in practice. We therefore introduce a combined estimator, denoted by $\hat{B}^*_{\lambda R}\left(\hat{Y}_h\right)$:

$$\hat{B}^*_{\lambda R}\left(\hat{Y}_h\right) = \lambda \hat{B}^*_{1R}\left(\hat{Y}_h\right) + (1-\lambda)\hat{B}^*_{0R}\left(\hat{Y}_h\right) \tag{19}$$

where the relative weight λ is determined by minimising the mean squared error of $\hat{B}^*_{\lambda R}\left(\hat{Y}_h\right)$. The exact procedure actually involves optimal weights at a more detailed level than indicated in (19); see Expression (25) in the Appendix. More details are given in van Delden et al. (2015a). The results of our case study in Section 4 and Section 5 below were obtained using this combined bootstrap estimator for the bias.

The bootstrap variance $\hat{V}^*_R\left(\hat{Y}_h\right)$ in (5) is also a biased estimator of $V\left(\hat{Y}_h\right)$ in (2), but this bias is expected to be small in practice compared to that of $\hat{B}^*_R\left(\hat{Y}_h\right)$ (cf. van Delden et al. 2015a for more details). Therefore we did not attempt to correct this bias in our case study; the results below were obtained using Estimator (5) for the variance.

Note that our bias correction is specifically derived for classification errors affecting a level estimate, so the approach cannot be applied directly to more difficult problems

(considering, for example, a combination of classification errors and measurement errors). A more general strategy for bias correction might be based on a 'double' bootstrap method (Efron and Tibshirani 1993; Hall and Maiti 2006).

3. Case Study: Data

The case study concerns estimates of quarterly turnover levels in the industry car trade (NACE rev. 2 code 45) for the first quarter (Q1) of 2012 until Q2 of 2014. The outcomes of the car trade are subdivided into nine industries. The quarterly turnover is estimated from a mixed-source production system (e.g., van Delden and de Wolf 2013).

Turnover in the small enterprises is derived from value-added tax (VAT) data. These enterprises are referred to as the complexity class *simple units*. On 1 January 2013 there were about one million simple units in the Netherlands, of which 28,605 were classified as car traders. The remaining units are observed in a census survey. There were 8,403 such enterprises within the whole domain of economic activities and 239 within the car trade (1 January 2013). For a subset of this group, there is a special business unit at Statistics Netherlands (CBS) with centralised data collection and data editing. This concerned 2,305 enterprises within the whole domain of economic activities and 49 within the car trade. This latter subset is referred to as the complexity class *most complex units*. The other units receiving survey data but not treated by this special business unit are referred to as the *complex units.*

The quarterly outcomes are published in different releases: 30 days (flash), 60 days (early), 90 days (late) and one year (final) after the end of the reference period. The computations in the current article concern the most recent releases available. For 2012 and 2013 this concerns the final release and for Q1 and Q2 of 2014 this concerns the late release. The available microdata covered nearly the complete target population. In late releases, quarterly nonrespondents are missing, as are units that report their VAT on a yearly basis. The latter group corresponds to 2–3% of the total turnover. Missing values are imputed. In the final release, the imputed quarterly turnover values of units that report VAT on a yearly basis are calibrated upon their reported yearly turnover values. We treat imputations here as if they are observed values, that is, we do not compute the effect of the imputation process on the accuracy.

The nine industries within the car trade vary considerably in the number of enterprises, total turnover and turnover per enterprise (van Delden et al. 2015a). In the first quarter of 2013, total turnover varies from 7,749 million euros (code 45112) to 51 million euros (code 45194). The division of total turnover in the different complexity classes also varies considerably across the nine industry codes (see Figure 1; the more detailed probability classes will be explained shortly). Note that throughout the article the industry classes are ordered from the largest to the smallest total turnover per industry.

The parameters of the classification-error model were estimated using three sources:

- We took an audit sample from the population of the *simple* enterprises within the car trade that existed on 1 July 2014 according to our BR. We randomly sampled 25 enterprises from each of the nine industries. Next, the true NACE codes were determined by two experts, examining the Chamber of Commerce information and Internet data and contacting the enterprise in case of doubt.

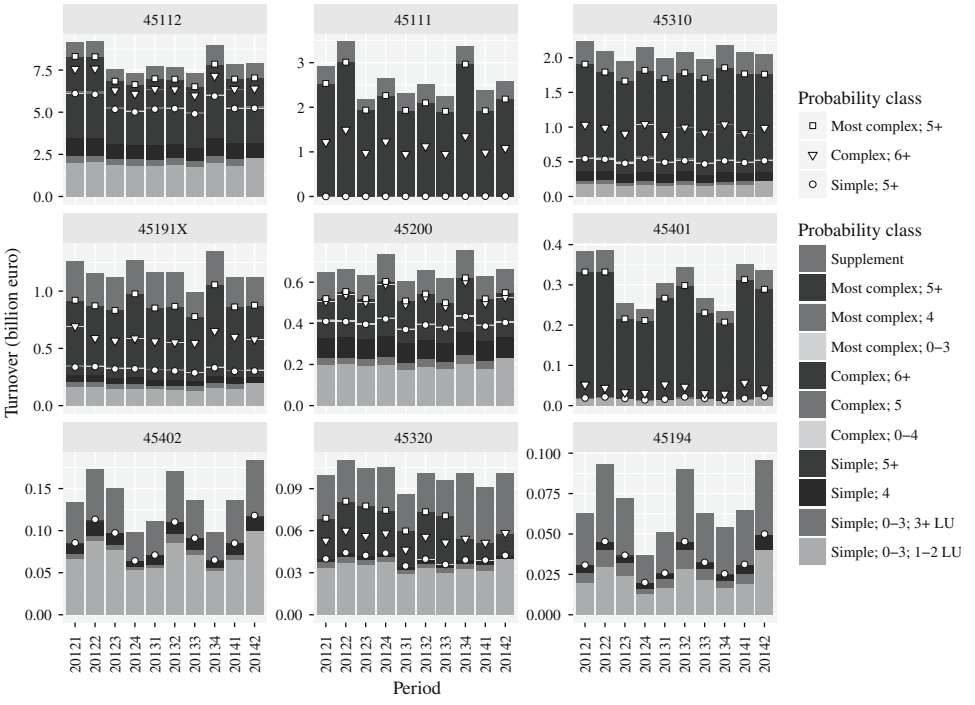


Fig. 1. Distribution of quarterly turnover among the different probability classes (Symbols appear at the upper side of the corresponding bar; Top bar is always the Supplement).

- For the *complex* and *most complex* enterprises we consulted experts at CBS who are responsible for the editing process of the car-trade industry and experts from a special business unit at CBS that deals with the large and complex units. We used expert knowledge for those enterprises, because quality studies reported in 2000 and 2003 that 97% of these enterprises were expected to have a correct three-digit NACE code (Burger et al. 2015). Therefore the transition probabilities for these units are close to 0 and 1, and estimating such small probabilities would have required a very large audit sample and too many resources. The experts were used to estimate the relative levels of classification error and the largest levels were set at five percent, which is in line with a Service Level Agreement that states that the three-digit NACE codes should be correct for 95% of the enterprises (Burger et al. 2015).
- In addition, we used data from our BR on the yearly transitions in NACE code of the enterprises for the years 2009–2014. From these data we computed the relative number of units that are observed in industry g in year t ($\hat{s}_i^t = g$) given they are observed in h in year $t-1$ $\left(\hat{s}_i^{t-1} = h\right)$ averaged over 2009–2014. The motivation behind this approach was given in Subsubsection 2.2.4. Based on the results of the temporal transitions, we have asked experts to appoint each cell (g, h) to a cluster $q \in \{1, \ldots, Q\}$, where cells within the same cluster have a comparable probability of misclassification.

Details about how these sources were used to estimate the probabilities are given in the next section.

4. Results

4.1. Estimated Probabilities

Diagonal elements. Recall that for the diagonal elements of the $H \times H$ submatrix we try to explain differences in classification-error probabilities between units from their properties. Based on consultations with experts, we identified the following variables that are available for all units in the population and that might affect the level of classification-error probabilities: observed industry, number of legal units, legal form, size class of the enterprise, and being observed in a sample survey (yes/no).

The audit sample contained no classification errors among the simple enterprises with size class 4 or larger (ten working persons or more). We therefore used the audit sample only to estimate the diagonal probabilities for the simple enterprises with size classes 0–3 (0–9 working persons).

We investigated all possible combinations of the background variables using subset selection. To compare the performance of the models, we computed the AIC and deviance values (based on log-likelihood). Table 2 displays the best-fitting models with one, two, and three predictor variables. The fourth column gives the p value of a chi-square test of decrease in deviance (cf. McCullagh and Nelder 1989). Among the three best-fitting models, the model with industry and legal units led to a significant ($p = 0.04$) increase in model fit compared to a model with only industry, whereas adding additional terms did not significantly improve model fit despite a small decrease in AIC. We also verified the model selection results by cross validation (not shown). Taking all results into account, we selected the model with industry and legal units to estimate the diagonal probabilities for the remainder of this study.

The estimated probabilities are given in the bottom two rows of Figure 2. The numbers in the labels "0–3, 4, 5, 5+ , 6+" stand for the size classes and 1–2 LU and 3+ LU stand for the number of legal units per enterprise. The diagonal probabilities of the upper nine rows of Figure 2 were based on experience of editing experts at CBS. Concerning the background variables affecting those probabilities, we limited ourselves to the complexity and size class of the units and supplementary editing (see below). From now on, the strata defined by these background variables, as shown in Figures 1 and 2, are referred to as probability classes.

The probability class 'supplement' in Figure 2 concerns the enterprises that are edited thoroughly by the statistical division at CBS responsible for the output. Enterprises that belong to the probability class 'supplement' have transition probabilities of 1.0 on the

Table 2. Three best-fitting logistic regression models for the audit sample, size classes 0–3. (Dev = Deviance; df = degrees of freedom).

	Model terms	Dev (df)	ΔDev (Δdf)	p value	AIC
0	NULL	257.27 (210)			259.27
1	Industry	170.94 (202)	86.34 (8)	<0.0001	188.94
2	Industry + Legal units	166.51 (201)	4.43 (1)	0.04	186.51
3	Industry + Legal units + Observed (Y/N)	164.36 (200)	2.15 (1)	0.14	186.36

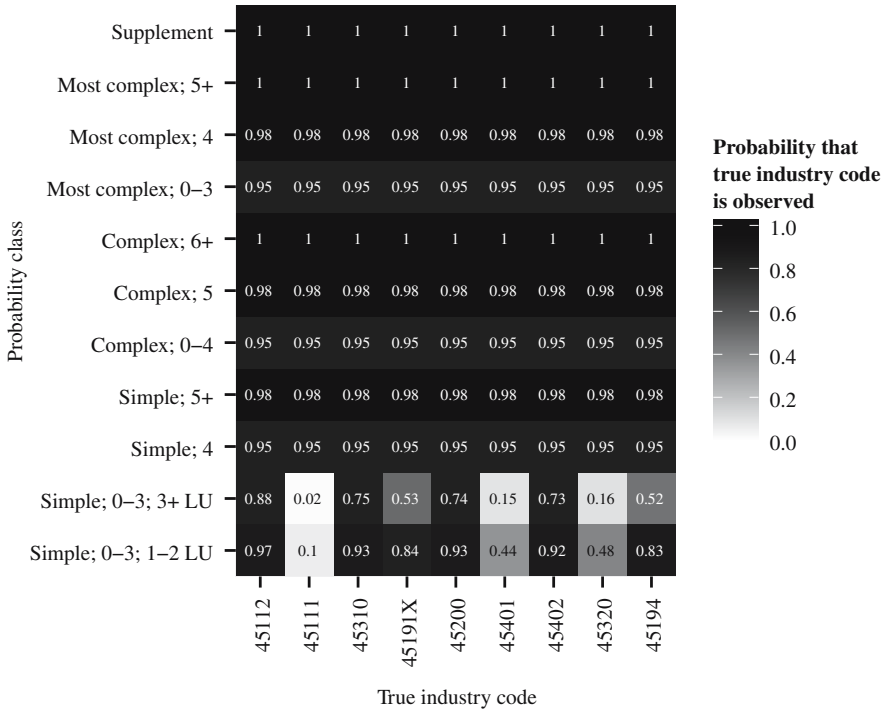


Fig. 2. Estimated transition probabilities for the diagonal elements.

main diagonal (first row of Figure 2), regardless of the further characteristics of the unit. For each target industry the size of this supplement was set to the 25 enterprises with the largest turnover, which approximately resembles the actual situation at the statistical division.

Off-diagonal elements. Using the average of the yearly transitions of the NACE codes over 2009–2014, experts appointed four clusters. Based on these $Q = 4$ clusters, we fitted a log-linear model to the off-diagonal numbers found in the audit sample, according to Equation (13). The model fitted well with a likelihood ratio of 85.92 with $p = 0.082$ at 69 degrees of freedom (df). The likelihood-ratio statistic compares the fit of the posited model to that of a saturated log-linear model, which reproduces the original table exactly (Bishop et al. 1975, 125); nonsignificant values indicate that all relevant factors are included in the model. There was one outlier that dominated the values for cluster 4. We therefore placed that outlying value in a separate fifth cluster. The model adjusted for this outlier had a likelihood ratio of 43.44 ($p = 0.991$ at 68 df). The adjusted model had expected numbers that fit the observed numbers in the audit sample better. Using those expected numbers and the sampling fractions n_{+h}/N_{+h}, the off-diagonal probabilities were estimated according to Equations (15) and (16) (Figure 3). Recall that the probabilities for the $(H+1)^{\text{th}}$ industry (column) were derived from Equation (17)–(18).

Results show that there are pairs of industries with relatively high conditional classification-error probabilities. For instance, a unit from industry 45310 (wholesale trade of motor vehicle parts and accessories) has a probability of 0.53 – given that it is

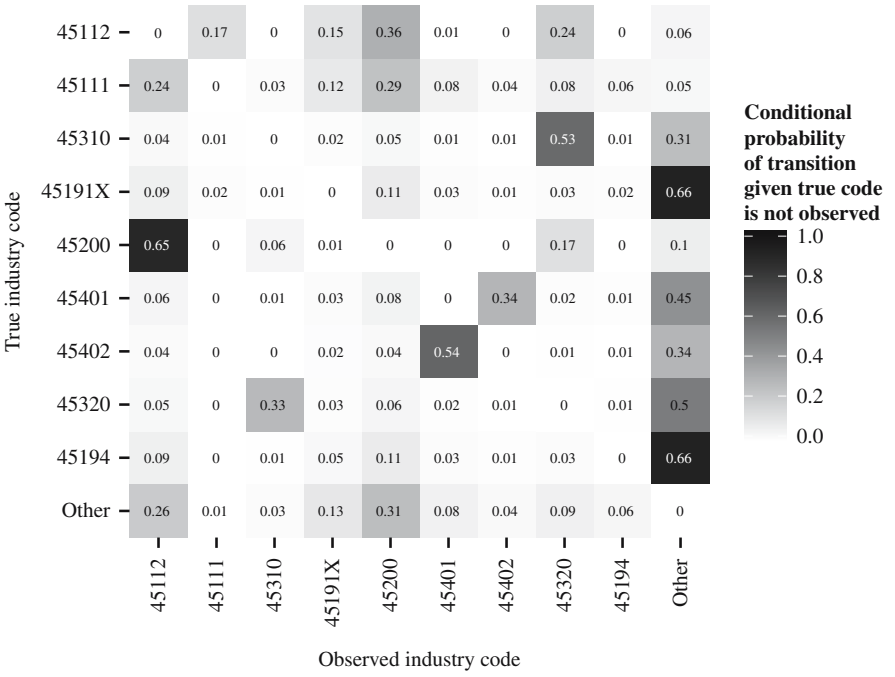


Fig. 3. Estimated transition probabilities for the off-diagonal elements. Each row adds up to 1.

misclassified – of being observed as 45320 (retail trade of motor vehicle parts and accessories). Likewise, misclassified units from industry 45320 have a probability of 0.33 of being observed as 45310. Similar high conditional probabilities of misclassification exist between the industries 45401 (wholesale trade in maintenance and repair of motor cycles) and 45402 (retail trade in maintenance and repair of motor cycles). Finally, note that in six of the nine car trade industries, misclassified units have a probability over 0.30 of being observed outside the car-trade.

Probabilities for the industries outside the car trade. We applied the approach of Subsubsection 2.2.5 to estimate the parameters of the second-level model. Details can be found in van Delden et al. (2015a).

4.2. Simulation of Accuracy

Having modelled the probabilities of classification errors for the data in our case study, we applied the bootstrap method from Subsection 2.1. We applied 10,000 bootstrap replicates. We implemented this method within the R environment for statistical computing. The code used for these simulations is available from the authors upon request.

We summarised the results in terms of the following accuracy measures, derived from (4) and (5):

- the relative bias (RB) $\hat{B}_R^*(\hat{Y}_h)/\hat{Y}_h$,
- the coefficient of variation (CV) $\sqrt{\hat{V}_R^*(\hat{Y}_h)}/\hat{Y}_h$,
- the relative root mean squared error (RRMSE) $\sqrt{\left\{\left[\hat{B}_R^*(\hat{Y}_h)\right]^2+\hat{V}_R^*(\hat{Y}_h)\right\}}/\hat{Y}_h$.

These results are shown in Figure 4 (expressed as percentages). The RRMSE varies from about 1.0% for the industries 45401 and 45310 to about 60% for industry 45320. The variance (CV) dominates in the industries 45191X, 45401, 45402, and 45194, in the other industries the bias dominates. The industries 45112 and 45310 both have a negative bias. A negative bias means that the values of bootstrap simulations $\left(\hat{Y}_{hr}^{*}\right)$ are smaller on average than the estimated value $(\hat{Y}_h)$, which in turn implies that $\hat{Y}_h$ underestimates the (unknown) true target value Y_h.

We found that industry 45320 has a very large RRMSE: on average 62% (Figure 4). This industry has a relatively large probability of classification error on the diagonal elements (Figure 2) of the complexity class "simple", and this class constitutes about one third of the total turnover in this industry (Figure 1). Industry 45111 has an even larger probability on classification errors in the complexity class "simple" (Figure 2) but does not have a large RRMSE. The latter is because the turnover of the simple enterprises in industry 45111 is very small compared to the other complexity classes (Figure 1). The RRMSE for the car trade as a whole is about 0.33% and was relatively stable over the ten periods (Figure 5). The CV was also relatively stable (about 0.29%). The RB varied most and ranged between -0.2% and -0.1%.

The RRMSE for the car trade as a whole is judged as acceptable by the owner of the production process, whereas that of industry 45320 is far too large. Fortunately, turnover levels for industry 45320 are not published separately but combined with industry 45310. The combined quarterly turnover-level estimates have an average RRMSE of 1.9% (see industry 45300 in Figure 2 by van Delden et al. 2015b). The least accurate industry that is

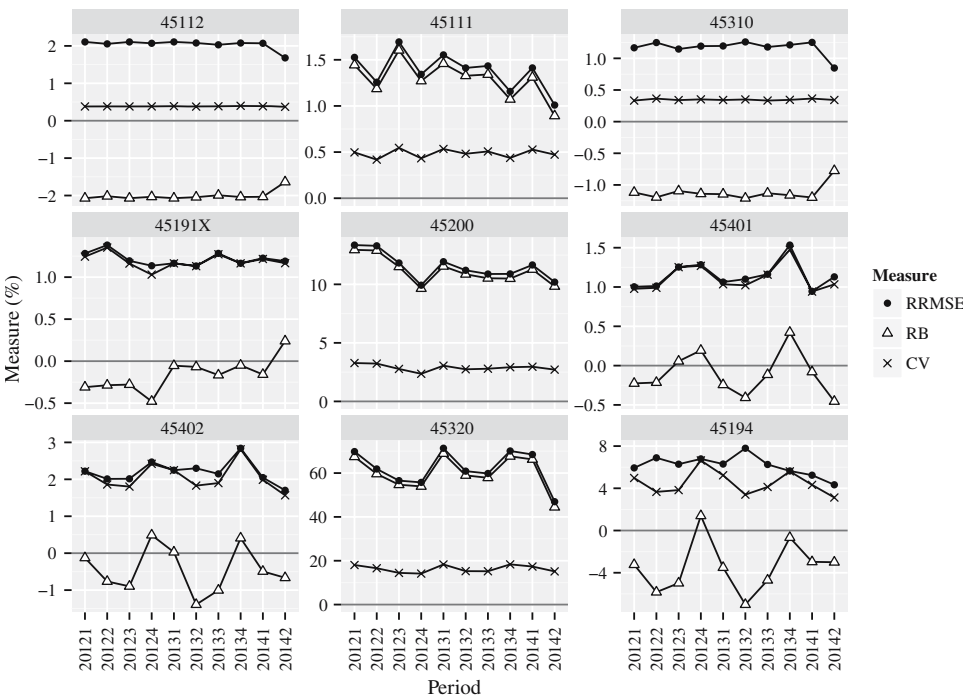


Fig. 4. RRMSE, RB and CV for ten periods per industry.

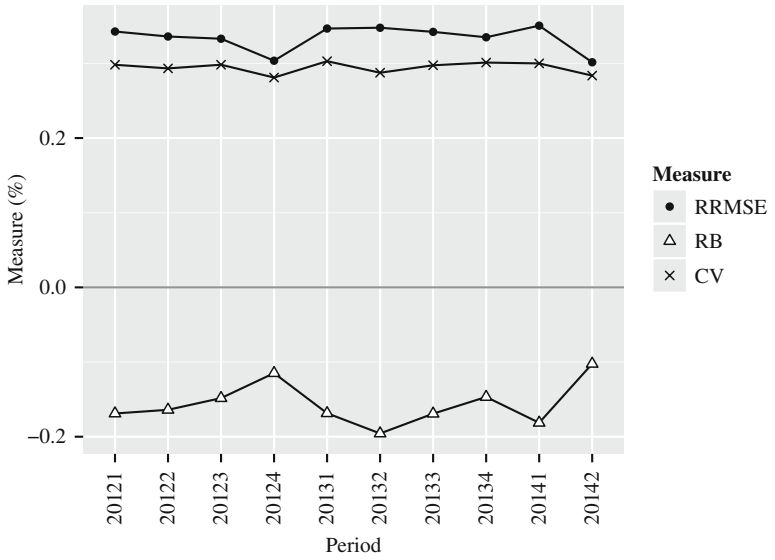


Fig. 5. RRMSE, RB and CV for ten periods for car trade as a whole.

actually published is 45200 with an RRMSE for the quarterly turnover slightly larger than ten percent. CBS aims to have a maximum uncertainty margin of three percent points on turnover levels. This means that in the car trade, an additional editing effort is needed to improve industry 45200. Below we investigate different selective editing strategies.

5. Editing Scenarios

5.1. Scenarios of Editing

We would also like to study to what extent the accuracy is improved when the editing effort is increased. An exact computation of those results is in fact only possible when we actually have a set of data that are free of classification errors. That information is needed because we need to know the true NACE code for each of the individual units. Since we do not have a data set for the whole population that is error free, we used an approximation. We assumed that with additional editing effort, those units that are checked and edited (on top of the starting situation) have a diagonal transition probability of 1, in other words a classification-error probability of zero. The edited units are called the "supplement" (Figure 1). They are called supplement because they are edited by the clerical reviewers of the production unit supplementary to the editing that is done by our central business unit on large and complex units. The exact difference between our approximation and the (true) effect of editing is explained in van Delden et al. (2015a). Nonetheless, we are convinced that our approximation is good enough to compare different editing scenarios in a qualitative way.

We compared four levels of supplementary editing, namely 0, 225, 450, and 675 edited enterprises in the car trade (relative editing effort 0, 1, 2, and 3). Since our results on accuracy were reasonably consistent over the ten quarters, we only computed the results

for one quarter: the first quarter of 2013. The second level, 225 units, corresponds reasonably well to the actual situation at CBS. We distinguished between two editing scenarios that differ in how those enterprises are allocated over the nine industries:

1. Fixed: each industry is allocated an equal number of enterprises for supplementary editing. So the four levels are equal to 0, 25, 50, and 75 enterprises per industry.
2. Pro rata: the number of enterprises to be edited per industry $\left(n_h^E\right)$ is in proportion to the product of $RMSE\left(\hat{Y}_h\right) = \sqrt{\left\{\left[\hat{B}_R^*\left(\hat{Y}_h\right)\right]^2 + \hat{V}_R^*\left(\hat{Y}_h\right)\right\}}$ and the population size per industry (N_h):

$$n_h^E = \frac{RMSE\left(\hat{Y}_h\right)N_h}{\sum_{h=1}^{H} RMSE\left(\hat{Y}_h\right)N_h} n^E, \tag{20}$$

where n^E denotes the total number of units to be selected for supplementary editing. Note that Equation (20) resembles the so-called Neyman allocation of a survey sample over its underlying industries (e.g., Cochran 1977, 98–99). Because of this analogy, one might expect the accuracy of the estimated turnover for the car trade as a whole to improve more under the pro-rata scenario than under the fixed scenario. For the $RMSE\left(\hat{Y}_h\right)$ values in Equation (20) we used the bootstrap estimates when 25 enterprises per industry were edited. Within each industry h, we selected the n_h^E units with the largest quarterly turnover for editing.

5.2. *Simulation of Editing*

The change in the accuracy measures with increased relative editing effort and the two editing scenarios showed several interesting results (Figure 6). First of all, as expected, the CV decreased with increased relative editing effort. Moreover, the absolute value of the RB) often decreased with increased editing effort. However, there were also many examples of situations where this relative bias increased. A prominent example is industry 45401, where the absolute RB clearly increased between the relative editing effort 1 and 2 for the fixed scenario, and between the relative editing effort 2 and 3 for the pro-rata scenario. The overall effect of the change of CV and RB with increased editing effort is that the RRMSE does not always decrease with increased editing effort.

We can understand this surprising phenomenon by analysing the transition of units among the industries. To this end we distinguish between inflow and outflow of turnover in industry h. Inflow of turnover occurs when units that were originally observed in another industry enter industry h in bootstrap replication r. Outflow of turnover occurs when units move to another industry from industry h where they were observed. The bias of a turnover estimate for an industry is the net result of the effects of the turnover inflow and outflow. Accordingly, when there are no classification errors (as a result of perfect editing of the units in all existing industries), the inflow and outflow components are zero and there is no bias. Likewise, when the transition probabilities happen to be such that turnover inflow and outflow to industry h are perfectly balanced, there is also no bias. In van Delden et al. (2015a) we describe and quantify the observed bias (and variance) patterns in each of the industries as the net result of inflow and outflow. In some industries we found a reasonable

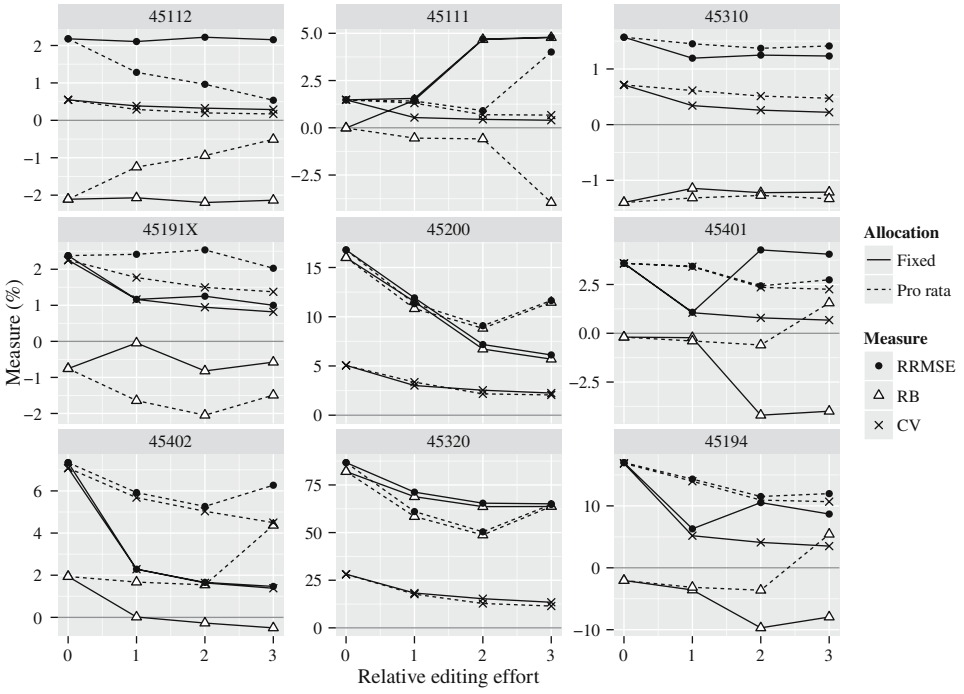


Fig. 6. Simulating the effect of editing on accuracy: the three measures (RRMSE, RB, and CV).

balance between inflow and outflow, while for others the total error is mainly determined by inflow (see Figure 4.4.5 in van Delden et al. 2015a). By changing the number of edited units in an industry, we can control the expected size of the outflow from that industry – that is, how many errors remain in industry h – but not the inflow. Due to this effect, the balance between outflow and inflow can become less favourable, leading to an increased bias.

In the above example, as the total amount of editing is increased, the absolute level of inflow in industry 45401 will decrease because the outflow from all the other car-trade industries will be reduced. Nonetheless, the balance in industry 45401 between out- and inflow on the bias becomes less favourable (van Delden et al. 2015a). In fact, with increased overall editing effort the turnover inflow in 45401 decreased at a smaller rate than the outflow, resulting in an increased bias. This effect is enhanced under the pro-rata scenario, because industry 45401 has the largest turnover inflow from industry 45402, which is more accurate than industry 45401 to begin with.

Figure 6 shows that in some industries the pro-rata scenario reduces the RRMSE further than the fixed scenario, while in other industries the opposite is the case. This is of course due to differences in editing effort per industry in the pro-rata scenario. Surprisingly, the decrease of the RRMSE for the car trade as a whole (sum of nine industries) is *larger* for the fixed scenario than for the scenario pro rata (Figure 7). This can be understood as follows. The pro-rata scenario, inspired by the Neyman allocation, assumes that the errors $\hat{Y}_h - Y_h$ are *independent* of each other. This assumption, however, does not hold in the

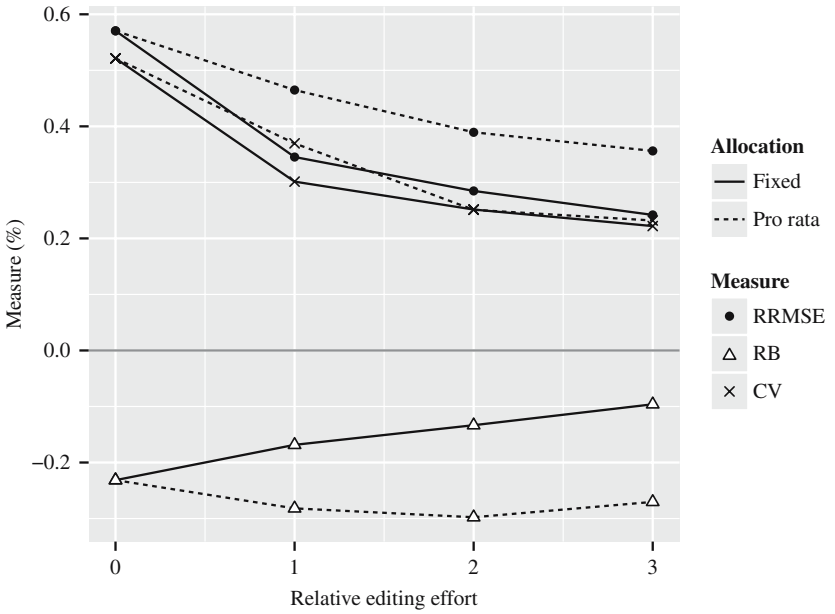


Fig. 7. Simulating the effect of editing on accuracy: overall effect on car trade.

case of classification errors, since many off-diagonal transition probabilities are larger than zero. We conclude that a simple 'fixed' scenario is more effective in reducing the overall RRMSE than the pro-rata approach. It remains to be seen whether a more efficient scenario than 'fixed' can be found, without introducing complexities that render the approach impractical.

6. Discussion

The long-term aim of our research is to develop a practical method for assessing as well as improving the accuracy of register-based estimates affected by nonsampling errors. In this article, we have estimated the accuracy of register-based outcomes for classification errors using a bootstrap method. Others have also used resampling to estimate the accuracy of statistical outcomes for certain error types, such as Zhang (2011), Lumme et al. (2015), and Chipperfield and Chambers (2015). A key challenge is to obtain good estimates of the parameters of the postulated error model.

How to handle the complex and most complex units in this respect is a difficult question. We have relied on expert knowledge when setting the diagonal probabilities of these units in our study. This is a relatively small group of units for which classification errors are rare. Furthermore, these units are not 'mutually interchangeable', given their large individual shares in the total turnover. Fundamentally, it may be asked whether a random classification-error model is appropriate for the group of complex and most complex units.

For the simple units where the model parameters can be estimated empirically, audit data can only be obtained at some additional cost. The question is then how to combine editing and estimation efficiently in practice. An option could be to use information

obtained during regular production instead of audit data to estimate the model parameters, similarly to the use of paradata in social surveys (Kreuter 2013). Maybe we can combine selective editing for the most influential units with a probability sample of less influential units. The result of this two-phase design can be used to estimate the bias and the variance of the resulting estimator as a result of the editing process (e.g., Ilves and Laitila 2009). Such an approach might also offer the possibility of extending the procedure to other industries. The development of a robust and efficient selective editing strategy for classification errors, which accounts for the in- and outflow components of the target variable due to misclassified units, is a point for future research.

Two key extensions are still needed to achieve our long-term aim. These are the extension to other types of estimators and the extension to other sources of nonsampling errors. The use of an overarching modelling framework in which the observations reflect measurements of unobserved (true) values, like in latent class models and like the Bayesian approach by Bryant and Graham (2015), might be helpful in this respect.

Appendix

Bias Correction

Correction for the Bias in $\hat{B}_R^*(\hat{Y}_h)$

We use the notation that was introduced in Section 2. In addition, let $\boldsymbol{a}_i$ denote the vector $(a_{1i}, \ldots, a_{H+1,i})^T$ that contains the values of the indicator variable $a_{hi} = I(s_i = h)$ of which one element per unit i is equal to 1. Similarly, we define $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{a}}_i^*$ on the basis of $\hat{s}_i$ and $\hat{s}_i^*$. Recall that $\mathbf{P}_i$ stands for the $(H+1)\times(H+1)$ matrix with the transition probabilities for unit i. Under the classification-error model, the expectation of $\hat{\boldsymbol{a}}_i$ for enterprise i equals $E(\hat{\boldsymbol{a}}_i) = \mathbf{P}_i^T \boldsymbol{a}_i$. Similarly it holds that $E(\hat{\boldsymbol{a}}_i^* | \hat{\boldsymbol{a}}_i) = \mathbf{P}_i^T \hat{\boldsymbol{a}}_i$. Denote the vectors with the true, observed and bootstrap values for the total turnover per industry as $\boldsymbol{y} = \sum_{i=1}^N \boldsymbol{a}_i y_i$, $\hat{\boldsymbol{y}} = \sum_{i=1}^N \hat{\boldsymbol{a}}_i y_i$ and $\hat{\boldsymbol{y}}^* = \sum_{i=1}^N \hat{\boldsymbol{a}}_i^* y_i$. Using an argument similar to that in Burger et al. (2015), the following expressions may be derived for the true bias and variance-covariance matrix of $\hat{\boldsymbol{y}}$ as an estimator for $\boldsymbol{y}$:

$$B(\hat{\boldsymbol{y}}) = E(\hat{\boldsymbol{y}}) - \boldsymbol{y} = \sum_{i=1}^N \left(\mathbf{P}_i^T - \mathbf{I}\right)\boldsymbol{a}_i y_i, \tag{21}$$

$$V(\hat{\boldsymbol{y}}) = \sum_{i=1}^N \left\{\mathrm{diag}\left(\mathbf{P}_i^T \boldsymbol{a}_i y_i^2\right) - \mathbf{P}_i^T \mathrm{diag}\left(\boldsymbol{a}_i y_i^2\right)\mathbf{P}_i\right\}, \tag{22}$$

where $\mathbf{I}$ stands for the $(H+1)\times(H+1)$-identity matrix. Here, we use the assumption that only the observed classifications $\hat{\boldsymbol{a}}_i$ may be erroneous, while the other components of $\hat{\boldsymbol{y}}$ are fixed.

In the bootstrap approach, the above bias and variance are estimated by the conditional bias and variance of $\hat{\boldsymbol{y}}*$ as an estimator for $\hat{\boldsymbol{y}}$. Letting $R \to \infty$ in Expressions (4) and (5), we would obtain:

$$\hat{B}_\infty^*(\hat{\boldsymbol{y}}) = B(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = E(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) - \hat{\boldsymbol{y}} = \sum_{i=1}^N \left(\mathbf{P}_i^T - \mathbf{I}\right)\hat{\boldsymbol{a}}_i y_i, \tag{23}$$

$$\hat{V}^*_{\infty}(\hat{\boldsymbol{y}}) = V(\hat{\boldsymbol{y}}^*|\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} \left\{ \mathrm{diag}\left(\mathbf{P}_i^T \hat{\boldsymbol{a}}_i y_i^2\right) - \mathbf{P}_i^T \mathrm{diag}\left(\hat{\boldsymbol{a}}_i y_i^2\right) \mathbf{P}_i \right\}; \tag{24}$$

cf. Burger et al. (2015). In our case study, we did not use these analytical formulas directly. We preferred to use Monte Carlo simulation to have more flexibility in the modelling of classification errors, in particular for industry $H+1$. (Note that the sum in Expressions (23) and (24) is over all units in the BR, including all industries outside the target set.)

Focussing on the bias, we see that $E\left\{\hat{B}^*_{\infty}(\hat{\boldsymbol{y}})\right\} = \sum_{i=1}^{N} \mathbf{P}_i^T \left(\mathbf{P}_i^T - \mathbf{I}\right) \boldsymbol{a}_i y_i$. This implies that $\hat{B}^*_{\infty}(\hat{\boldsymbol{y}})$ is biased as an estimator for $B(\hat{\boldsymbol{y}})$; the same follows for $\hat{B}^*_R(\hat{\boldsymbol{y}})$ based on a finite number of replications.

Now assume that the matrix $\mathbf{P}_i^T$ can be inverted and denote its inverse as $\mathbf{Q}_i = \left(\mathbf{P}_i^T\right)^{-1}$. It follows directly that $\hat{\boldsymbol{b}}_i = \mathbf{Q}_i \hat{\boldsymbol{a}}_i$ is an unbiased estimator for $\boldsymbol{a}_i$:

$$E(\hat{\boldsymbol{b}}_i) = E(\mathbf{Q}_i \hat{\boldsymbol{a}}_i) = \mathbf{Q}_i E(\hat{\boldsymbol{a}}_i) = \mathbf{Q}_i \mathbf{P}_i^T \boldsymbol{a}_i = \boldsymbol{a}_i.$$

Similarly for $\hat{\boldsymbol{b}}^*_i = \mathbf{Q}_i \hat{\boldsymbol{a}}^*_i$ it holds that $E\left(\hat{\boldsymbol{b}}^*_i | \hat{\boldsymbol{b}}_i\right) = E\left(\hat{\boldsymbol{b}}^*_i | \hat{\boldsymbol{a}}_i\right) = \mathbf{Q}_i \mathbf{P}_i^T \hat{\boldsymbol{a}}_i = \hat{\boldsymbol{a}}_i$. Analogously to $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}*$, we can define the turnover-related vectors $\hat{\boldsymbol{z}} = \sum_{i=1}^{N} \hat{\boldsymbol{b}}_i y_i$ and $\hat{\boldsymbol{z}}* = \sum_{i=1}^{N} \hat{\boldsymbol{b}}^*_i y_i$. Now, consider the conditional bias of $\hat{\boldsymbol{z}}*$ as an estimator for $\hat{\boldsymbol{z}}$:

$$B(\hat{\boldsymbol{z}}*|\hat{\boldsymbol{z}}) = E(\hat{\boldsymbol{z}}*|\hat{\boldsymbol{z}}) - \hat{\boldsymbol{z}} = \sum_{i=1}^{N} \left\{ E\left(\hat{\boldsymbol{b}}^*_i | \hat{\boldsymbol{b}}_i\right) - \hat{\boldsymbol{b}}_i \right\} y_i = \sum_{i=1}^{N} (\hat{\boldsymbol{a}}_i - \hat{\boldsymbol{b}}_i) y_i.$$

It follows that $E\{B(\hat{\boldsymbol{z}}*|\hat{\boldsymbol{z}})\} = \sum_{i=1}^{N} \{E(\hat{\boldsymbol{a}}_i) - E(\hat{\boldsymbol{b}}_i)\} y_i = \sum_{i=1}^{N} \left(\mathbf{P}_i^T - \mathbf{I}\right) \boldsymbol{a}_i y_i = B(\hat{\boldsymbol{y}})$. Hence $B(\hat{\boldsymbol{z}}*|\hat{\boldsymbol{z}})$ is an unbiased estimator for the bias of $\hat{\boldsymbol{y}}$.

In our case study the population is divided into a limited number of probability classes (PCs) with the same transition matrix. We can exploit this to compute $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{z}}*$ in an efficient manner. Divide the population into the PCs of units $U_1, \ldots, U_K$, where the transition matrix for the k^{th} PC is denoted by $\mathbf{P}_k$, with the corresponding inverse being $\mathbf{Q}_k = \left(\mathbf{P}_k^T\right)^{-1}$. Now $\hat{\boldsymbol{z}}$ can be computed according to:

$$\hat{\boldsymbol{z}} = \sum_{i=1}^{N} \hat{\boldsymbol{b}}_i y_i = \sum_{k=1}^{K} \sum_{i \in U_k} \hat{\boldsymbol{b}}_i y_i = \sum_{k=1}^{K} \mathbf{Q}_k \sum_{i \in U_k} \hat{\boldsymbol{a}}_i y_i = \sum_{k=1}^{K} \mathbf{Q}_k \hat{\boldsymbol{y}}_k \equiv \sum_{k=1}^{K} \hat{\boldsymbol{z}}_k,$$

with $\hat{\boldsymbol{y}}_k = \sum_{i \in U_k} \hat{\boldsymbol{a}}_i y_i$ the vector of industry-turnover totals for the k^{th} PC. Analogously, $\hat{\boldsymbol{z}}*$ can be computed as $\hat{\boldsymbol{z}}* = \sum_{k=1}^{K} \hat{\boldsymbol{z}}^*_k \equiv \sum_{k=1}^{K} \mathbf{Q}_k \hat{\boldsymbol{y}}^*_k$, with $\hat{\boldsymbol{y}}^*_k = \sum_{i \in U_k} \hat{\boldsymbol{a}}^*_i y_i$. Some other practical issues related to the computation of the bootstrap estimator and its bias correction are discussed in Appendix A2 of van Delden et al. (2015a).

Similarly to the bias, the bootstrap estimator of the variance is also biased. In section A4 van Delden et al. (2015a) derive a formula for this bias, explain how it can be corrected and argue that this bias is likely to be small. We therefore did not apply the bias correction for the variance.

Adjusted Bias Correction for "increased Accuracy"

The corrected bootstrap estimator for the bias $B(\hat{\boldsymbol{z}}*|\hat{\boldsymbol{z}})$ is unbiased, but may yield inaccurate estimates of $B(\hat{\boldsymbol{y}})$ in practice. Unbiased bootstrap estimation of $B(\hat{\boldsymbol{y}})$ may come at the cost of an increased variance, to such a degree that the bias correction is not an

improvement in all cases. Results on simulated data (not shown here) suggest that the bias-corrected bootstrap estimator tends to be unstable when some of the probabilities of classification errors are large.

In fact, it turns out that when some of the diagonal probabilities in $\mathbf{P}_k$ are much smaller than 1, the so-called condition number $\text{cond}(\mathbf{P}_k^T) = ||\mathbf{P}_k^T||\ ||\mathbf{Q}_k||$ can become much larger than 1. Here, the symbol ||.|| denotes a matrix norm. Since $\hat{\mathbf{y}}_k^* = \mathbf{P}_k^T \hat{\mathbf{z}}_k^*$, it follows from a standard result in numerical analysis that

$$\left|\text{rel. change}\left(\hat{\mathbf{z}}_k^*\right)\right| \leq \text{cond}\left(\mathbf{P}_k^T\right) \times \left|\text{rel. change}\left(\hat{\mathbf{y}}_k^*\right)\right|,$$

where rel. change(.) denotes a relative change in the value of its argument (e.g., Stoer and Bulirsch 2002, 211). Hence, when $\text{cond}(\mathbf{P}_k^T)$ is large, a small uncertainty in the simulated values of $\hat{\mathbf{y}}_k^*$ can be propagated as a large uncertainty in the derived values of $\hat{\mathbf{z}}_k^*$. This provides a heuristic explanation for why the bias-corrected bootstrap estimator (based on $\hat{\mathbf{z}}_k^*$) can be less accurate than the original bootstrap estimator (based on $\hat{\mathbf{y}}_k^*$) in situations where some units have a relatively large probability of being misclassified.

In Appendix A3 of van Delden et al. (2015a) an alternative correction method is proposed that uses a combined estimator

$$\hat{\mathbf{B}}_\lambda^* = \sum_{k=1}^{K} \left\{ \lambda_k \hat{\mathbf{B}}_{1k}^* + (1 - \lambda_k) \hat{\mathbf{B}}_{0k}^* \right\}, \tag{25}$$

where $\hat{\mathbf{B}}_{0k}^* = B\left(\hat{\mathbf{y}}_k^* | \hat{\mathbf{y}}_k\right)$ and $\hat{\mathbf{B}}_{1k}^* = B\left(\hat{\mathbf{z}}_k^* | \hat{\mathbf{z}}_k\right)$ denote the original and bias-corrected bootstrap estimators of the bias of $\hat{\mathbf{y}}_k$. It is shown there how the weights $\lambda_k \in [0, 1]$ can be obtained by minimising the mean square error of the estimated bias.

7. References

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley and Sons. Doi: http://dx.doi.org/10.1002/0471458740.

Bishop, Y.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Bryant, J.R. and P. Graham. 2015. "A Bayesian Approach to Population Estimation with Administrative Data." *Journal of Official Statistics* 31: 475–487. Doi: http://dx.doi.org/10.1515/JOS-2015-0028.

Burger, J., A. van Delden, and S. Scholtus. 2015. "Sensitivity of Mixed-Source Statistics to Classification Errors." *Journal of Official Statistics* 31: 489–506. Doi: http://dx.doi.org/10.1515/jos-2015-0029.

Chipperfield, J. and R. Chambers. 2015. "Using the Bootstrap to Analyse Binary Data Obtained via Probabilistic Linkage." *Journal of Official Statistics* 31: 397–414. Doi: http://dx.doi.org/10.1515/JOS-2015-0024.

Christensen, J.L. 2008. "Questioning the Precision of Statistical Classification of Industries." Paper presented at the DRUID Conference on Entrepreneurship and Innovation, 17–20 June 2008, Copenhagen. Available at: http://www2.druid.dk/conferences/viewpaper.php?id=3419&cf=29 (accessed April 2016).

Cochran, W.G. 1977. *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons.

Delden, A. van, S. Scholtus, and J. Burger. 2015a. "Quantifying the Effect of Classification Errors on the Accuracy of Mixed-Source Statistics." Discussion Paper 2015-10. Available at: https://www.researchgate.net/publication/281450992_Quantifying_the_effect_of_classification_errors_on_the_accuracy_of_mixed-source_statistics (accessed April 2016).

Delden, A. van, S. Scholtus, and J. Burger. 2015b. "Effect of Classification Errors on the Accuracy of Business Statistics." Paper presented at the European Establishment Statistics Workshop, 7–9 September 2015, Poznan. Available at: https://enbes.wikispaces.com/file/view/Effect+of+classification+errors+on+the+accuracy+of+business+statistics+Arnout+van+Delden,+Sander+Scholtus+and+Joep+Burger.pdf (accessed April 2016).

Delden, A. van and P.P. de Wolf. 2013. "A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data." Paper presented at the New Techniques and Technologies for Statistics (NTTS) conference, 5–7 March 2013, Brussels. Available at http://ec.europa.eu/eurostat/cros/sites/crosportal/files/NTTS2013%20Proceedings_0.pdf (accessed April 2016).

Delden, A. van, S. Scholtus, P.P. de Wolf, and J. Pannekoek. 2014. "Methods to Assess the Quality of Mixed-Source Estimates." Internal report PPM-2014-09-26-ADLN-SSHS-PWOF-JPNK, Statistics Netherlands, The Hague.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC. Doi: http://dx.doi.org/10.1007/978-1-4899-4541-9.

Groves, R., F. Fowler, M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*, 2nd ed. New York: John Wiley and Sons.

Hall, P. and T. Maiti. 2006. "On Parametric Bootstrap Methods for Small Area Prediction." *Journal of the Royal Statistical Society B* 68: 221–238.

Ilves, M. and T. Laitila. 2009. "Probability-Sampling Approach to Editing." *Austrian Journal of Statistics* 38: 171–182. Doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00541.x.

Kreuter, F., ed. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. New York: John Wiley and Sons. Doi: http://dx.doi.org/10.1002/9781118596869.

Kuha, J. and C. Skinner. 1997. "Categorical Data Analysis and Misclassification." In *Survey Measurement and Process Quality*, edited by L.E. Lyberg, P.P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 633–670. New York: John Wiley and Sons. Doi: http://dx.doi.org/10.1002/9781118490013.

Lumme, S., R. Sund, A.H. Leyland, and I. Keskmäki. 2015. "A Monte Carlo Method to Estimate the Confidence Intervals for the Concentration Index Using Aggregated Population Register Data." *Health Services and Outcomes Research Methodology* 15: 82–98. Doi: http://dx.doi.org/10.1007/s10742-015-0137-1.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall. Doi: http://dx.doi.org/10.1007/978-1-4899-3242-6.

Stoer, J. and R. Bulirsch. 2002. *Introduction to Numerical Analysis*, 3rd ed. New York: Springer. Doi: http://dx.doi.org/10.1007/978-0-387-21738-3.

Waal, T. de, J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. New York: John Wiley and Sons.

Zhang, L.-C. 2005. "On the Bias in Gross Labour Flow Estimates Due to Nonresponse and Misclassification." *Journal of Official Statistics* 21: 591–604.

Zhang, L.-C. 2011. "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics* 27: 415–432.

Zhang, L.-C. 2012a. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Zhang, L.-C. 2012b. "On the Accuracy of Register-Based Census Employment Statistics." Paper presented at the European Conference on Quality in Official Statistics, May 30–June 1 2012, Athens. Available at: http://www.q2012.gr/articlefiles/sessions/23.4_Zhang_AaccuracyRegisterStatistics.pdf (accessed April 2016).

Received October 2015
Revised April 2016
Accepted May 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 643–660, http://dx.doi.org/10.1515/JOS-2016-0033

Detecting Fraudulent Interviewers by Improved Clustering Methods – The Case of Falsifications of Answers to Parts of a Questionnaire

Samuel De Haas[1] *and Peter Winker*[1]

Falsified interviews represent a serious threat to empirical research based on survey data. The identification of such cases is important to ensure data quality. Applying cluster analysis to a set of indicators helps to identify suspicious interviewers when a substantial share of all of their interviews are complete falsifications, as shown by previous research. This analysis is extended to the case when only a share of questions within all interviews provided by an interviewer is fabricated. The assessment is based on synthetic datasets with a priori set properties. These are constructed from a unique experimental dataset containing both real and fabricated data for each respondent. Such a bootstrap approach makes it possible to evaluate the robustness of the method when the share of fabricated answers per interview decreases. The results indicate a substantial loss of discriminatory power in the standard cluster analysis if the share of fabricated answers within an interview becomes small. Using a novel cluster method which allows imposing constraints on cluster sizes, performance can be improved, in particular when only few falsifiers are present. This new approach will help to increase the robustness of survey data by detecting potential falsifiers more reliably.

Key words: Survey data falsifications; partial falsifications; cluster analysis; constraint cluster analysis; bootstrap.

1. Introduction

Survey data are a central ingredient of empirical research in economics, other social sciences, and medicine. The quality of any analysis of surveys depends on the quality of the survey data. A huge literature exists on potential pitfalls linked to issues of sampling and the construction of questionnaires which might have a negative impact on data quality. The issue of potential falsifications by the interviewers, however, has received less attention, although anecdotal reports date back to Crespi (1945). In fact, the prevalence of such behavior might be higher than commonly assumed. For a recent survey of the literature, see Bredl et al. (2013). They conclude that the share might be typically below five percent for large scale surveys with intensive supervision, while it might reach levels exceeding 50% in smaller surveys with limited supervision and difficult framework conditions such as inaccessibility of respondents or binding quota requirements.

[1] University of Giessen, Chair of Industrial Organisation, Regulation and Antitrust, and Chair of Statistics and Econometrics, Licher Strasse 64, 35394 Giessen, Germany. Emails: Samuel.De-Haas@wirtschaft.uni-giessen.de and Peter.Winker@wirtschaft.uni-giessen.de

Acknowledgments: Financial support through the DFG in project WI 2024/5-4 is gratefully acknowledged. We are indebted to the associate editor handling our submission and three anonymous referees for their comments which helped to improve the presentation of our results substantially.

Consequently, approaches focusing both on prevention and deterrence are required. Concerning prevention, this might include appropriate interviewer training, payment, and motivation (Gwartney 2013). Approaches for deterrence include close supervision and controls in the field and of the collected data.

Bredl et al. (2012) proposed a method for the analysis of collected data, which employs a clustering procedure on multivariate indicators calculated at interviewer level. The method has been tested successfully on real and experimental datasets (Menold et al. 2013). The method was not meant to replace other methods used for quality management such as reinterviews (Forsman and Schreiner 1991), but rather to focus them on a subset of interviewers exhibiting conspicuous patterns in the data they contribute. The number of accessible real and experimental datasets with identified falsifications is limited, as there are no incentives to report such cases in real surveys (for an exception see Finn and Ranchhod 2013). Although reporting identified fabrications in survey data might be considered a clear signal of successful supervision and quality management, it might also mislead the reader into challenging the integrity of the data. For this reason, identified falsifications in survey data are typically removed before the data are made available for further research without explicit indication. As a consequence, the robustness of the method with regard to the choice of indicators and the structure of the dataset (number of interviewers, share of falsifiers) cannot be assessed solely by using the few datasets available. Insights into this problem can be gained by generating synthetic data from real or experimental data using a bootstrap-based approach as described by Storfinger and Winker (2013).

The bootstrap approach has been used to analyze the performance of the clustering procedure for partial falsifications, that is, the situation when interviewers provide some real interviews and add falsifications, for example, to complete a quota (De Haas and Winker 2014). The present article complements this earlier work with an analysis of partial falsifications within questionnaires, that is, for the situation when interviewers collect part of the data from the respondents and complete the questionnaire themselves afterwards. Anecdotal evidence suggests that there are different reasons for both types of partial falsifications in real surveys. Examples are that part of the questionnaire comprises embarrassing questions that the respondents refuse to answer or questions which are time-consuming when filled in with the respondent. As in previous work (De Haas and Winker 2014), we are interested in the effects of shrinking shares of fabricated answers by a deviant interviewer on the performance of the clustering procedure.

We add a novel clustering tool for the present analysis, which allows us to impose a priori constraints on the (expected) maximum number of falsifiers. This approach is motivated by the finding that the unconstrained clustering method tends to produce a substantial share of false alarms, especially when the share of falsifiers is low or only partial falsifications are provided (De Haas and Winker 2014). Using the constrained approach might improve the discriminatory power of the method, in particular, when the share of falsifiers is small. Finally, we add Matthew's correlation coefficient (Matthews 1975) as an alternative summary measure. It complements the standard measures used for the quality of the assignment of interviewers to the two groups of honets and supposedly deviant interviewers, namely oversights and false alarms.

The article is organized as follows. In Section 2 we introduce the methods used for the identification of falsifications, in particular the indicators constructed at the interviewer

level, the standard clustering procedure, and the new variant imposing a size constraint on the falsifier cluster. The experiment providing the data is described in Section 3 together with the bootstrap procedure for generating synthetic data with a specific structure. The results are summarized in Section 4, while Section 5 concludes and provides an outlook onto further steps of the research on partial falsifications.

2. Methods

The data-based identification of potential falsifications uses properties of the data which differ between real and falsified interviews. The indicators used for that purpose should ideally be independent of a specific questionnaire. At the same time, they should be unknown to the interviewers to avoid strategic falsifications. Several indicators have been proposed and used previously (Schäfer et al. 2005; Kemper et al. 2011; Storfinger and Opper 2011; Bredl et al. 2012; Menold et al. 2013; Kemper and Menold 2014; Menold and Kemper 2014; De Haas and Winker 2014). To allow for a comparison of the results, we use the same indicators as De Haas and Winker (2014). A full list of these indicators with a short description is provided in Appendix A.

In the following, the indicator acquiescent responding style (ARS), which has been used previously by Kemper et al. (2011), illustrates the idea of using indicators to separate real and fabricated interviews. It is constructed based on pairs of items which address similar issues, but differ in using either a positive or negative wording, respectively. Consequently, fully rational respondents should choose opposite answers. However, it is commonly observed that respondents tend to prefer to agree with a given statement and to some extent provide inconsistent answers to such pairs of questions. While some interviewers might be aware of these phenomena, it may be impossible for them to judge the extent of such acquiescent behavior by real respondents. In fact, results from previous studies show that falsifiers tend to exhibit less acquiescence in their fabricated interviews (Menold et al. 2013). The indicator ARS used for the present application is based on five pairs of such items and measures the relative agreement frequency. Here, agreement frequency is defined as the share of the answer options "fully correct" and "fairly correct".

While most indicators can be calculated for each questionnaire, it appears doubtful that they would allow for a discrimination at the level of individual interviews. Typically, the number of questions linked to each indicator is rather small unless the questionnaires are extensively long. For this reason, as in previous research (Menold et al. 2013; De Haas and Winker 2014), we focus our analysis on the interviewer level. Thus, the values of all indicators are calculated based on all interviews for each interviewer. We will consider an interviewer as deviant ("falsifier") if at least one of her or his interviews or parts thereof are not obtained from real respondents, but are fabricated by the interviewer her- or himself. Obviously, given the aggregation at the interviewer level, detection might become more difficult if the share of fabricated (parts of) interviews is low.

Differences between real and false interviews might show up simultaneously in several indicators. If these indicators are not perfectly correlated (for most of the indicators used here, the pairwise correlation is found to be smaller than 0.2), exploiting the multivariate structure in a cluster analysis is expected to outperform splits based on a single indicator. This idea is supported by the findings presented in earlier research (Bredl et al. 2012;

Menold et al. 2013). Furthermore, using the multivariate distribution of several indicators makes it more difficult for a deviant interviewer to generate data meeting the properties of real data closely enough to pass through undetected.

For the cluster analysis, each interviewer k is represented by a vector of indicator values $\mathbf{i}_{k,j}, k = 1, \ldots, K$ and $j = 1, \ldots, J$. K denotes the total number of interviewers and J the number of indicators. Prior to performing the cluster analysis each indicator value is standardized, resulting in

$$\tilde{\mathbf{i}}_{k,j} = \frac{\mathbf{i}_{k,j} - \bar{\mathbf{i}}_{.,j}}{\sqrt{\text{var}(\mathbf{i}_{.,j})}}.$$

Often, the task of clustering a set of vectors like $\tilde{\mathbf{i}}_k = (\tilde{\mathbf{i}}_{k,1}, \ldots, \tilde{\mathbf{i}}_{k,J})$ is tackled by using hierarchical (agglomerative or divisive) clustering methods (Baragona et al. 2011, 199ff). However, the sequential approach of agglomerative procedures might not result in a global optimum for the assignment. Consequently, we apply a global clustering approach. While existing methods such as k-means (Baragona et al. 2011, 211) could be employed in this case, they also do not guarantee convergence to a global optimum given their iterative local search. Here, we use Threshold Accepting (TA) because it is a globally convergent optimization heuristic. It has been shown that under certain conditions it will converge to the global optimum when the number of search steps goes to infinity (Althöfer and Koschnik 1991; Winker 2001). Other optimization heuristics might also be used in this context, for example the clustering method based on genetic algorithms described by Baragona et al. (2011, 219ff). A further advantage of using this approach is the possibility to add constraints. In our application, this option will be used to limit the size of the cluster corresponding to potential falsifications.

We start with the description of the unconstrained version of the algorithm as used previously by Storfinger and Winker (2013) and De Haas and Winker (2014): it aims at minimizing an objective function which is calculated as the sum of the pairwise Euclidean distances within the clusters. Hence, the goal consists in reducing the heterogeneity with regard to the values of the various indicators within each group. The algorithm is initialized with a randomly drawn assignment of all elements (interviewers) into two groups. Afterwards, for a preset number of iterations, one randomly chosen element is regrouped in each iteration. The resulting new clusters are accepted as long as the value of the objective function is improved (decreases) or at least does not increase by more than a predefined threshold. In order to find a global optimum, or at least a close approximation, this local search step has to be repeated many times. An obvious drawback of the optimization-based clustering method as compared to traditional clustering algorithms is its higher computational cost. On a standard desktop computer (i7-3770 CPU, 3.40 GHz, 8 GB RAM) a single run of the Matlab implementation with 2,500,000 iterations takes about 110 seconds to finish. This becomes relevant in a simulation study as the present one, when the clustering problem has to be solved thousands of times, while it is not a major issue for a single application to a single real dataset in a survey setting.

After the clustering algorithm is completed, the identification of the two clusters is based on the assumptions about the indicator values for honest versus cheating interviewers (see Table 2 in Appendix A). Therefore, it can be decided automatically

which of the two subgroups represents the falsifiers and the honest interviewers, respectively. In order to perform this identification step unsupervised for each bootstrap simulation, we sum up the standardized mean values for every indicator over all interviewers in each cluster. To this end the signs of the indicators are adjusted in such a way that higher values always point to the group containing the potential falsifiers.

Employing this clustering procedure may result in two potential types of errors. Honest interviewers might be incorrectly added to the group labeled "deviant interviewers". Such a misassignment is called "false alarm" or "false negative". On the other hand, some falsifiers might be allocated to the group labeled "honest interviewers". This type of error is called "oversight" or "false positive". To provide a summary measure of the extent of such misclassifications, Matthews's correlation coefficient (MCC) is used (Matthews 1975). The MCC is calculated as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}, \tag{1}$$

where TP denotes the number of true positives, that is, the number of interviewers correctly assigned to the group labeled as "honest interviewers". TN is the number of true negatives, that is, of correctly identified falsifiers. FP is the number of false positives (oversights) and FN the number of false negatives (false alarms) (for a comprehensive overview of alternative measures used to evaluate the quality of binary classifications see also Verbiest et al. 2015). By construction, MCC takes on values between -1 and 1. If all interviewers are correctly assigned, it takes on the value one, for a random, that is, noninformative assignment, the values should be close to zero, while when exactly the wrong assignment is given, MCC takes on the value -1.

The objective function used lays highest emphasis on the homogeneity within its clusters based on pairwise Euclidian distances. Implicitly, this objective functions favors clusterings with rather similar group sizes. Therefore, it might be expected to perform better if the share of falsifiers is about 50%, while it tends to generate a large number of false alarms for a low share of falsifiers in the dataset. Consequently, imposing some additional constraints on the size of the falsifier cluster might be beneficial, in particular if a modest share of falsifications is expected and the cost of controls is high. It might help to concentrate controls on those interviewers with the highest risk. Obviously, if some falsifications are found in this group, there is a risk that even a larger share of interviewers is actually deviating and the analysis might need to be repeated allowing for a larger size of the falsifier cluster.

Technically, the constraint is imposed in the following way: the same optimization heuristic as described above is used. The objective function is augmented by an additional term, described as "penalty term". This part of the objective function takes on the value zero as long as the number of interviewers assigned to the potential falsifiers cluster is smaller than or equal to the predefined limit. If more interviewers are assigned to this cluster, the term becomes positive and is an increasing function of the differences between the number of potential falsifiers and the predefined value. Due to this penalty term, the value of the objective function decreases when the falsifier cluster becomes smaller as long as its size is above the predefined limit. Consequently, as the algorithm proceeds and the weight of the penalty increases, the current solution becomes more and more likely to

satisfy the imposed constraint, which should be met by the final solution. For a description of alternative ways to handle constraints in the framework of an optimization heuristic, see Gilli et al. (2011, 352ff). The computational complexity of the algorithm is higher when the constraint is taken into account, as the evaluation of the penalty term requires the identification of the falsifier cluster in each iteration.

Pseudocode 1 shows the pseudocode of the algorithm taking a constraint on the size of the falsifier cluster into account.

The algorithm is run for a defined setting of number of interviewers, share of falsifiers and extent of falsifications (1:). Then, a substantial number B of bootstrap replications is performed (2: to 8:). For each replication a new synthetic dataset with the defined properties is generated (3:) (for the details of the bootstrap procedure, see Section 3 below). Based on the dataset, the vector of indicator values $\tilde{\mathbf{i}}_k$ is calculated for every interviewer (4:). Next, two clusters are identified by means of the constrained version of the optimization heuristic described above, ensuring that the cluster labeled falsifier cluster does not contain more than the predefined number of interviewers (5:). The performance of this clustering is evaluated based on the MCC. After all bootstrap replications have been carried out, the overall performance can be summarized, For example, by the mean of the MCC for a given set-up (8:).

In order to evaluate this method, different values of the constraint on the size s of the falsifier cluster are combined with different simulation setups, that is, different shares of falsifiers and falsified questions. In reality, however, one would need a rough idea of the falsifiers' share in the underlying dataset to introduce a plausible constraint. Optimally, a preselection criterion like the total value of pairwise distances is used to compare different constraints in a pretest. The corresponding results could be used to define the most preferable constraint. The development of such a pretest is left for future research.

Before turning to the data and results, a special case of the above procedure is introduced. It consists in limiting the size of the falsifier cluster to one. In this case, the global optimization algorithm can be replaced by an exact enumeration procedure to find the one interviewer resulting in the optimum for the objective function if labeled as a falsifier. Obviously, this approach is well suited neither for obtaining an estimate of the extent of falsifications in the sample nor for the identification of a larger share of such

Pseudocode 1 Pseudocode of bootstrap procedure for cluster analysis.

1: Define parameter settings: number of bootstrap replications B; size restriction for group of potential falsifiers s, parameters $n_1, \ldots, n_5$
2: **for** $b = 1$ to B **do**
3: Create artificial dataset: n_1 honest interviewers with n_2 real questionnaires; n_3 falsifiers with n_4 fabricated and n_5 real answers in all of their questionnaires
4: Calculate indicators for all interviewers
5: Conduct clustering analysis imposing the group size restriction s for the group of falsifiers
6: Store performance of clustering procedure and indicators for given dataset
7: **end for**
8: Summarize statistical information of performance over all B datasets

falsifiers. However, given the low computational cost, it might be a sensible first step in quality control to identify this extreme interviewer and conduct a follow-up on his or her data. We will also report results on the quality of this simplified procedure for the detection of only one potential falsifier in Section 4. The evaluation of this procedure will not be based on MCC, but simply on the frequency over all bootstrap replications, with which a falsifier is actually found in this one element cluster.

3. Data

To assess the methods' performance, a substantial number of datasets from surveys would be required. To this end both interviews collected by honest interviewers and interviews collected by faking interviewers should be contained and identified a priori. Although anecdotal evidence suggests a substantial prevalence of deviant behavior in surveys, such datasets are rarely available (Bredl et al. 2013). Typically, identified falsifications are removed from the dataset prior to further analysis. Publications based on the cleaned data do not contain information about the falsifications as it might provide a negative signal on the data quality. Falsifications which have remained undetected are still present in datasets, but cannot be used for the evaluation of our methods either, as no known benchmark is available.

For the present study, we resort to the results of a large-scale experiment conducted at the University of Giessen in 2012 providing both real and fabricated data (Menold et al. 2013). 78 students of the University of Giessen were recruited as interviewers. In a first stage of the experiment, they each conducted about ten real interviews using a questionnaire comprising sociodemographic information and questions about study subjects and on attitudes. The respondents were recruited randomly by the interviewers among other students. The quality of these real interviews was verified by controlling the tape recordings of all interviews. In a second stage of the experiment, each student was asked to fabricate another ten interviews in the laboratory. As input for these falsifications, the students were provided with a short sociodemographic profile of one of the respondents from the real data who was not interviewed by themselves. Making use of this profile, the students were asked to generate data which should replicate a real interview as close as possible. An additional monetary incentive for generating high quality falsifications was provided which was distributed to those interviewers generating data that could not be assigned to the falsifier cluster by the method proposed in Bredl et al. (2012). Given this explicit incentive, the interviewers' knowledge from conducting real interviews first and their knowledge about the group from which the respondents have been recruited, the experimental setup promotes the generation of high-quality falsifications. Hence, these falsifications might be more difficult to detect compared to "quick and dirty" approaches, which might be more common in some real settings if interviewers are aware of missing or weak supervision. In fact, the quality of assessment of interviewers by the methods discussed in the previous section has been substantially higher in the few applications to real data (Bredl et al. 2012; Storfinger and Winker 2013). Therefore, we consider the data used in this article as a worst-case (difficult to detect) scenario, but will discuss potential limitations of the dataset in the concluding section.

In our set-up, for each respondent we obtain a real interview conducted by one interviewer and a fabricated one provided by a different interviewer. Thus, starting with

the fabricated interviews conducted by one interviewer, in each interview some questions can be replaced by the real answers provided by the respondent in a real interview. This way, it becomes possible to generate synthetic datasets which contain interviews composed of actual answers to some questions and the falsifications provided by the interviewer to the other ones (in contrast, in De Haas and Winker (2014) complete real and fabricated interviews have been used for one interviewer). The share of these falsifications can be controlled when generating the synthetic data. Obviously, for the group of nondeviant interviewers, only data from the interviews actually conducted are employed. As a consequence, it is possible to control both the share of falsifiers and the extent to which their fabricated interviews comprise real and false data.

We are interested in how well the methods described in the previous section perform depending on the share of falsifiers and the extent of falsifications within the fabricated interviews. Obviously, this performance will depend on the specific selection of data from our experiment and, as the consequence of a random choice, has to be considered stochastic. Thus it is not sufficient to consider single synthetic datasets, and instead the analysis has to be repeated a large number of times for different random selections to allow for systematic conclusions.

Given that hardly any appropriate data are available from real surveys and generating suitable data through thousands of separate experiments of the type described above is not feasible either, the well-known resampling method known as bootstrap (Efron 1979, 1982) is used to generate synthetic datasets with the defined properties. The procedure comprises the following steps for each bootstrap iteration, that is, for the generation of a single synthetic dataset to be used in the analysis: first, a predefined number of interviewers is chosen randomly (with replacement) from all interviewers. This group will represent the honest interviewers. This means, for each of these interviewers a fixed number of real questionnaires is selected (with replacement) from the actual interviews conducted during the experiment by the corresponding interviewer. Finally, based on the resampled questionnaires, the indicator values are obtained. Second, another group of interviewers is generated in the same way as for the honest interviewers by selecting randomly (with replacement) a predefined number of interviewers from all interviewers. This second group is meant to represent falsifiers. Therefore, for each interviewer in this group, the actual data are compiled in a modified way. A fixed share of fabricated questions is assumed for all questionnaires selected for this interviewer. These partial falsifications are obtained by starting with a randomly selected (with replacement) fabricated interview provided by the respective interviewer during the experiment. Then, a number of questions corresponding to one minus the share of fabricated questions is randomly selected within the questionnaire. The answer to these questions is replaced by the corresponding real data collected by another interviewer, which provided the profile for the falsification. Finally, the indicator values for the falsifiers are calculated. This procedure allows to generate a large number of synthetic datasets with well-defined properties regarding the number of interviewers, the share of falsifiers, the number of questionnaires per interviewer and the extent of partial falsifications for the falsifiers. This makes it possible to analyze the performance of the clustering method for different scenarios based on – in the present study – 1,000 different samples for each scenario.

4. Results

Given this article's main focus on semifalsifications and the potential gains in discriminatory power of imposing size constraints on the falsifier cluster, a smaller set of experiments compared to the design in De Haas and Winker (2014) is conducted. Thus we take into account the substantially higher computational burden due to the repeated application of the TA heuristic for different constraints on the size of the falsifier cluster. In order to preserve comparability, the set of experiments is chosen as a subset of the original design. The details of the design are summarized in Table 1.

For all experiments, a number of 150 interviewers – similar to the original setting of the experiment – is used. In previous analysis, it was found that reducing the total number of interviewers typically helps to improve the discriminatory power. Thus the results presented here might be considered as lower bounds for the performance to be expected when the number of interviewers is small. Furthermore, the actual share of falsifiers in the dataset might affect the discriminatory power. We consider two settings for the share of falsifiers, namely six percent as a low and 50% as a substantial value. Obviously, the case of 50% falsifications might be considered an extreme setting. Finally, to study the impact of only partial falsifications, both a setting with 50% fabricated data in each questionnaire and one with 100% fabrication as in the experimental data and also as analyzed by Storfinger and Winker (2013) is used for comparison. In Appendix B, we consider shares of falsifiers of two percent, ten percent and 20%, respectively, and provide results for further shares of falsified questions, in particular for 25%, 70%, 75%, 80%, 85%, 90%, and 95%. Given that the main qualitative findings support those obtained for the main design, we do not comment on these additional cases in the text.

For the new parameter "size of the falsifier cluster", several values are also considered. Given that this size is unknown in real applications, a value of two percent stands for a low expectation about the prevalence of faking that is below the actual shares considered in the bootstrap simulations (six percent and 50%). The algorithm should exhibit a low false alarm rate in this case. The value of six percent corresponds to the actual number of falsifiers in one setting and still is much lower than the actual number for the other. 25% is an assumption that is high for the low-share setting, which has to result in a high rate of false alarms, while it is still low for the high faker-share setting. The highest value of 50% corresponds to the high faker setting. In addition to these four values, the algorithm is also run without restriction of the size of the falsifier cluster for comparability with the results in Storfinger and Winker (2013) and De Haas and Winker (2014).

A final setting, which might turn out to be interesting for practical applications, when the focus is just on finding some fakers, but not necessarily many or all of them, is given by

Table 1. Simulation settings for bootstrap runs.

Dimension	Original sample	Values for bootstrap			
No. of interviewers	156				150
No. of questionnaires per interviewer	~10				20
Share of falsifiers	50%			6%	50%
Faking share	100%			50%	100%
Constraint on size of falsifier cluster	n.a.	2%	6%	25%	50%

a constraint of the falsifier cluster to size one. This setting does not require an explicit optimization, as the global optimum can be easily identified by the full enumeration of all possible cases (i.e., by checking one interviewer at a time to represent the falsifier cluster, and selecting the one resulting in the best value of the objective function). It could be extended recursively by applying the method again after removing the interviewer identified as a potential falsifier in the previous step. Alternatively, one might look at the aggregated indicator values per interviewer and label the interviewer(s) exhibiting the largest values as potential falsifiers. For this approach, it is required to know a priori the direction of deviations of indicator values in falsifications, which might not always be obvious. Both approaches are not listed in Table 1, but we will also report results for these screening tools.

The results for the MCC values for all settings except the last mentioned screening devices are summarized in Figure 1. The bars labeled "w/o" provide the results for the unconstrained clustering, that is, using the same methodology as De Haas and Winker (2014). For the case of complete falsifications (graphs in the right column labeled "100% falsified questions") the results are comparable to the same setting in De Haas and Winker (2014). They confirm that the method has a strong potential to identify falsifiers, in particular if the share of falsifiers is high (lower right graph). If only few fakers are present (upper right graph), the tendency of the unconstrained clustering procedure towards clusters of similar sizes, results in a large share of false alarms and, consequently, in an MCC value close to zero.

The performance of the unconstrained clustering method deteriorates when only partial falsifications are considered (left column of Figure 1). While the performance remains weak for the case with few fakers (upper left graph), it becomes substantially worse for the high faker setting (lower left graph). In fact, the MCC value shrinks from about 0.5 to only 0.1. Again, this finding is qualitatively similar to those obtained by De Haas and Winker (2014) for the case of partial falsifications in the sense that a faker produces some completely real and some completely fabricated interviews.

Against this background, it is of interest to see to what extent the restricted cluster algorithm can improve the discrimination between honest and faking interviewers. When

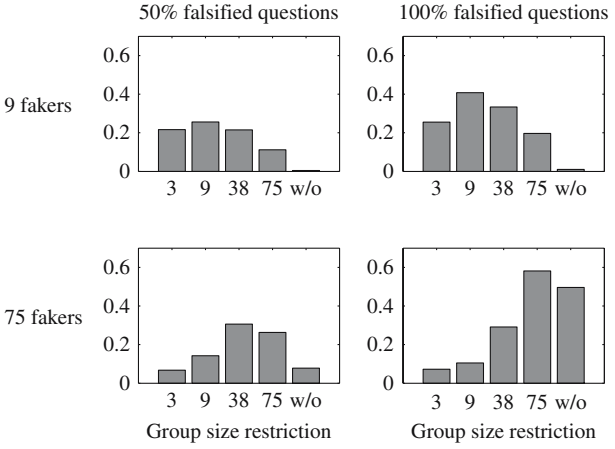


Fig. 1. Mean of MCC values over all bootstrap replications for different settings.

the correct restriction is imposed, that is, a number equal to the actual number of falsifiers (nine for the upper and 75 for the lower row), the MCC values always take on their maximum value, exhibiting a substantial improvement compared to the unconstrained version. These improvements are most pronounced for a low number of falsifications and for the case with partial falsifications. As long as the actual number of falsifiers is low, even wrong assumptions about this number still result in substantially improved MCC values both for partial and complete falsifications (upper row). In the second setting with a large number of falsifiers, however, imposing overly restrictive values for the size of the falsifier cluster (i.e., three or nine instead of the actual number of 75) results in a performance worse than the one of the unconstrained algorithm at least for completely fabricated questionnaires. In the case of partial falsifications, even the restriction to only nine falsifiers results in a slight improvement of the MCC value as compared to the unconstrained setting.

More insights into these results can be obtained by looking separately at the frequencies of oversights and false alarms, which are reported for the same settings in Figure 2. As expected, imposing a small size for the cluster containing the potential falsifiers results in a substantial share of oversights, in particular if the actual number of fakers is high. However, at the same time, the frequency of false alarms is remarkably low, suggesting that the method might be helpful to identify at least a subset of all falsifiers with some precision as long as the imposed size constraints are close to or smaller than the actual number of falsifiers.

We finish with a look at the "screening tools". Given that in all settings described in Table 1 at least one (partial) falsifier is present in the sample, when imposing a falsifier cluster of size one, one would hope that the method always spots one of the actual falsifiers. In fact, the simple procedure comes close to this result for the first interviewer marked as potential falsifier when considering only the more challenging setting of partial falsifications. If only nine falsifiers are present, in 81.6% of the bootstrap samples an

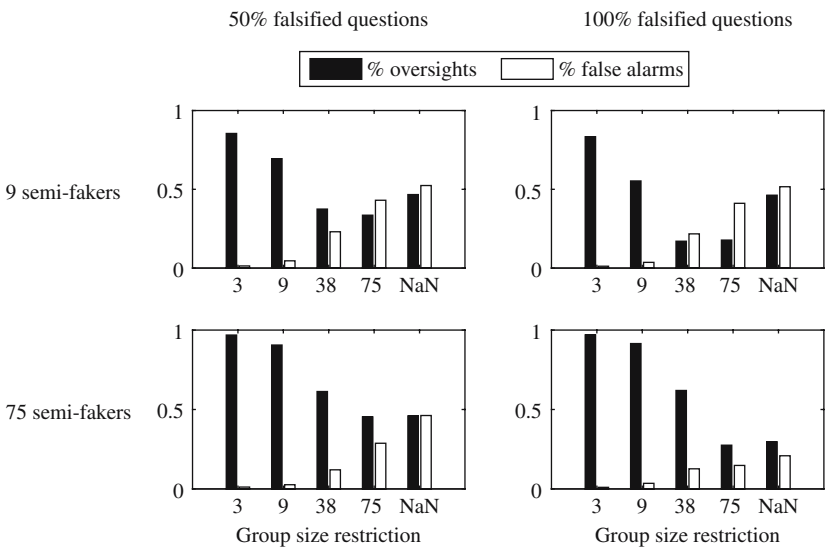


Fig. 2. Frequencies of oversights and false alarms over all bootstrap replications for different settings.

actual falsifier is found, while this share is 98.1% in the case of 75 falsifiers. The alternative approach of selecting the interviewers with highest aggregate indicator values results in shares of 65.5% and 99.9%, respectively. Obviously, if the sample does not contain a single falsifier, the one signal generated by the screening tools will always be a false alarm. Thus, it should not be used to "identify" falsifiers, but rather to select one case for a careful follow up if available resources do not allow for a more comprehensive quality check.

5. Conclusions and Outlook

The present article has analyzed how the identification of falsifiers by means of a data-driven clustering procedure is negatively affected if only partial falsifications are presented in the sense that falsifiers use partially real information and complement this with fabricated data. As expected, the performance deteriorates substantially when the share of fabricated answers decreases. This happens to a similar extent as for the setting when falsifiers generate some of their questionnaires completely (De Haas and Winker 2014).

The situation that interviewers rather fabricate some of their assignments or parts of the questionnaire than delivering only complete falsifications is considered as quite typical in real survey settings. Although empirical evidence is limited, completing partial interviews after a break-off by the respondent or just collecting basic sociodemographic information from the respondent and fabricating lengthy or sensitive parts of the questionnaire might represent situations resulting in partial fabrications. Therefore, to deal with the shortcomings of the previously proposed clustering method in this situation, a new clustering procedure is proposed, which allows imposing an a priori restriction on the number of falsifiers in the corresponding cluster. It is shown that imposing such restrictions improves the performance substantially. This holds in particular if the share of falsifiers is low, only partial falsifications are present, and the assumed share of falsifiers is close to the actual number.

As an extreme setting of this restricted clustering approach, a falsifier cluster of size one is also considered. While it is obvious that this method cannot produce a good overall assignment in a case where several falsifiers are present, it appears to be a valuable screening tool, as in all settings for most individual bootstrap replications a falsifier was correctly identified.

Given the high cost of classical methods of quality management such as reinterviews (Forsman and Schreiner 1991), it is recommended to apply the method presented here to select a small number of interviewers exhibiting conspicuous patterns. Thus the (restricted) size of the cluster containing the interviewers flagged for follow up might be set according to available resources for reinterviews or based on previous experience with prevalence of fabrications in a specific survey setting.

The present study has two major limitations, which might be overcome in future research. First, it might be argued that the experimental setting used to generate the data for the present analysis represents a worst-case setting in the sense that high-quality falsifications are obtained given the strong incentives for good fabrications in the experiment. At the same time, one might argue that the quality of falsifications was poor, given that all interviewers were students with no or limited previous experience as

interviewers. The prevailing effect might be evaluated making use of further experimental datasets. Second, falsifications in real data might differ from those obtained in an experimental setting. Therefore, further analysis based on real data such as in Bredl et al. (2012) is required. Besides dealing with these limitations, future research will also address some methodological issues. Alternative objective functions for the cluster construction will be considered, which might improve the performance in particular for the case of a low share of falsifiers, as the present method privileges clusters of about equal size. Furthermore, the usage of cross validation techniques to find a good a priori value for the size of the falsifier cluster is left for future analysis. Finally, probabilistic clustering methods are natural competitors in the case of partial falsifications and will also be a subject of our future work.

Appendix A

Indicators

Table 2 provides a summary of the indicators used to differentiate between data generated by interviewers following the prescribed procedure and data coming from faking interviewers. It provides the name of the indicator, a brief explanation of how it is constructed, the expectation about the sign of the deviation in its value between honest interviewers and falsifiers, a short argument explaining this expectation and a reference to the specific indicator. A more detailed description of all indicators used in the present study can be found in Menold et al. (2013) and De Haas and Winker (2014).

Table 2. Summary of indicators used to identify data potentially stemming from falsifications.

Indicator	Brief explanation	Higher value for	Rationale	Reference
Extreme Responding Style	Frequency of choosing extreme responses on rating scale	Honest	Falsifiers try to avoid extreme responses	Porras and English (2004)
Middle Responding Style	Frequency of choosing middle category (for uneven number of categories)	Falsifiers	Falsifiers try to avoid extreme responses	Storfinger and Opper (2011)
Acquiescent Responding Style	Agreement responses regardless of positive or negative item wording	Honest	Honest often agree regardless of content	Messick (1967)
Nondifferentiation	Standard deviation across all items	Falsifiers	Falsifiers use stereotypes	Reuband (1990)
News	Frequency of choosing fictitious categories (read magazines)	Falsifiers	Falsifiers try to save time and effort	Menold et al. (2013)
Filter	Frequency of choosing answers which allow skipping of further questions	Falsifiers	Falsifiers try to save time and effort	Hood and Bushery (1997)
Participation	Asking for past political activities	Honest	Falsifiers underestimate those activities	Menold et al. (2013)
Semi-Open	Frequency of choosing "other, please specify"	Honest	Falsifiers try to save time and effort	Hood and Bushery (1997)
Open	Frequency of providing answers to open questions	Honest	Falsifiers try to save time and effort	Bredl et al. (2012)
Rounding	Frequency of rounded numbers to numerical open questions	Falsifiers	Numerical information should be exactly known for real respondent	Tourangeau et al. (1997)
Primacy	Frequency of choosing the first two categories (visual presentation)	Honest	Honest tend to choose first option that seems satisfactory	Krosnick and Alwin (1987)
Recency	Frequency of choosing the last category (oral presentation)	Honest	Limited capacity of short-term memory	Krosnick and Alwin (1987)

Appendix B

Results for Other Shares of Falsified Questions

As additional information complementing Figure 1, we provide results for shares of falsified questions of 25%, 70%, 75%, 80%, 85%, 90%, and 95% in this appendix. Figure 3 shows the results for a situation when three falsifiers are present (two percent of all interviewers), while the results for 15 falsifiers (10%) and 30 falsifiers (20%) are exhibited in Figures 4 and 5, respectively.

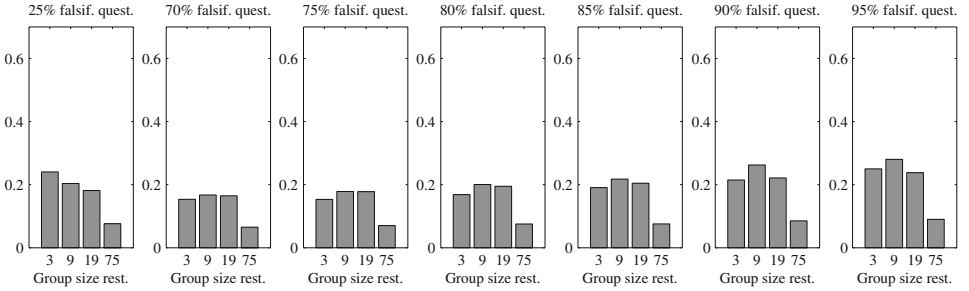


Fig. 3. Mean of MCC values over all bootstrap replications with three falsifiers.

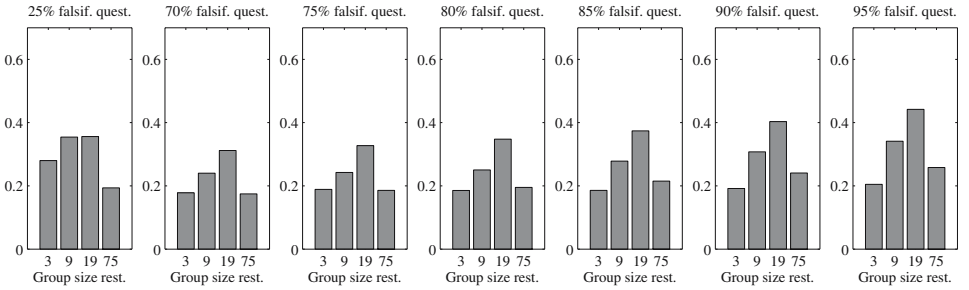


Fig. 4. Mean of MCC values over all bootstrap replications with 15 falsifiers.

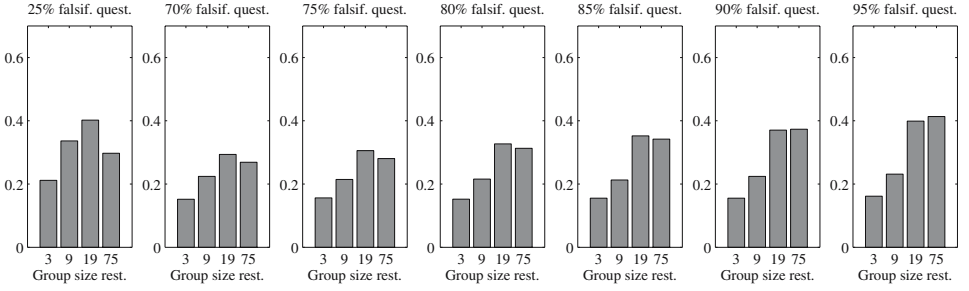


Fig. 5. Mean of MCC values over all bootstrap replications with 30 falsifiers.

6. References

Althöfer, I. and K.-U. Koschnik. 1991. "On the Convergence of Threshold Accepting." *Applied Mathematics and Optimization* 24: 183–195. Doi: http://dx.doi.org/10.1007/BF01447741.

Baragona, R., F. Battaglia, and I. Poli. 2011. *Evolutionary Statistical Procedures*. Statistics and Computing. Heidelberg: Springer.

Bredl, S., N. Storfinger, and N. Menold. 2013. "A Literature Review of Methods to Detect Fabricated Survey Data." In *Interviewers' Deviations in Surveys - Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 3–24. Frankfurt am Main: Peter Lang.

Bredl, S., P. Winker, and K. Kötschau. 2012. "A Statistical Approach to Detect Interviewer Falsification of Survey Data." *Survey Methodology* 38: 1–10.

Crespi, L. 1945. "The Cheater Problem in Polling." *The Public Opinion Quarterly* 9: 431–445.

De Haas, S. and P. Winker. 2014. "Identification of Partial Falsifications in Survey Data." *Statistical Journal of the IAOS* 30: 271–281. Doi: http://dx.doi.org/10.3233/SJI-140834.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7: 1–26. Doi: http://dx.doi.org/10.1214/aos/1176344552.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 38. Doi: http://dx.doi.org/10.1137/1.9781611970319.

Finn, A. and V. Ranchhod. 2013. "Genuine Fakes: The Prevalence and Implications of Fieldworker Fraud in a Large South African Survey." SALDRU Working Papers 115, Southern Africa Labour and Development Research Unit, University of Cape Town. Available at: http://ideas.repec.org/p/ldr/wpaper/115.html (accessed October 22, 2015).

Forsman, G. and I. Schreiner. 1991. "The Design and Analysis of Reinterview: An Overview." In *Measurement Errors in Surveys*, edited by P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman, 279–301. Chichester: Wiley. Doi: http://dx.doi.org/10.1002/9781118150382.ch15.

Gilli, M., D. Maringer, and E. Schumann. 2011. *Numerical Methods and Optimization in Finance*. Waltham, MA: Academic Press.

Gwartney, P. 2013. "Mischief Versus Mistakes: Motivating Interviewers to not Deviate." In *Interviewers' Deviations in Surveys - Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 195–215. Frankfurt am Main: Peter Lang.

Hood, C. and M. Bushery. 1997. "Getting More Bang from the Reinterviewer Buck: Identifying 'at Risk' Interviewers." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, August 10th to 14th 1997, Anaheim, CA, 820–824. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1997_141.pdf (accessed October 22, 2015).

Kemper, C. and N. Menold. 2014. "Nuisance or Remedy? The Utility of Stylistic Responding as an Indicator of Data Fabrication in Surveys." *Methodology: European*

Journal of Research Methods for the Behavioral and Social Sciences 10: 92–99. Doi: http://dx.doi.org/10.1027/1614-2241/a000078.

Kemper, C., V. Trofimow, B. Rammstedt, and N. Menold. 2011. "Indicators for the ex post Detection of Faking in Survey Data Constructed from Responses to the Big Five Inventory-10 (BFI-10)." Poster presented at the 11th European Conference on Psychological Assessment, date of conference, Riga, Latvia. Available at: http://www.ecpa11.lu.lv/files/Kemper Christoph.pdf (accessed October 22, 2015).

Krosnick, J. and D. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement." *Public Opinion Quarterly* 51: 201–219. Doi: http://dx.doi.org/10.1086/269029.

Matthews, B. 1975. "Comparison of the Predicted and Observed Secondary Structure of t4 Phage Lysozyme." *Biochimica et Biophysica Acta* 405: 442–451. Doi: http://dx.doi.org/10.1016/0005-2795(7590109-9).

Menold, N. and C. Kemper. 2014. "How Do Real and Falsified Data Differ? Psychology of Survey Response as a Source of Falsification Indicators in Face-to-Face Surveys." *International Journal of Public Opinion Research* 26: 41–65. Doi: http://dx.doi.org/10.1093/ijpor/edt017.

Menold, N., P. Winker, N. Storfinger, and C. Kemper. 2013. "A Method for ex-post Identification of Falsifications in Survey Data." In *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 25–47. Frankfurt am Main: Peter Lang.

Messick, S. 1967. "The Psychology of Acquiescence, an Interpretation of Research Evidence." In *Response Set in Personality Assessment*, edited by I. Berg. Chicago: Aldine Publishing Company. Doi: http://dx.doi.org/10.1002/j.2333-8504.1966.tb00357.x.

Porras, J. and N. English. 2004. "Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, August 8th to 12th 2004, Toronto, 4223–4228. Available at: http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000879.pdf (accessed October 23, 2015).

Reuband, K.-H. 1990. "Interviews, die keine sind, 'Erfolge' und 'Mißerfolge' beim Fälschen von Interviews." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 42: 706–733.

Schäfer, C., J. Schräpler, K. Müller, and G. Wagner. 2005. "Automatic Identification of Faked and Fraudulent Interviews in the German SOEP." *Schmollers Jahrbuch* 125: 183–193.

Storfinger, N. and M. Opper. 2011. "Datenbasierte Indikatoren für potentiell abweichendes Interviewerverhalten." Discussion Paper 58, ZEU, September 2011, Giessen. Available at: http://geb.uni-giessen.de/geb/volltexte/2012/8559/pdf/ZeuDiscPap58.pdf (accessed October 23, 2015).

Storfinger, N. and P. Winker. 2013. "Assessing the Performance of Clustering Methods in Falsification Using Bootstrap." In *Interviewers' Deviations in Surveys - Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 49–65. Frankfurt am Main: Peter Lang.

Tourangeau, R., K. Rasinski, J. Jobe, B. Jared, T. Smith, and W. Pratt. 1997. "Sources of Error in a Survey on Sexual Behavior." *Journal of Official Statistics* 13: 341–365.

Verbiest, N., K. Vermeulen, and A. Teresdai. 2015. "Evaluation of Classification Methods." In *Data Classification – Algorithms and Applications*, edited by C. Aggarwal, 633–655. Boca Raton, FL: CRC Press.

Winker, P. 2001. *Optimization Heuristics in Econometrics: Applications of Threshold Accepting*. Chichester: Wiley.

Received April 2015
Revised October 2015
Accepted November 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 661–692, http://dx.doi.org/10.1515/JOS-2016-0034

Empirical Best Prediction Under Unit-Level Logit Mixed Models

Tomáš Hobza[1] *and Domingo Morales*[2]

The article applies unit-level logit mixed models to estimating small-area weighted sums of probabilities. The model parameters are estimated by the method of simulated moments (MSM). The empirical best predictor (EBP) of weighted sums of probabilities is calculated and compared with plug-in estimators. An approximation to the mean-squared error (MSE) of the EBP is derived and a bias-corrected MSE estimator is given and compared with parametric bootstrap alternatives. Some simulation experiments are carried out to study the empirical behavior of the model parameter MSM estimators, the EBP and plug-in estimators and the MSE estimators. An application to the estimation of poverty proportions in the counties of the region of Valencia, Spain, is given.

Key words: Poverty; method of moments; logit mixed models; empirical best predictor; mean-squared error; bootstrap.

1. Introduction

This article deals with the estimation of weighted sums of probabilities in domains where the sample size is not large enough to obtain reliable direct estimates. Small-area estimation (SAE) deals with this problem by introducing model-based or model-assisted estimators. See the monographs of Rao (2003) and Rao and Molina (2015), and the reviews of Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002, 2013), and Jiang and Lahiri (2006) for an introduction to SAE.

The binomial-logit mixed models are generalized linear mixed models (GLMM) that take into account the between-domains variability that it is not explained through auxiliary variables by introducing random effects. The random effects are usually assumed to be normally distributed. Inferences based on GLMMs have some computational difficulties because the likelihood may involve high-dimensional integrals which cannot be evaluated analytically. This article uses the method of simulated moments (MSM), introduced by Jiang (1998), to fit the proposed model. This method approximates the method of moments (MM), is computationally attractive, and gives consistent estimators of model parameters.

[1] Department of Mathematics, Czech Technical University in Prague, Trojanova 13, 12000 Prague 2, Czech Republic. Email: hobza@fjfi.cvut.cz
[2] Operations Research Center, Miguel Hernández University of Elche, Avda. de la Universidad s/n, 03202 Elche, Spain. Email: d.morales@umh.es

Acknowledgments: The author thanks the Office of Social, Demographic and Economic Statistics of the Valencian Government for providing the real data employed in the application of this article. This study was partially supported by the Spanish grant MTM2015-64842-P and by the Czech grant SGS15/214/OHK4/3T/14.

The article derives empirical best predictors (EBP) for estimating weighted sums of probabilities, where the weights are known quantities taken from administrative registers or census files. This is to say, the term "weighting" is distinct from sample survey weighting based on the survey design. The EBPs are sums of predicted values of probabilities (sometimes called soft estimates) rather than sums of target-variable predictions that must be 0 or 1 (hard estimates). The statistical methodology is taken and adapted from Jiang and Lahiri (2001) and Jiang (2003), where EBPs of functions of fixed effects and small-area-specific random effects were developed in the context of logistic mixed models and GLMM respectively. Furthermore, plug-in estimators are considered and empirically studied in simulation experiments.

The MSE is a standard accuracy measure for point estimators. Jiang and Lahiri (2001) and Jiang (2003) studied the approximation of the MSE of the EBP in the context of binary data and GLMM. Their approach is based on Taylor series expansions. They further gave a second-order bias-corrected estimator of the MSE. We adapt the MSE calculations given by these authors to the case of EBPs of weighted sums of probabilities. We give two analytical estimators of the MSE approximation, without and with a bias-correction term. As these MSE estimators are computationally expensive in practice, we consider the parametric bootstrap estimator introduced by Gonzalez-Manteiga et al. (2007) in the context of logistic mixed models. This approach was later extended by Gonzalez-Manteiga et al. (2008a,b) to nested error regression models and to multivariate area-level models respectively. As the simple parametric bootstrap method tends to underestimate the MSE when domain sizes are too small, we also give a double-bootstrap bias-corrected estimator by following Hall and Maiti (2006a,b).

Jiang and Lahiri (2001) and Jiang (2003) introduced the basic statistical methodology for EBPs of functions of fixed and random effects. They also studied the large sample properties of the EBPs and MSE estimators. However, they did not carry out simulation experiments to empirically investigate the behavior of the EBPs and MSE estimators in the standard small-area estimation setup, that is when the domain sample sizes are small. They also did not present applications to real data. This article in part covers this gap.

We carry out simulation experiments to investigate the behavior of the EBPs of weighted sums of probabilities and the corresponding MSE estimators. We investigate computational and numerical issues appearing in the implementation of the EBP methodology. The article also presents an application to 2012 data from the Spanish living conditions survey (SLCS2012) in the region of Valencia. The target of the application is the estimation of poverty proportions at county level.

Other SAE models for the estimation of poverty proportions are currently available. Without being exhaustive, we cite some basic references here. Chambers et al. (2012) introduce an M-quantile regression approach for binary data. Farrell et al. (1997) give bootstrap adjustments for empirical Bayes interval estimates of small-area proportions. Malec et al. (1997) gave some small-area inference methods for binary variables. By using unit-level linear mixed models, Molina et al. (2014) and Molina and Rao (2010) give several procedures for estimating poverty indicators. There is also Elbers et al. (2003) and its extensions, who use full unit record census data, rather than building a model-based census substitute from cross tabulations. Based on temporal and spatio-temporal area-level

models, Esteban et al. (2012a,b), Marhuenda et al. (2013), and Morales et al. (2015) give EBLUPs of poverty proportions.

The article is organized as follows. Section 2 introduces the unit-level binomial-logit mixed model and the employed fitting algorithm. Section 3 presents the EBPs (which in this article are synthetic because they assume random effects given small area are zero mean) and the plug-in estimators of weighted sums of probabilities. Section 4 gives an approximation to the MSE of the EBP and four estimators. The first two MSE estimators are plug-in derivations of the MSE approximation, without and with bias correction term. The third and fourth MSE estimators are based on parametric bootstrap. Section 5 presents three simulation experiments. The first simulation studies the behavior of the MSM fitting algorithm. The second simulation compares the performances of the EBPs and the plug-in estimators. The third simulation deals with the MSE estimators proposed in Section 4. All use relatively small populations in comparison with real-life applications. Section 6 applies the developed methodology to data from the SLCS2012 using model-based synthetic estimators with small-area level random effects only. The target is the estimation of poverty proportions at county level. Section 7 gives a discussion and some conclusions.

The article contains four appendixes. Appendix A.1 gives the components of the updating equation of the MSM algorithm. Appendix A.2 contains the proof of Proposition 4.1. Appendix A.3 presents some routines for MSE calculation and Appendix A.4 describes approximations of some derivatives needed for MSE calculation.

The article employs the notation $\boldsymbol{a} = \underset{1\le i\le I}{\text{col}}(a_i)$ and $\boldsymbol{b} = \underset{1\le i\le I}{\text{col}'}(b_i)$ for column and row vectors of size I respectively.

2. The Model

This section introduces a unit-level binomial-logit mixed model and its MSM fitting algorithm. Let D be the number of small areas or domains, with $d = 1, \ldots, D$. Let $\{v_d : d = 1, \ldots, D\}$ be a set of independent and identically distributed (i.i.d.) $N(0, 1)$ random effects, which is a reasonable assumption for sufficiently large domains. Note that in this model, random effects at finer levels are not considered. In matrix notation, we have $\boldsymbol{v} = (v_1, \ldots, v_D)' \sim N_D(\boldsymbol{0}, \boldsymbol{I}_D)$, where $\boldsymbol{I}_D$ is the $D \times D$ unit matrix. The target variable y_{dj} represents the jth sample observation from domain d and its distribution, conditioned to the random effect v_d, is

$$y_{dj}|_{v_d} \sim \text{Bin}(m_{dj}, p_{dj}), \quad d = 1, \ldots, D, \quad j = 1, \ldots, n_d, \tag{2.1}$$

where m_{dj} is a known size parameter. The binomial distribution is typically employed for counting numbers of successes. In official statistics, the sample units can be individuals, households, companies and so on. For individuals, we can investigate the presence of a characteristic of interest and the corresponding size parameter is one. For households and companies, we might be interested in counting the number of household members with lactose intolerance or the number of employees with a high salary. Then the size parameter is the number of household members or company workers, respectively. In all cases, we can assume that the size parameters are known quantities for the sampled elements. Note that in the model there is assumed to be no clustering at survey primary sampling unit level or at household level.

For the natural parameter, we assume

$$\eta_{dj} = \log \frac{p_{dj}}{1-p_{dj}} = \boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d, \quad d = 1, \ldots, D, \quad j = 1, \ldots, n_d, \tag{2.2}$$

where $\phi > 0$ is a variance parameter, $\boldsymbol{\beta} = \underset{1\le k\le p}{\text{col}}(\beta_k)$ is the column vector of regression parameters and $\boldsymbol{x}_{dj} = \underset{1\le k\le p}{\text{col}'}(x_{djk})$ is a row vector of auxiliary variables. We also assume that the y_{dj}'s are independent conditioned to $\boldsymbol{v}$. It holds that

$$\begin{aligned} P(y_{dj}|\boldsymbol{v}) = P(y_{dj}|v_d) &= \binom{m_{dj}}{y_{dj}} p_{dj}^{y_{dj}}(1-p_{dj})^{m_{dj}-y_{dj}}, \\ p_{dj} &= \frac{\exp\{\boldsymbol{x}_{dj}\beta + \phi v_d\}}{1+\exp\{\boldsymbol{x}_{dj}\beta + \phi v_d\}}. \end{aligned} \tag{2.3}$$

Let us define $\boldsymbol{y} = \underset{1\le d\le D}{\text{col}}(\boldsymbol{y}_d)$, where $\boldsymbol{y}_d = \underset{1\le j\le n_d}{\text{col}}(y_{dj})$. The marginal distribution of $\boldsymbol{y}$ is

$$P(\boldsymbol{y}) = \prod_{d=1}^{D} \int_R P(\boldsymbol{y}_d|v_d) f(v_d) dv_d,$$

where f is the standard normal probability density function and

$$P(\boldsymbol{y}_d|v_d) = \prod_{j=1}^{n_d} P(y_{dj}|v_d) = \prod_{j=1}^{n_d} \binom{m_{dj}}{y_{dj}} \frac{\exp\{y_{dj}(\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d)\}}{[1+\exp\{\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d\}]^{m_{dj}}}$$

$$= \exp\left\{ \sum_{j=1}^{n_d} \log \binom{m_{dj}}{y_{dj}} + \sum_{j=1}^{n_d} y_{dj}(\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d) - \sum_{j=1}^{n_d} m_{dj} \log[1+\exp\{\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d\}] \right\}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi)'$ be the vector of model parameters. To fit the unit-level binomial-logit mixed model, we employ the MSM algorithm suggested by Jiang (1998). A natural set of equations for applying the method of moments is

$$0 = f_k(\boldsymbol{\theta}) = M_k(\boldsymbol{\theta}) - \hat{M}_k = \sum_{d=1}^{D}\sum_{j=1}^{n_d} E_{\boldsymbol{\theta}}[y_{dj}] x_{djk} - \sum_{d=1}^{D}\sum_{j=1}^{n_d} y_{dj} x_{djk}, \quad k = 1, \ldots p,$$

$$0 = f_{p+1}(\boldsymbol{\theta}) = M_{p+1}(\boldsymbol{\theta}) - \hat{M}_{p+1} = \sum_{d=1}^{D} E_{\boldsymbol{\theta}}\left[y_{d.}^2\right] - \sum_{d=1}^{D} y_{d.}^2,$$

where $y_{d.} = \sum_{j=1}^{n_d} y_{dj}$. For solving this system of nonlinear equations, which involve matching mean and sums of squares, the Newton-Raphson updating formula is

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \boldsymbol{H}^{-1}(\boldsymbol{\theta}^{(r)}) \boldsymbol{f}(\boldsymbol{\theta}^{(r)}), \tag{2.4}$$

where $\theta_1 = \beta_1, \ldots, \theta_p = \beta_p$, $\theta_{p+1} = \phi$ and

$$\boldsymbol{\theta} = \underset{1\le k\le p+1}{\text{col}}(\theta_k), \quad \boldsymbol{f}(\boldsymbol{\theta}) = \underset{1\le k\le p+1}{\text{col}}(f_k(\boldsymbol{\theta})), \quad \boldsymbol{H}(\boldsymbol{\theta}) = \left(\frac{\partial f_k(\boldsymbol{\theta})}{\partial \theta_\ell}\right)_{k,\ell=1,\ldots,p+1}.$$

Appendix A.1 gives the components of vector $\boldsymbol{f}$ and matrix $\boldsymbol{H}$ appearing in (2.4). As the expectations appearing in the components cannot be explicitly calculated, they are

approximated by Monte Carlo simulation. As the algorithm seed for $\boldsymbol{\beta}$, we use $\boldsymbol{\beta}^{(0)} = \tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is the maximum-likelihood estimator under the model without random effects. In that model, the natural parameters are

$$\eta_{dj} = \boldsymbol{x}_{dj}\beta, \quad d = 1, \ldots, D, \quad j = 1, \ldots, n_d.$$

Concerning the variance parameter, we use

$$\phi^{(0)} = \left(\frac{1}{n} \sum_{d=1}^{D} \sum_{j=1}^{n_d} \left(\tilde{\eta}_{dj} - \hat{\eta}_{d.}^{dir} \right)^2 \right)^{1/2}, \quad \hat{\eta}_{d.}^{dir} = \log \frac{\hat{p}_{d.}^{dir}}{1 - \hat{p}_{d.}^{dir}},$$

$$\hat{p}_{d.}^{dir} = \frac{1}{n_d} \sum_{j=1}^{n_d} \frac{y_{dj}}{m_{dj}}, \quad \tilde{\eta}_{dj} = \boldsymbol{x}_{dj}\tilde{\boldsymbol{\beta}},$$

where $n = \sum_{d=1}^{D} n_d$ is the total sample size. A bootstrap algorithm to estimate $\text{var}(\hat{\boldsymbol{\theta}})$ is

1. Fit the Model (2.1)–(2.2) to the sample and calculate $\hat{\boldsymbol{\theta}}$.
2. Generate bootstrap samples $\left\{ y_{dj}^{(b)} : d = 1, \ldots, D, j = 1, \ldots, n_d \right\}$, $b = 1, \ldots, B$, from the fitted model. Fit the Model (2.1)–(2.2) to the bootstrap samples and calculate $\hat{\boldsymbol{\theta}}^{(b)}$, $b = 1, \ldots, B$, and $\bar{\boldsymbol{\theta}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{\theta}}^{(b)}$.
3. Output: $\widehat{\text{var}}_B(\hat{\boldsymbol{\theta}}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}})'$.

3. Empirical Best Prediction

Let us assume that the unit-level binomial-logit mixed model (2.1)–(2.2), with random effects at domain level only, holds for all the units of a population U partitioned into D domains $U_1, \ldots, U_D$ of sizes $N_1, \ldots, N_D$. The best predictor (BP) of $p_{dj} = p_{dj}(\boldsymbol{\theta}, v_d)$ and of the sum of probabilities $\mu_d = \mu_d(\boldsymbol{\theta}, v_d) = \sum_{j=1}^{N_d} p_{dj}$ is

$$\hat{p}_{dj}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[p_{dj} | \boldsymbol{y}_d] = \frac{\int_R \frac{\exp\{\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d\}} P(\boldsymbol{y}_d | v_d) f(v_d) dv_d}{\int_R P(\boldsymbol{y}_d | v_d) f(v_d) dv_d} = \frac{A_{dj}}{C_d},$$

$$\hat{\mu}_d(\boldsymbol{\theta}) = \sum_{j=1}^{N_d} \hat{p}_{dj}(\boldsymbol{\theta}),$$

where f is the standard normal probability density function and

$$A_{dj} = \int_R \frac{\exp\{\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{\boldsymbol{x}_{dj}\boldsymbol{\beta} + \phi v_d\}} \exp\left\{ \phi y_{d.} v_d - \sum_{i=1}^{n_d} m_{di} \log[1 + \exp\{\boldsymbol{x}_{di}\boldsymbol{\beta} + \phi v_d\}] \right\} f(v_d) dv_d,$$

$$C_d = \int_R \exp\left\{ \phi y_{d.} v_d - \sum_{i=1}^{n_d} m_{di} \log[1 + \exp\{\boldsymbol{x}_{di}\boldsymbol{\beta} + \phi v_d\}] \right\} f(v_d) dv_d. \tag{3.1}$$

The EBP of p_{dj} and μ_d are $\hat{p}_{dj}(\hat{\boldsymbol{\theta}})$ and $\hat{\mu}_d(\hat{\boldsymbol{\theta}})$ and they can be approximated by Monte Carlo simulation. If at least one of the auxiliary variables is continuous, the calculation of $\hat{\mu}_d(\hat{\boldsymbol{\theta}})$ requires the availability of census files with the values of $\boldsymbol{x}_{dj}$ for all the units $j \in U_d$. Two data files (survey and census), with the same auxiliary variables, are then needed to

calculate the EBP of μ_d in this setup. In the real data case of Section 6, the x-values are only available for the sample units. The nonavailability of census data is the standard situation for the living conditions public statistics in most countries. This is why we study the special case where the covariates are categorical and take a finite number of values. For this last case, the EBPs can be calculated without having a full unit-record census file, with the caveats that using various cross-tabulations implies a specific, untested census structure, and excludes the use of continuous variables in the survey-based binomial-logit mixed model. However, the external-to-sample information can be obtained more easily from cross tabulations, as detailed unit-record data have limited availability.

Let us assume that $\boldsymbol{x}_{dj} \in \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$ for all d and j. Define

$$\bar{\mu}_d = \frac{\mu_d}{N_d}, \quad \mu_d = \sum_{j=1}^{N_d} p_{dj} = \sum_{k=1}^{K} N_{dk} q_{dk}, \quad q_{dk} = \frac{\exp\{\boldsymbol{z}_k \boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{\boldsymbol{z}_k \boldsymbol{\beta} + \phi v_d\}} \tag{3.2}$$

and $N_{dk} = \#\{\, j \in U_d : \boldsymbol{x}_{dj} = \boldsymbol{z}_k\}$ is the known size of the covariate class $\boldsymbol{z}_k$ at the domain d. The target of this section is the estimation of the parameters defined in (3.2).

The BP of the random effect, v_d, and of the weighted sums of probabilities, μ_d and $\bar{\mu}_d$, are

$$\hat{v}_d(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[v_d|\boldsymbol{y}_d] = \frac{\int_R v_d P(\boldsymbol{y}_d|v_d) f(v_d) dv_d}{\int_R P(\boldsymbol{y}_d|v_d) f(v_d) dv_d} = \frac{A_d^v}{C_d}, \tag{3.3}$$

$$\hat{\mu}_d(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\mu_d|\boldsymbol{y}_d] = \sum_{k=1}^{K} N_{dk} E_{\boldsymbol{\theta}}[q_{dk}|y_d] =: \psi_d(y_{d.}, \boldsymbol{\theta}), \quad \hat{\bar{\mu}}_d(\boldsymbol{\theta}) = \frac{\hat{\mu}_d(\boldsymbol{\theta})}{N_d}, \tag{3.4}$$

where the symbol =: stands for notation,

$$E_{\boldsymbol{\theta}}[q_{dk}|\boldsymbol{y}_d] = \frac{\int_R \frac{\exp\{\boldsymbol{z}_k \boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{\boldsymbol{z}_k \boldsymbol{\beta} + \phi v_d\}} P(\boldsymbol{y}_d|v_d) f(v_d) dv_d}{\int_R P(\boldsymbol{y}_d|v_d) f(v_d) dv_d} = \frac{A_{dk}^z}{C_d}$$

and

$$A_{dk}^z = \int_R \frac{\exp\{\boldsymbol{z}_k \boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{\boldsymbol{z}_k \boldsymbol{\beta} + \phi v_d\}} \exp\left\{ \phi y_{d.} v_d - \sum_{i=1}^{n_d} m_{di} \log[1 + \exp\{\boldsymbol{x}_{di} \boldsymbol{\beta} + \phi v_d\}] \right\} f(v_d) dv_d,$$

$$A_d^v = \int_R v_d \exp\left\{ \phi y_{d.} v_d - \sum_{i=1}^{n_d} m_{di} \log[1 + \exp\{\boldsymbol{x}_{di} \boldsymbol{\beta} + \phi v_d\}] \right\} f(v_d) dv_d,$$

$$C_d = \int_R \exp\left\{ \phi y_{d.} v_d - \sum_{i=1}^{n_d} v_{di} \log[1 + \exp\{\boldsymbol{x}_{di} \boldsymbol{\beta} + \phi v_d\}] \right\} f(v_d) dv_d. \tag{3.5}$$

The EBPs of v_d, μ_d and $\bar{\mu}_d$ are $\hat{v}_d = \hat{v}_d(\hat{\boldsymbol{\theta}})$, $\hat{\mu}_d(\hat{\boldsymbol{\theta}}) = \psi_d(y_{d.}, \hat{\boldsymbol{\theta}})$ and $\hat{\bar{\mu}}_d = \hat{\mu}_d(\hat{\boldsymbol{\theta}})/N_d$, respectively. They can be approximated by Monte Carlo simulation. For $\hat{v}_d(\hat{\boldsymbol{\theta}})$ and $\hat{\mu}_d(\hat{\boldsymbol{\theta}})$, the algorithm steps are

1. Estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\phi})'$ and generate $v_d^{(s)}$ i.i.d. $N(0,1)$, $v_d^{(S+s)} = -v_d^{(s)}$, $s = 1, \ldots, S$, $d = 1, \ldots, D$.
2. Calculate $\hat{v}_d = \hat{A}_d^v / \hat{C}_d$ and $\hat{\mu}_d(\hat{\boldsymbol{\theta}}) = \sum_{k=1}^K N_{dk} \hat{q}_{dk}$, where $\hat{q}_{dk} = \hat{A}_{dk}^z / \hat{C}_d$ and

$$\hat{A}_{dk}^z = \frac{1}{2S} \sum_{s=1}^{2S} \frac{\exp\{z_k \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}\}}{1 + \exp\{z_k \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}\}} \exp\left\{ \hat{\phi} y_{d.} v_d^{(s)} - \sum_{i=1}^{n_d} m_{di} \log\left[1 + \exp\{\boldsymbol{x}_{di} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}\}\right] \right\},$$

$$\hat{A}_d^v = \frac{1}{2S} \sum_{s=1}^{2S} v_d^{(s)} \exp\left\{ \hat{\phi} y_{d.} v_d^{(s)} - \sum_{i=1}^{n_d} m_{di} \log\left[1 + \exp\{\boldsymbol{x}_{di} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}\}\right] \right\},$$

$$\hat{C}_d = \frac{1}{2S} \sum_{s=1}^{2S} \exp\left\{ \hat{\phi} y_{d.} v_d^{(s)} - \sum_{i=1}^{n_d} m_{di} \log\left[1 + \exp\{\boldsymbol{x}_{di} \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}\}\right] \right\}.$$

Note that the census unit-record data $\{\boldsymbol{x}_{dj}\}$ is not needed to construct the EBPs (3.3) and (3.4), provided an implicit structure is assumed for the census data and no continuous variables are to be included in the survey-based model. Here we employ a set of marginal tables containing the population sizes N_{dk}. If there is no sample, that is $n_d = 0$ for a given domain d, then $\hat{v}_d(\hat{\boldsymbol{\theta}}) = 0$ and $\hat{\mu}_d(\hat{\boldsymbol{\theta}})$ can be approximated using a synthetic estimate based on the sampled domains. The method is as follows.

1. As above.
2. Calculate $\hat{\mu}_d(\hat{\boldsymbol{\theta}}) = \sum_{k=1}^K N_{dk} \hat{q}_{dk0}$, where

$$\hat{q}_{dk0} = \frac{1}{2S} \sum_{s=1}^{2S} \left\{ \frac{\exp\{z_k \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}\}}{1 + \exp\{z_k \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}\}} \right\}.$$

The plug-in and the synthetic plug-in estimators of p_{dj}, μ_d and $\bar{\mu}_d = \mu_d / N_d$ do not require running Monte Carlo simulation algorithms. They are

$$\hat{p}_{dj}^{in} = \frac{\exp\{\boldsymbol{x}_{dj} \hat{\boldsymbol{\beta}} + \hat{v}_d\}}{1 + \exp\{\boldsymbol{x}_{dj} \hat{\boldsymbol{\beta}} + \hat{v}_d\}}, \quad \hat{\mu}_d^{in} = \sum_{k=1}^K N_{dk} \frac{\exp\{z_k \hat{\boldsymbol{\beta}} + \hat{v}_d\}}{1 + \exp\{z_k \hat{\boldsymbol{\beta}} + \hat{v}_d\}}, \quad \hat{\bar{\mu}}_d^{in} = \frac{\hat{\mu}_d^{in}}{N_d}, \tag{3.6}$$

$$\hat{p}_{dj}^{syn} = \frac{\exp\{\boldsymbol{x}_{dj} \hat{\boldsymbol{\beta}}\}}{1 + \exp\{\boldsymbol{x}_{dj} \hat{\boldsymbol{\beta}}\}}, \quad \hat{\mu}_d^{syn} = \sum_{k=1}^K N_{dk} \frac{\exp\{z_k \hat{\boldsymbol{\beta}}\}}{1 + \exp\{z_k \hat{\boldsymbol{\beta}}\}}, \quad \hat{\bar{\mu}}_d^{syn} = \frac{\hat{\mu}_d^{syn}}{N_d}. \tag{3.7}$$

Remark 3.1. We can further define the population proportion $\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$. If $m_{dj} = 1$, $d = 1, \ldots, D$, $j = 1, \ldots, N_d$, it holds that

$$E[\bar{Y}_d | v_d] = \frac{1}{N_d} \sum_{j=1}^{N_d} p_{dj} = \sum_{k=1}^K w_{dk} q_{dk} = \bar{\mu}_d, \quad w_{dk} = \frac{N_{dk}}{N_d},$$

$$\mathrm{var}[\bar{Y}_d | v_d] = \frac{1}{N_d^2} \sum_{j=1}^{N_d} p_{dj}(1 - p_{dj}) = \frac{1}{N_d} \sum_{k=1}^K w_{dk} q_{dk}(1 - q_{dk}) = \frac{\kappa_d}{N_d}, \quad 0 \le \kappa_d \le 1/4.$$

By applying the Tchebysheff's inequality, it holds that

$$P(|\bar{Y}_d - \bar{\mu}_d| < \varepsilon | v_d) \geq 1 - \frac{\text{var}[\bar{Y}_d | v_d]}{\varepsilon^2} = 1 - \frac{\kappa_d}{N_d \varepsilon^2}. \tag{3.8}$$

Equation (3.8) implies, for example, that $P(|\bar{Y}_d - \bar{\mu}_d| < \varepsilon | v_d) \geq 0.9$ if $\varepsilon = 10^{-2}$ and $N_d = \kappa_d 10^5$. Further, if $q_{dk} \approx 0.20$ for all k, then $\kappa_d \approx 0.16$ and $N_d \approx 16{,}000$ is smaller than the domain sizes appearing in the application to real data presented in Section 6. Thus, $\bar{\mu}_d$ can be assumed to be a good approximation of $\bar{Y}_d$.

4. The MSE of the EBP

This section presents an approximation and gives four estimators of the MSE of the EBP of $\mu_d = \mu_d(\boldsymbol{\theta}, v_d)$. We assume that $m_{dj} = 1$, $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, and that all the n_ds are bounded to be finite. This is the situation of the application to real data of Section 6. A consequence of the last assumption is that the total sample size n and the number of domains D are of the same order.

The MSE of the EBP can be decomposed into the following form

$$\begin{aligned} MSE(\hat{\mu}_d) &= E[(\hat{\mu}_d(\hat{\boldsymbol{\theta}}) - \mu_d(\boldsymbol{\theta}, v_d))^2] = E[(\{\hat{\mu}_d(\hat{\boldsymbol{\theta}}) - \hat{\mu}_d(\boldsymbol{\theta})\} + \{\hat{\mu}_d(\boldsymbol{\theta}) - \mu_d(\boldsymbol{\theta}, v_d)\})^2] \\ &= E[(\hat{\mu}_d(\hat{\boldsymbol{\theta}}) - \hat{\mu}_d(\boldsymbol{\theta}))^2] + E[(\hat{\mu}_d(\boldsymbol{\theta}) - \mu_d(\boldsymbol{\theta}, v_d))^2]. \end{aligned}$$

The second term of $MSE(\hat{\mu}_d)$ is

$$\begin{aligned} g_d(\boldsymbol{\theta}) &= E[(\hat{\mu}_d(\boldsymbol{\theta}) - \mu_d(\boldsymbol{\theta}, v_d))^2] = E\big[\hat{\mu}_d^2(\boldsymbol{\theta})\big] + E\big[\mu_d^2(\boldsymbol{\theta}, v_d)\big] - 2E[\hat{\mu}_d(\boldsymbol{\theta}) E[\mu_d(\boldsymbol{\theta}, v_d) | \boldsymbol{y}_d]] \\ &= E\big[\mu_d^2(\boldsymbol{\theta}, v_d)\big] - E\big[\hat{\mu}_d^2(\boldsymbol{\theta})\big]. \end{aligned}$$

The first and second terms of $g_d(\boldsymbol{\theta})$ are

$$E\big[\mu_d^2(\boldsymbol{\theta}, v_d)\big] = \int_R \left(\sum_{k=1}^{K} N_{dk} \frac{\exp\{z_k \boldsymbol{\beta} + \phi v_d\}}{1 + \exp\{z_k \boldsymbol{\beta} + \phi v_d\}} \right)^2 f(v_d)\, dv_d,$$

$$E\big[\hat{\mu}_d^2(\boldsymbol{\theta})\big] = E\big[\psi_d^2(y_{d.}, \boldsymbol{\theta})\big] = \sum_{j=0}^{n_d} \psi_d^2(j, \boldsymbol{\theta}) p_d(j, \boldsymbol{\theta}),$$

where ψ_d was defined in (3.4),

$$p_d(j, \boldsymbol{\theta}) = P(y_{d.} = j) = \sum_{\boldsymbol{y}_d \in S_{n_d, j}} P(\boldsymbol{y}_d) = \sum_{\boldsymbol{y}_d \in S_{n_d, j}} \int_R P(\boldsymbol{y}_d | v_d) f(v_d)\, dv_d \tag{4.1}$$

$$= \sum_{\boldsymbol{y}_d \in S_{n_d, j}} \left\{ \exp\left\{ \sum_{i=1}^{n_d} y_i \boldsymbol{x}_{di} \boldsymbol{\beta} \right\} \int_R \exp\left\{ j\phi v_d - \sum_{i=1}^{n_d} \log[1 + \exp\{\boldsymbol{x}_{di} \boldsymbol{\beta} + \phi v_d\}] \right\} f(v_d) dv_d \right\}$$

and $S_{n_d, j} = \{\boldsymbol{y}_d = (y_1, \ldots, y_{n_d}) \in \{0,1\}^{n_d} : y_{d.} = y_1 + \cdots + y_{n_d} = j\}$.

Concerning the first term of $MSE(\hat{\mu}_d)$, we have

$$\hat{\mu}_d(\hat{\boldsymbol{\theta}}) - \hat{\mu}_d(\boldsymbol{\theta}) = \psi_d(y_{d.}, \hat{\boldsymbol{\theta}}) - \psi_d(y_{d.}, \boldsymbol{\theta}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(y_{d.}, \boldsymbol{\theta})\right)' (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

$$+ \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \psi_d(y_{d.}, \boldsymbol{\theta})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2).$$

Hereafter the symbols $o(.), O(.)$ are understood in an appropriate sense, for example in probability. We assume that the $\boldsymbol{x}_{dj}$s fulfill the regularity condition (23) of Jiang (2003). Then, we have $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| = O(D^{-1/2})$ and

$$E\left[(\hat{\mu}_d(\hat{\boldsymbol{\theta}}) - \hat{\mu}_d(\boldsymbol{\theta}))^2\right] = \frac{1}{D} E\left[\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(y_{d.}, \boldsymbol{\theta})\right)' \sqrt{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right)^2\right] + o(1/D).$$

Now consider $\hat{\boldsymbol{\theta}}_{d-}$, an estimator based on $\boldsymbol{y}_{d-} = (\boldsymbol{y}_{d'})_{d' \neq d}$, and write $\hat{\mu}_{d-} = \psi_d(y_{d.}, \hat{\boldsymbol{\theta}}_{d-})$. By the independence of $\boldsymbol{y}_d$ and $\boldsymbol{y}_{d-}$, we have

$$a_d(\boldsymbol{\theta}) = E\left[\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(y_{d.}, \boldsymbol{\theta})\right)' \sqrt{D}(\hat{\boldsymbol{\theta}}_{d-} - \boldsymbol{\theta})\right)^2\right]$$

$$= \sum_{j=1}^{n_d} E\left[\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(y_{d.}, \boldsymbol{\theta})\right)' \sqrt{D}(\hat{\boldsymbol{\theta}}_{d-} - \boldsymbol{\theta})\right)^2 \middle| y_{d.} = j\right] p_d(j, \boldsymbol{\theta})$$

$$= \sum_{j=1}^{n_d} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(j, \boldsymbol{\theta})\right)' \boldsymbol{V}_d(\boldsymbol{\theta}) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(j, \boldsymbol{\theta})\right) p_d(j, \boldsymbol{\theta}),$$

where $\boldsymbol{V}_d(\boldsymbol{\theta}) = DE\left[(\hat{\boldsymbol{\theta}}_{d-} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{d-} - \boldsymbol{\theta})' | y_{d.} = j\right] = DE\left[(\hat{\boldsymbol{\theta}}_{d-} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{d-} - \boldsymbol{\theta})'\right]$. Therefore,

$$MSE(\hat{\mu}_{d-}) = g_d(\boldsymbol{\theta}) + \frac{1}{D} a_d(\boldsymbol{\theta}) + o(1/D).$$

If the $\boldsymbol{x}_{dj}$s also fulfill the regularity conditions (24) and (25) of Jiang (2003), then we may replace $\hat{\boldsymbol{\theta}}_{d-}$ by $\hat{\boldsymbol{\theta}}$, an estimator of $\boldsymbol{\theta}$ based on all data, and we obtain

$$MSE(\hat{\mu}_d) = g_d(\boldsymbol{\theta}) + \frac{1}{D} c_d(\boldsymbol{\theta}) + o(1/D), \tag{4.2}$$

where

$$c_d(\boldsymbol{\theta}) = \sum_{j=1}^{n_d} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(j, \boldsymbol{\theta})\right)' \boldsymbol{V}(\boldsymbol{\theta}) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \psi_d(j, \boldsymbol{\theta})\right) p_d(j, \boldsymbol{\theta}), \quad \boldsymbol{V}(\boldsymbol{\theta}) = DE\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\right].$$

A plug-in estimator of $MSE(\hat{\mu}_d)$ is

$$mse^P(\hat{\mu}_d) = g_d(\hat{\boldsymbol{\theta}}) + \frac{1}{D} c_d(\hat{\boldsymbol{\theta}}).$$

We have that $E[c_d(\hat{\boldsymbol{\theta}}) - c_d(\boldsymbol{\theta})] = o(1)$. However $E[g_d(\hat{\boldsymbol{\theta}}) - g_d(\boldsymbol{\theta})]$ is not of order $o(D^{-1})$.

Let $\hat{\boldsymbol{\theta}}$ be a truncated MM estimator. This is to say

$$\hat{\beta}_k = \begin{cases} -L_n & \text{if } \tilde{\beta}_k < -L_n, \\ \tilde{\beta}_k & \text{if } -L_n < \tilde{\beta}_k < L_n, \\ L_n & \text{if } \tilde{\beta}_k > L_n, \end{cases} \quad \hat{\phi} = \begin{cases} \tilde{\phi} & \text{if } \tilde{\phi} \le L_n, \\ L_n & \text{if } \tilde{\phi} > L_n, \end{cases}$$

where $\tilde{\boldsymbol{\theta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p, \tilde{\phi})'$ is an MM estimator. Under regularity conditions (23)–(25) of Jiang (2003) it can be proved that $E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] = O(D^{-1})$ holds for the MM and for the truncated MM estimator (see Jiang 2003, 123). In what follows, we assume that $E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] = O(D^{-1})$ holds. By the Taylor expansion, we have

$$g_d(\hat{\boldsymbol{\theta}}) = g_d(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}} g_d(\boldsymbol{\theta})\right)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} g_d(\boldsymbol{\theta})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(D^{-1}),$$

and hence

$$E[g_d(\hat{\boldsymbol{\theta}})] = g_d(\boldsymbol{\theta}) + \frac{1}{D} b_d(\boldsymbol{\theta}) + o(D^{-1}),$$

where

$$b_d(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}} g_d(\boldsymbol{\theta})\right)' DE[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] + \frac{1}{2} E\left[\sqrt{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} g_d(\boldsymbol{\theta})\right)\sqrt{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right]. \quad (4.3)$$

Proposition 4.1 gives an approximation to the bias term b_d when $\hat{\boldsymbol{\theta}}$ is the truncated MM estimator. Appendix A.2 gives the proof.

Proposition 4.1. Let $\hat{\boldsymbol{\theta}}$ be the truncated MM estimator. Under regularity conditions (23)–(25) of Jiang (2003), it holds that $b_d(\boldsymbol{\theta}) = B_d(\boldsymbol{\theta}) + o(1)$, where

$$B_d(\boldsymbol{\theta}) = \frac{1}{2}\left\{E[r_{D,d}] - \left(\frac{\partial}{\partial \boldsymbol{\theta}} g_d(\boldsymbol{\theta})\right)'\left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{M}(\boldsymbol{\theta})\right)^{-1} E[\boldsymbol{q}_D]\right\},$$

$$r_{D,d} = \Delta_D' \boldsymbol{R}_d(\boldsymbol{\theta}) \Delta_D, \quad \boldsymbol{R}_d(\boldsymbol{\theta}) = \left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\right)'\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} g_d(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{M}(\boldsymbol{\theta})\right)^{-1},$$

$$\boldsymbol{q}_D = \underset{1 \le k \le p+1}{\text{col}}(q_{Dk}), \quad \boldsymbol{M}(\boldsymbol{\theta}) = \underset{1 \le k \le p+1}{\text{col}}(M_k(\boldsymbol{\theta})), \quad \hat{\boldsymbol{M}} = \underset{1 \le k \le p+1}{\text{col}}(\hat{M}_k),$$

$$q_{Dk} = \Delta_D' \boldsymbol{Q}_k(\boldsymbol{\theta}) \Delta_D, \quad \boldsymbol{Q}_k(\boldsymbol{\theta}) = \left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\right)'\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} M_k(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \theta} \boldsymbol{M}(\boldsymbol{\theta})\right)^{-1},$$

$$\Delta_D = \sqrt{D}(\hat{\boldsymbol{M}} - \boldsymbol{M}(\boldsymbol{\theta})), \quad \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{M}(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}_{k_2}} \boldsymbol{M}_{k_1}(\boldsymbol{\theta})\right)_{k_1, k_2 = 1, \ldots, p+1}.$$

Finally, an order $o(D^{-1})$ theoretical estimator of the MSE with bias correction is

$$\widehat{MSE}(\hat{\mu}_d) = g_d(\hat{\boldsymbol{\theta}}) + \frac{1}{D} c_d(\hat{\boldsymbol{\theta}}) - \frac{1}{D} B_d(\hat{\boldsymbol{\theta}})$$

and the practical estimators, without and with bias correction, are

$$mse^0(\hat{\mu}_d) = \hat{g}_d(\hat{\boldsymbol{\theta}}) + \frac{1}{D} \hat{c}_d(\hat{\boldsymbol{\theta}}), \quad mse^1(\hat{\mu}_d) = mse^0(\hat{\mu}_d) - \frac{1}{D} \hat{B}_d(\hat{\boldsymbol{\theta}}). \tag{4.4}$$

Appendix A.3 gives the Monte Carlo approximations $\hat{g}_d(\hat{\boldsymbol{\theta}})$ and $\hat{c}_d(\hat{\boldsymbol{\theta}})$ of $g_d(\hat{\boldsymbol{\theta}})$ and $c_d(\hat{\boldsymbol{\theta}})$ respectively, and the bootstrap estimator $\hat{B}_d(\hat{\boldsymbol{\theta}})$ of $B_d(\boldsymbol{\theta})$. Appendix A.4 presents formulas of some derivatives needed to evaluate the abovementioned approximations.

Another approach to estimating the MSE is to use a parametric bootstrap. The following procedure calculates a bootstrap and a double-bootstrap bias-corrected estimator of $MSE(\hat{\mu}_d)$.

1. Fit the model to the sample and calculate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\phi})'$.
2. Repeat B_1 times ($b_1 = 1, \ldots, B_1$):
   (a) For $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, generate $v_d^{*(b_1)}$ i.i.d. N(0,1) and calculate

$$p_{dj}^{*(b_1)} = \frac{\exp\{\boldsymbol{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{*(b_1)}\}}{1 + \exp\{\boldsymbol{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{*(b_1)}\}}, \quad y_{dj}^{*(b_1)} \sim \text{Bin}\left(m_{dj}, p_{dj}^{*(b_1)}\right),$$

$$\mu_d^{*(b_1)} = \mu_d\left(\hat{\boldsymbol{\theta}}, v_d^{*(b_1)}\right) = \sum_{k=1}^{K} N_{dk} q_{dk}^{*(b_1)}, \quad q_{dk}^{*(b_1)} = \frac{\exp\{z_k\hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{*(b_1)}\}}{1 + \exp\{z_k\hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{*(b_1)}\}}.$$

   (b) For each bootstrap sample, calculate $\hat{\boldsymbol{\theta}}^{*(b_1)}$ and the EBP $\hat{\mu}_d^{*(b_1)} = \hat{\mu}_d(\hat{\boldsymbol{\theta}}^{*(b_1)})$.
   (c) Repeat B_2 times ($b_2 = 1, \ldots, B_2$):
   i. For $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, generate $v_d^{*(b_1,b_2)}$ i.i.d. N(0,1) and calculate

$$p_{dj}^{*(b_1,b_2)} = \frac{\exp\left\{\boldsymbol{x}_{dj}\hat{\boldsymbol{\beta}}^{*(b_1)} + \hat{\phi}^{*(b_1)} v_d^{*(b_1,b_2)}\right\}}{1 + \exp\left\{\boldsymbol{x}_{dj}\hat{\boldsymbol{\beta}}^{*(b_1)} + \hat{\phi}^{*(b_1)} v_d^{*(b_1,b_2)}\right\}},$$

$$y_{dj}^{*(b_1,b_2)} \sim \text{Bin}\left(m_{dj}, p_{dj}^{*(b_1,b_2)}\right),$$

$$\mu_d^{*(b_1,b_2)} = \mu_d\left(\hat{\boldsymbol{\theta}}^{*(b_1)}, v_d^{*(b_1,b_2)}\right) = \sum_{k=1}^{K} N_{dk} q_{dk}^{*(b_1,b_2)},$$

$$q_{dk}^{*(b_1,b_2)} = \frac{\exp\left\{z_k\hat{\boldsymbol{\beta}}^{*(b_1)} + \hat{\phi}^{*(b_1)} v_d^{*(b_1,b_2)}\right\}}{1 + \exp\left\{z_k\hat{\boldsymbol{\beta}}^{*(b_1)} + \hat{\phi}^{*(b_1)} v_d^{*(b_1,b_2)}\right\}}.$$

   ii. For each bootstrap sample, calculate $\hat{\boldsymbol{\theta}}^{*(b_1,b_2)}$ and the EBP $\hat{\mu}_d^{*(b_1,b_2)} = \hat{\mu}_d(\hat{\boldsymbol{\theta}}^{*(b_1,b_2)})$.
   iii. Output: $mse_d^{*(b_1)} = \frac{1}{B_2}\sum_{b_2=1}^{B_2}\left(\hat{\mu}_d^{*(b_1,b_2)} - \mu_d^{*(b_1,b_2)}\right)^2$.

3. Output:

$$mse^*(\hat{\mu}_d) = \frac{1}{B_1}\sum_{b_1=1}^{B_1}\left(\hat{\mu}_d^{*(b_1)} - \mu_d^{*(b_1)}\right)^2, \quad mse^{**}(\hat{\mu}_d) = 2mse^*(\hat{\mu}_d) - \frac{1}{B_1}\sum_{b_1=1}^{B_1} mse_d^{*(b_1)}.$$

5. Simulation Experiments

In this section we present three simulation experiments. They are fully model-based and are linked to the case study of Section 6, but with much smaller sample and population sizes. All of them use the same simulation environment, which can be described in the following way. Take $N_d = 1{,}000$, $n_d = 5, 10, 20, 40$, $D = 30$. Note that, to enable computation, N_d is more than an order of magnitude smaller than specified in the approximation in probability used in (3.8). For $d = 1, \ldots, D$, $j = 1, \ldots, n_d$, generate regressors, x_{dj1} and x_{dj2}, classifying individuals into one of three possible classes (e.g., inactive, unemployed, and employed), so that they take on values $(x_{dj1}, x_{dj2}) \in \{(0,0),(0,1),(1,0)\}$ with probabilities 0.3, 0.2, and 0.5, respectively. Generate $v_d \sim N(0,1)$, $d = 1, \ldots, D$. Take $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (1/3, -3/2, 1/2)$ and $\phi = 1/2$. For $d = 1, \ldots, D, j = 1, \ldots, n_d$, generate the target variable

$$y_{dj} \sim \text{Bin}(m_{dj}, p_{dj}), \quad p_{dj} = \frac{\exp\{\beta_0 + x_{dj1}\beta_1 + x_{dj2}\beta_2 + \phi v_d\}}{1 + \exp\{\beta_0 + x_{dj1}\beta_1 + x_{dj2}\beta_2 + \phi v_d\}}, \quad m_{dj} = 1,$$

where $y_{dj} = 1$ $(= 0)$ indicates that individual j of domain d is (not) below the poverty line and p_{dj} is the corresponding binomial probability. We choose $D = 30$ as a round figure close to the number $D = 34$ of domains in the real data.

As some of the theoretical results are asymptotic, we investigate cases with small to medium sample sizes. The selected scenario resembles the application to real data. For computational reasons, the population size $N = 30{,}000$ is much smaller than the population size $N = 4{,}990.277$ of the study case. Nevertheless, the simulations are illustrative and give useful information about how the methodology works in practice. Simulation 1 investigates the behavior of the model parameter estimators. Simulation 2 calculates the bias and the MSE of the EBP and the plug-in estimators under different scenarios. Simulation 3 compares the introduced MSE estimators.

5.1. Simulation 1

The target of Simulation 1 is to check the behavior of the fitting algorithm. The steps of Simulation 1 are

1. Repeat K $=$ 1,000 times $(k = 1, \ldots, K)$.
   1.1. Generate a sample of size $n = \sum_{d=1}^{D} n_d$. Calculate $\hat{\beta}_0^{(k)}$, $\hat{\beta}_1^{(k)}$, $\hat{\beta}_2^{(k)}$ and $\hat{\phi}^{(k)}$.
2. Output: For $\theta \in \{\beta_0, \beta_1, \beta_2, \phi\}$, calculate the empirical bias and the root-MSE

$$BIAS = \frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}^{(k)} - \theta), \quad RMSE = \left(\frac{1}{K}\sum_{k=1}^{K}(\hat{\theta}^{(k)} - \theta)^2\right)^{1/2}.$$

Table 1. BIAS (left) and RMSE (right) for $D = 30$.

n	150	300	600	1,200	150	300	600	1,200
n_d	5	10	20	40	5	10	20	40
$\hat{\beta}_0$	0.004	−0.004	−0.003	−0.001	0.274	0.242	0.180	0.156
$\hat{\beta}_1$	−0.002	−0.005	0.006	0.004	0.346	0.301	0.223	0.158
$\hat{\beta}_2$	−0.012	0.010	0.010	0.007	0.474	0.372	0.261	0.187
$\hat{\phi}$	−0.103	−0.058	−0.042	−0.022	0.359	0.270	0.177	0.116

Table 1 presents the obtained results. This table shows that the empirical bias and root-MSE of the MSM estimators of the model parameters decrease as the sample size increases. This is coherent with the consistency property given by Theorem 1 of Jiang (1998).

5.2. *Simulation 2*

The target of Simulation 2 is to investigate the behavior of the EBP, the plug-in (IN) and synthetic plug-in estimators (SYN). The steps of Simulation 2 are

1. Repeat $K = 10{,}000$ times ($k = 1, \ldots, K$)
   1.1. Generate the population in the same way as described at the beginning of this section and calculate

$$\bar{\mu}_d^{(k)} = \frac{1}{N_d}\sum_{j=1}^{N_d} p_{dj}^{(k)}, \quad \bar{Y}_d^{(k)} = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}^{(k)}, \quad \nabla_d^{(k)} = \bar{\mu}_d^{(k)} - \bar{Y}_d^{(k)}.$$

   1.2. For $d = 1, \ldots, D$, select a simple random sample s_d (without replacement) of size n_d and calculate $\hat{\boldsymbol{\theta}}^{(k)} = \left(\hat{\beta}_0^{(k)}, \hat{\beta}_1^{(k)}, \hat{\beta}_2^{(k)}, \hat{\phi}^{(k)}\right)$ and $\tilde{\boldsymbol{\theta}}^{(k)} = \left(\tilde{\beta}_0^{(k)}, \tilde{\beta}_1^{(k)}, \tilde{\beta}_2^{(k)}\right)$ under the logit model with and without random effects.

   1.3. Calculate $\hat{v}_d^{(k)} = \hat{v}_d(\hat{\boldsymbol{\theta}}^{(k)})$ and

$$\hat{\bar{\mu}}_d^{ebp,k} = \frac{\hat{\mu}_d(\hat{\boldsymbol{\theta}}^{(k)})}{N_d}, \quad \hat{\bar{\mu}}_d^{in,k} = \frac{\hat{\mu}_d^{in}(\hat{\boldsymbol{\theta}}^{(k)}, \hat{v}_d^{(k)})}{N_d}, \quad \hat{\bar{\mu}}_d^{syn,k} = \frac{\hat{\mu}_d^{syn}(\tilde{\boldsymbol{\theta}}^{(k)})}{N_d}, \quad d = 1, \ldots, D.$$

2. For each $\hat{\bar{\mu}}_d \in \left\{\hat{\bar{\mu}}_d^{ebp}, \hat{\bar{\mu}}_d^{in}, \hat{\bar{\mu}}_d^{syn}\right\}$, calculate

$$B_d = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{\mu}}_d^{k} - \bar{\mu}_d^{(k)}\right), \quad E_d = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{\mu}}_d^{k} - \bar{\mu}_d^{(k)}\right)^2, \quad d = 1, \ldots, D. \tag{5.1}$$

Figures 1, 2, and 3 present the boxplots of the empirical biases, B_ds, and mean-square errors, E_ds, of the EBP, the IN and the SYN defined in (3.4), (3.6), and (3.7) respectively. The SYN is based on the model without random effects and it is calculated from the corresponding parameter estimates $\tilde{\boldsymbol{\theta}}$. Figures 1 and 3 show that the MSE of the EBP is lower than the MSE of the SYN, overall for domain samples sizes greater than ten. Furthermore, the EBP has slightly lower bias than the SYN. Further, the EBP has slightly lower bias and MSE than the IN.

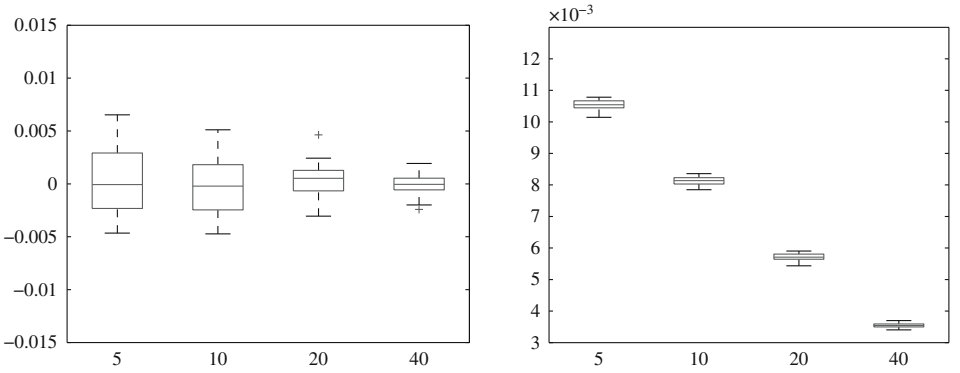


Fig. 1. Boxplots of B_d (left) and E_d (right) defined in (5.1) for EBP and $n_d \in \{5, 10, 20, 40\}$.

Let us illustrate the difference between domain parameters $\bar{\mu}_d^{(k)}$ and $\bar{Y}_d^{(k)}$ in the simulations. As the calculated differences do not depend on d in the simulated scenario, Table 2 presents the quantiles of $\nabla_d^{(k)}$ for $N_d = 1{,}000$ and $d = 1$. In the interquartile range, the absolute difference is lower than 10^{-2}.

5.3. Simulation 3

The target of Simulation 3 is to investigate the behavior of the four MSE estimators, the analytic estimator $mse^0(\hat{\mu}_d)$, the analytic estimator with bias correction $mse^1(\hat{\mu}_d)$, the bootstrap estimator $mse^*(\hat{\mu}_d)$ and the double-bootstrap bias-corrected estimator $mse^{**}(\hat{\mu}_d)$ of the EBP. The number of first-stage bootstrap resamples is $B_1 = 100$. By following Erciulescu and Fuller (2014), the number of second-stage bootstrap resamples is $B_2 = 1$. The steps of Simulation 3 are

1. Repeat $K = 1{,}000$ times ($k = 1, \ldots, K$)
   1.1. Generate the population in the same way as described at the beginning of this section.
   1.2. For $d = 1, \ldots, D$, select a simple random sample s_d (without replacement) of size n_d. Calculate $mse_d^{0(k)} = mse^0(\hat{\mu}_d)/N_d^2$, $mse_d^{1(k)} = mse^1(\hat{\mu}_d)/N_d^2$, $mse_d^{*(k)} = mse^*(\hat{\mu}_d)/N_d^2$ and $mse_d^{**(k)} = mse^{**}(\hat{\mu}_d)/N_d^2$, $d = 1, \ldots, D$.

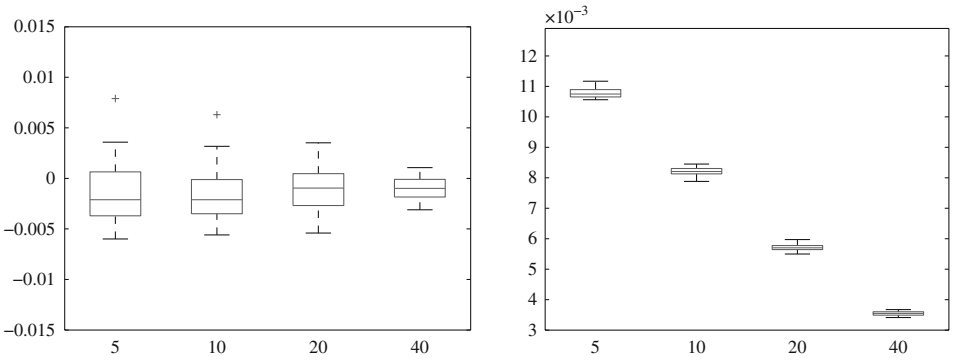


Fig. 2. Boxplots of B_d (left) and E_d (right) defined in (5.1) for IN and $n_d \in \{5, 10, 20, 40\}$.

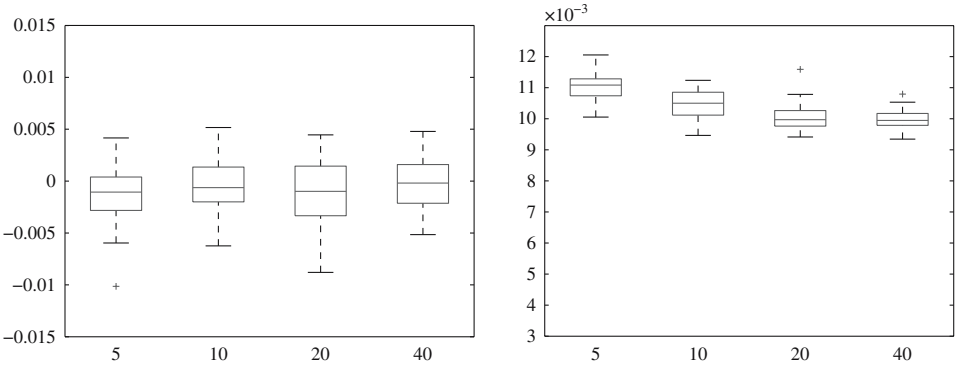


Fig. 3. Boxplots of B_d (left) and E_d (right) defined in (5.1) for SYN and $n_d \in \{5, 10, 20, 40\}$.

2. For every $mse \in \{mse^0, mse^1, mse^*, mse^{**}\}$, calculate

$$mse_d = \frac{1}{K}\sum_{k=1}^{K} mse_d^{(k)}, \quad e_d = \frac{1}{K}\sum_{k=1}^{K}\left(mse_d^{(k)} - E_d^{ebp}\right)^2, \quad d = 1, \ldots, D,$$

where E_d^{ebp}, $d = 1, \ldots, D$, is taken from the output of Simulation 2.

Figure 4 presents the plots of the four MSE estimators. This figure shows that the bias correction used in mse_d^1 as well as in mse_d^{**} does not work for domain sample sizes around $n_d = 10$. It starts to be effective if $n_d \geq 20$, when both mse_d^1 and mse_d^{**} seem to estimate the real value of the simulated mean-squared error E_d^{ebp} quite well.

Figure 5 contains the boxplots of the empirical MSEs, e_d^0, e_d^1, e_d^* and e_d^{**}, of the MSE estimators mse_d^0, mse_d^1, mse_d^* and mse_d^{**}. The analytic estimators, mse_d^0 and mse_d^1, perform better than the parametric bootstrap estimators mse_d^*, mse_d^{**}. Further, the bias-corrected estimator, mse_d^1, has the lowest MSEs. It can be also seen that in the case of the double-bootstrap estimator mse_d^{**} the gained bias correction comes at the cost of increased variability of this estimator.

Remark 5.1. The approximation of terms g_d, c_d and B_d used correspondingly in the mse_d^0 and mse_d^1 estimators entails some implementational and computational difficulties. First of all, to be able to calculate the B_d term one has to implement all formulas for the needed derivatives, which are partly presented in Appendix A.4. A greater difficulty is that the calculation of the probabilities (4.1) and their derivatives appearing in the g_d, c_d and B_d terms is computationally expensive if $n_d \geq 40$, because calculating the sum in (4.1) requires an iterative enumeration of all the elements of the subset $S_{n_d,j} \subset \{0,1\}^{n_d}$ and the size of $S_{n_d,j}$ is $\binom{n_d}{j}$, which increases exponentially with n_d when j is around $n_d/2$, for

Table 2. Quantiles of $\nabla_d^{(k)}$ for $N_d = 1{,}000$, $d = 1$.

∇_1	min	Q_1	Q_2	Q_3	max
Quantiles	−0.0549	−0.0098	−0.0003	0.0093	0.0559

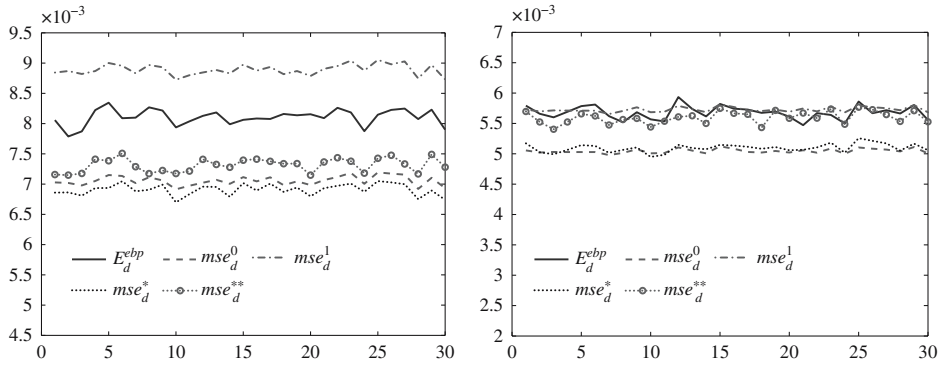


Fig. 4. Plots of E_d^{ebp}, mse_d^0, mse_d^1 and mse_d^* for $n_d = 10$ (left) and $n_d = 20$ (right).

example if $n_d = 40$ the sum over the set $S_{40,20}$ has around $1.4 \cdot 10^{11}$ summands which must be evaluated and counted up for each domain d. This difficulty can be solved in some manner in a simulation experiment when the sizes n_d are small and the same for all d so that the set of all possible $\boldsymbol{y}_d \in S_{n_d,j}$ can be stored once in memory or on a disc. But in practical application when the sizes n_d are different and some of them are large, there are complications with implementation and computation time.

The parametric bootstrap MSE estimators, mse_d^*, mse_d^{**}, avoid the computational problems of their analytic counterparts, mse_d^0 and mse_d^1, and present quite good behavior. To illustrate the performance of the bootstrap estimator mse_d^* for higher values of n_d, in Figure 6 we present results of this estimator for $n_d = 40$. We can observe that for this sample size, the estimator mse_d^* is practically unbiased and the double bootstrap is not needed.

6. Application to SLCS Data

Alleviating poverty is one of the main social tasks in the European Union (EU). Following the instructions of EUROSTAT, European countries implement a Living Conditions

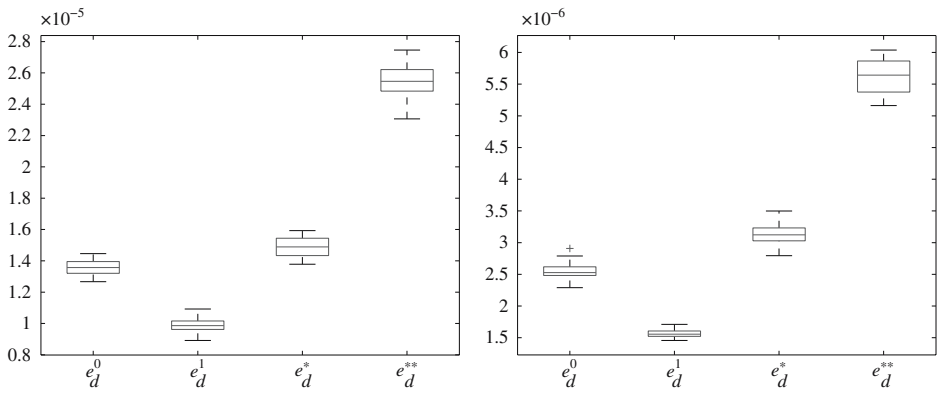


Fig. 5. Plots of e_d^0, e_d^1 and e_d^* for $n_d = 10$ (left) and $n_d = 20$ (right).

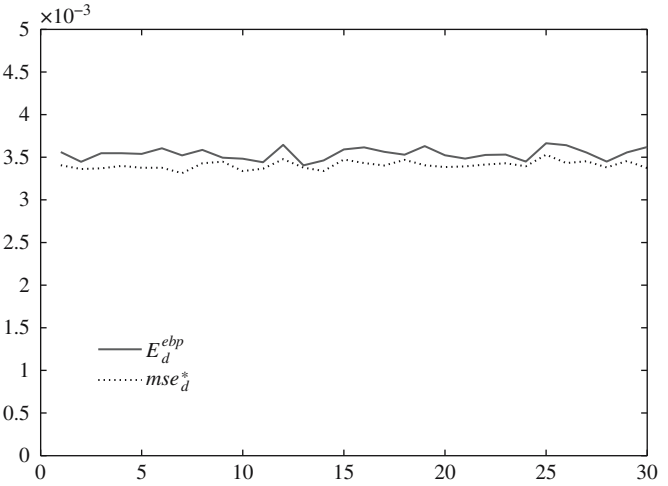


Fig. 6. Plot of E_d^{ebp} and mse_d^* for $n_d = 40$.

Survey to estimate poverty indicators. We use the SLCS2012 data from the Autonomous Community of Valencia (East of Spain). This region has three provinces, Alicante, Castellón and Valencia, encoded as 3, 12, and 46 by the Spanish Statistical Office. The provinces are partitioned into 9, 8, and 17 comarcas (counties) respectively, but only 8, 4, and 14 appear in the SLCS2012. The target domains are the counties, there are thus $D = 34$ domains, but not all of them appear in the sample. The SLCS2012 sample size is $n = 2,678$. The SCLS2012 is a two-stage area sampling design with census section as the primary units and main family addresses as the ultimate sampling units. The sampling frame is the Population Census updated from the Municipal Register.

The SLCS2012 gives information about the equivalent personal incomes, which are obtained by dividing the total household income by the equivalent total of household members. This total is calculated as a weighted sum assigning weights 1 to the first adult, 0.5 to remaining adults and 0.3 to children under 14 years of age. The weighting for obtaining the equivalent household size and income in the SLCS2012 file is done by following the instructions of EUROSTAT. The weights are based on socioeconomic theory and are not sampling weights.

The Spanish Statistical Office builds the data files of the SLCS2012 and assigns the same household equivalent income to all the household members. Because of this fact, individual-level models are not as explanatory as they hypothetically could be. One could think of fitting models at the household level. However, we cannot follow such approach because the domain-level aggregated auxiliary variables, which are used to construct a model-based census, are only available for individuals and not for households.

EUROSTAT defines the poverty line as the 60% of the median of the equivalent personal incomes in the whole country. A person is classified as poor if their equivalent personal income (denoted as E_{dj} for individual j of domain d) is lower than the poverty line. The poverty proportion is the proportion of people below the poverty line. The 2012 poverty line is $z = 6,840$ (in euros per annum per person) for the region of Valencia.

The poverty proportions at the domain levels are

$$\bar{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}, \quad \text{where } y_{dj} = I(E_{dj} < z), \quad d = 1, \ldots, D, \quad j = 1, \ldots, N_d. \tag{6.1}$$

This article estimates domain poverty proportions by using the EBPs of the corresponding weighted sums of probabilities based on unit-level logit mixed models, with random effects at domain level only. This approach requires unit-level survey data for fitting the models, and cross-classified domain level census data for constructing the EBPs. At the unit level, data is taken from the SLCS2012 and the target variable indicates whether individuals are below the poverty line (or not). As the target variable is dichotomic, we employ logit regression models.

In addition to the SLCS2012 data, we take auxiliary aggregated data from the 2012 Labour Force Survey (SLFS2012) file, which contains survey data about the labor market. The sizes of domains crossed by labor status (employed, unemployed, inactive and below 15 years old) are taken from this file. Note that by summing up in the labor categories we obtain the estimated domain sizes. We have taken the estimated domain sizes from SLFS2012 in the estimation of the EBPs. The 2012 population size for the region of Valencia, estimated from SLFS2012, is $N = 4{,}990.277$.

We remark that we have estimated the population sizes N_{dk} using SLFS2012 data and we have ignored their variability. As we have not got good covariates at the county level from Spanish administrative registers, we have instead employed SLFS2012 data and have taken the selected covariates as true aggregated values. This is a drawback of this application to real data, as it leads to underestimates of the MSE. Nevertheless, the sample size of the SLFS2012 is much higher than the one of the SLCS2012. This is why we have followed this practical approach.

For $d = 1, \ldots, D$, $j = 1, \ldots, N_d$, we assume that $y_{dj}|_{v_d} \sim \text{Bin}(1, p_{dj})$, where $v_1, \ldots, v_D$ are i.i.d. $N(0, 1)$ and

$$\text{logit}(p_{dj}) = \beta_0 + \beta_1 \,\text{employed}_{dj} + \beta_2 \,\text{unemployed}_{dj} + \beta_3 \,\text{inactive}_{dj} + \phi v_d. \tag{6.2}$$

Table 3 presents the estimates of the model parameters and the corresponding p-values. We observe that the more people are employed and inactive, the smaller is the probability of being below the poverty line, and the more people are unemployed, the bigger is this probability.

Pearson residuals are calculated for each domain and covariate class. Conditionally on v_d, the sum of y_{dj} over the domain d and the covariate class k is binomially distributed with parameters (N_{dk}, p_{dk}), where N_{dk} is the number of observations in the area d and the covariate class k and p_{dk} is the corresponding model probability.

Figure 7 (left) presents a dispersion graph of residuals. The residuals are mainly located in the interval $(-2, 2)$ and they do not present any visible nonrandom pattern. Figure 7 (right) presents a boxplot of the residuals. The marked residual 7 is an outlier under the standard normal distribution. We have also fitted Model (6.2) to the data without the observations belonging to the subset (domain crossed by covariate class) marked by 7. The obtained parameter estimates do not differ significantly from the one appearing in Table 3. Therefore, we calculate the EBPs with the parameter estimates given by this table.

Table 3. Estimates of model parameters.

	estimate	standard error	p-value
β_0	− 1.1773	0.1473	1.33E-15
β_1	− 0.8354	0.1522	4.02E-08
β_2	0.5379	0.1559	0.000558
β_3	− 0.4040	0.1445	0.005177
ϕ	0.3986		

In order to check the assumption that the random effects at domain level have the standard normal distribution, we calculated the EBP $\hat{v}_d$ of the random effects v_d and we present their Q-Q plot in Figure 8. We can observe quite good agreement and moreover, the Kolmogorov-Smirnov test does not reject the hypothesis $H_0 : F_{\hat{v}_d} = F_{N(0,1)}$ with p-value equal to 0.9653. Because random effects at finer levels are not incorporated into the model, these cannot be tested.

Table 4 presents the direct (dir) and EBP (ebp) estimates of the domain poverty proportions, the corresponding variance (var) and bootstrap MSE (mse) estimates and the relative root mean-squared errors in % (rrmse). We have employed the simple parametric bootstrap with $B = 500$ resamples. The columns labeled n and $\hat{N}$ contain the SLCS2012 sample sizes and the SLFS2012-based estimated population sizes (number of individuals), respectively. The columns labeled prov and com indicate the provinces and the counties, respectively.

Table 4 shows that the EBPs have lower MSEs conditional on the model being correct than the direct estimates in areas with small sample sizes ($n_d < 90$) and comparable or slightly higher MSEs in the rest of domains.

Table 5 presents the EBPs (ebp) of the domain poverty proportions, the corresponding MSE estimates (mse) and the relative root mean-squared errors in % (rrmse) for counties with zero sample size in SLCS2012. For these counties, the direct estimators are not calculable, as there is no sample.

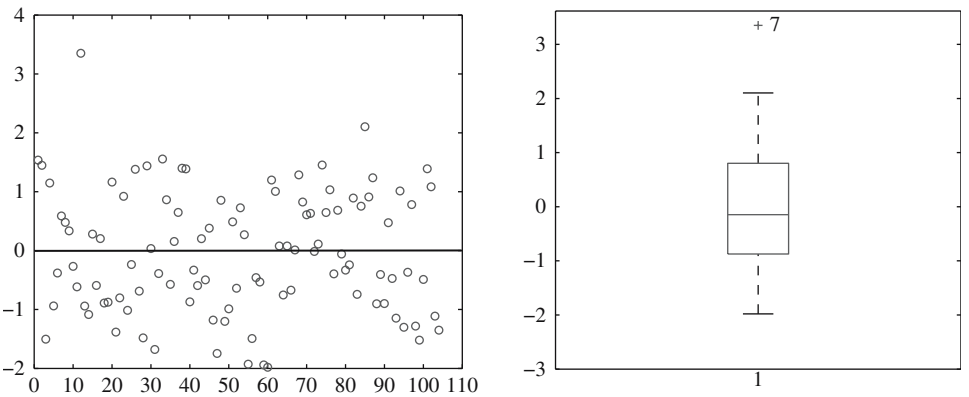


Fig. 7. Dispersion graph (left) and boxplot (right) of the residuals.

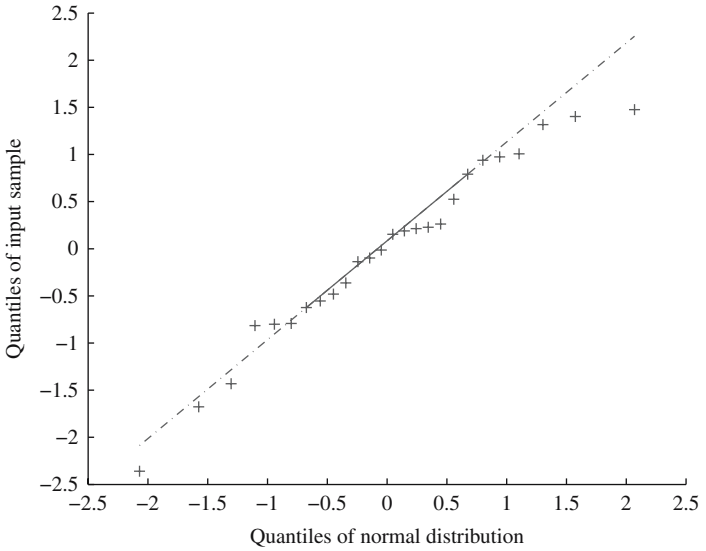


Fig. 8. Q-Q plot of $\hat{v}_d$ with respect to $N(0, 1)$ distribution.

Figure 9 (left) plots EBP and direct poverty-incidence estimates and Figure 9 (right) plots the estimated MSEs of poverty-incidence estimates. These figures show that the EBPs have a lower estimated MSE than the direct estimates.

7. Discussion and Conclusions

Binomial-logit regression models are a flexible class of modelling dichotomic and count variables. This work estimates poverty proportions in counties of the region of Valencia, Spain, by using a model-based unit-level logit mixed model with a domain-level random error (but without reference to survey clustering or survey weighting), and using predictors of weighted sums of probabilities. We fit the model by the method of simulated moments. We consider the EBP and two plug-in estimators and we compare them in a simulation study, based (due to computational complexity) on a small sample and a very small census.

The assumed binomial-logit model with normally distributed domain random effects is widely used in small-area estimation when the parameter of interest is a proportion or a sum of probabilities. The model can be extended from $v_d \sim$ Normal to $v_d \sim F$, where F is a cumulative distribution function with support on the real line. The extension to heteroscedasticity can also be formulated. If we do this however, we would have to study which of the properties of the binomial logit-normal models still hold in the new model, or how they are modified. Working with a well-studied model has the advantage of having mathematical (e.g., asymptotic) results on the model parameter estimators. In particular, some properties established by Jiang (1998, 2003) and Jiang and Lahiri (2001) are employed in the article to derive the approximation to the model-based MSE of the EBP, conditional on a specified model and without random effects at levels finer than domain.

For the EBP, we derive a model-based MSE and introduce four estimators. The first two are analytic estimators without and with bias correction of the second order. The third and

Table 4. Direct and EBP poverty-proportions estimates.

prov	com	n	$\hat{N}$	dir	var	rrmse	ebp	mse	rrmse
46	17	12	35415	0.3085	0.0164	41.52	0.1957	0.0035	30.23
12	7	17	28715	0.3010	0.0113	35.32	0.2231	0.0031	25.11
46	12	18	88625	0.0000	0.0076		0.1384	0.0032	40.94
46	22	18	30672	0.5010	0.0101	20.04	0.2405	0.0030	22.76
3	31	37	151846	0.0133	0.0051	536.71	0.1149	0.0029	46.60
46	18	37	55440	0.1103	0.0049	63.65	0.1582	0.0027	32.76
46	23	45	75157	0.2490	0.0051	28.65	0.1982	0.0026	25.87
46	21	47	53869	0.2454	0.0036	24.53	0.2528	0.0028	21.06
12	3	56	86580	0.2993	0.0034	19.56	0.2796	0.0026	18.34
3	34	66	282076	0.1216	0.0032	46.40	0.1659	0.0025	29.90
46	24	66	86138	0.1251	0.0025	39.82	0.1615	0.0025	31.14
46	25	75	177618	0.1711	0.0031	32.49	0.2009	0.0021	23.04
3	28	82	76849	0.3962	0.0027	13.05	0.1817	0.0021	25.31
3	29	85	228083	0.0868	0.0025	57.58	0.1270	0.0020	35.56
12	6	88	182172	0.1823	0.0020	24.69	0.1855	0.0024	26.42
3	27	92	125445	0.2995	0.0019	14.39	0.2688	0.0022	17.33
46	16	109	139697	0.2995	0.0022	15.58	0.2617	0.0021	17.44
46	11	109	179960	0.1984	0.0016	20.37	0.1543	0.0022	30.71
12	5	117	251429	0.3141	0.0029	17.19	0.1812	0.0021	25.05
46	20	124	260449	0.0223	0.0012	158.11	0.0747	0.0019	58.38
46	13	125	184704	0.1265	0.0014	29.11	0.1456	0.0019	30.25
3	30	147	239247	0.1774	0.0016	22.31	0.2021	0.0016	19.60
3	33	154	266020	0.2967	0.0011	11.20	0.2898	0.0017	14.25
46	14	213	373164	0.2527	0.0010	12.52	0.2509	0.0012	13.76
3	32	298	467602	0.1830	0.0006	13.24	0.1568	0.0011	21.00
46	15	441	781222	0.2134	0.0005	10.08	0.1926	0.0013	18.41

fourth estimators are based on a parametric bootstrap. We analyze the behavior of the proposed estimators in a small simulation study. Estimating the bias correction term is computationally intensive and the results of the analytic estimators without and with bias correction are quite similar for very small domain sample sizes, although this conclusion may require modification if cluster and household level random effects are included or survey weighting used. The MSE analytical estimators are consistent, but they are not practical for medium- and large-domain sample sizes, because the calculation of the term

Table 5. EBP poverty-proportions estimates for counties without sample ($n_d = 0$).

prov	com	$\hat{N}$	ebp	mse	rrmse
3	26	22310	0.1943	0.0034	30.13
12	1	7819	0.1680	0.0024	29.36
12	2	6868	0.1712	0.0029	31.70
12	4	16109	0.1780	0.0022	26.23
12	8	5529	0.1781	0.0035	33.11
46	9	1826	0.1879	0.0029	28.52
46	10	14400	0.1964	0.0036	30.69
46	19	7222	0.1920	0.0029	28.28

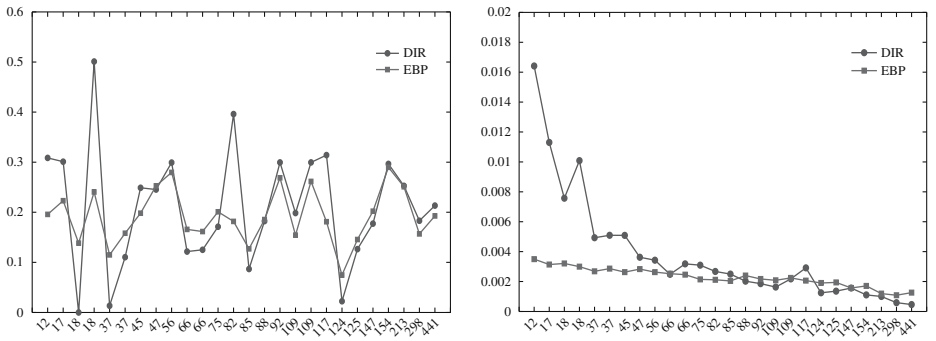


Fig. 9. EBP and direct poverty-proportions estimates (left) and estimated MSEs (right).

g_d involves iteratively finding the $\binom{n_d}{j}$ subsets of size j of a set of size n_d. This problem requires a high-speed computer with large RAM memory for $n_d \geq 40$, so would be difficult to implement for most real-world applications where samples in each surveyed domain are almost always much larger.

Our simulations are computationally intensive in two senses. First, they are carried out at the unit level as we deal with unit-level models and not at the area level. Second and more importantly, all the considered methodology is computationally expensive. The fitting method (MSM) proposed by Jiang (1998) has nice asymptotic properties and allows approximating the MSE of the EBP. However, it requires solving a nonlinear system of equations by using a Newton-Raphson algorithm, where the components of the updating formula have to be approximated by Monte Carlo simulation in each algorithm step. Therefore, it is not a high-speed procedure. The calculation of the EBPs for $d = 1, \ldots, D$, requires evaluating integrals by Monte Carlo simulation. This is again time consuming. The calculation of the MSE estimators (analytic and bootstrap) are even more time consuming than the calculation of the EBPs.

The computational burden is not a major problem for the application of the model to real data (with only one sample) provided domain sizes are kept small, but it is a major problem for the 1,000 samples in the simulation (which is why sample and population sizes have been kept very small there). For example, we parallelized Simulation 3 in several computers and they ran for two weeks before obtaining the results. This is why we have simplified the simulation scenario and have implemented a reduced-in-size version. The small sample and population sizes necessary for the simulation do however reduce general applicability of the results.

Jiang and Lahiri (2001) did not present any application to data. This article shows that the EBP methodology may with further development be applicable to real small-area estimation problems. One exception is Jiang and Lahiri's MSE estimator, which would instead need an alternative computationally feasible approach. As a good alternative, we suggest the introduced parametric bootstrap procedure, which does not have the same computational drawbacks, is easy to implement, and has generally good behavior.

In the application to poverty data from the SLCS2012, we use the EBPs to estimate poverty proportions. We take the model-based MSE for a synthetic estimator estimated via the parametric bootstrap as a performance measure. The units are taken to be individuals,

not households, and the model is fitted without adjustment for the complex structure of the survey design. From the point of view of modelling the survey data, it would be better to work with households as sampling units (instead of individuals), because some survey variables (like income) are household variables. The problem appears when calculating the EBPs. In the Spanish case, we do not have a full unit-record census file and we instead need to rely on categorical covariates only, taken from census cross tabulations. Consequently, the population sizes available in covariates classes crossed by domains are individual based. The corresponding information is not available for households. This is the reason we have treated individuals as sampling units.

The introduced model has only one random effect for domains. In future, we might consider hierarchical models with random effects on the different levels of the hierarchy. This would allow more accurate modelling, including household effects. The model proposed in this article is however purely model-based, without reference to the structure of the complex survey design, so does not include stratification, clustering or survey-based weighting. Therefore, model assumptions need to be comprehensively checked to the extent possible by using diagnostic tools, like graphical methods, testing procedures and residual analysis. For a simple random sample, at unit level the $\{y_{dj}\}$ are independent, conditional on the model and the domain-level random effects. The same is true if we consider stratified random sampling with domains nested within strata. For more complex designs (like a design clustered within domains) this assumption will not hold.

Another issue that remains to be studied is how survey weighting could be included in the analysis. For example, how best to introduce the survey weights in the model-based methodology to reduce the design bias of predictors? The difficult problems linked with survey-design issues for small-area predictors based on binomial-logit mixed models will require further research.

Appendix

A.1. Components of MM Newton-Raphson Algorithm

The MM Newton-Raphson algorithm is specified if we calculate the expectations appearing in $\boldsymbol{f}(\boldsymbol{\theta})$ and its partial derivatives. The expectation of y_{dj} is

$$E_{\boldsymbol{\theta}}[y_{dj}] = E_v[E_{\boldsymbol{\theta}}[y_{dj}|\boldsymbol{v}]] = E_v[m_{dj}p_{dj}]$$

and the corresponding derivatives are

$$\frac{\partial E_{\boldsymbol{\theta}}[y_{dj}]}{\partial \beta_k} = E_v[m_{dj}p_{dj}(1-p_{dj})x_{djk}], \quad \frac{\partial E_{\boldsymbol{\theta}}[y_{dj}]}{\partial \phi} = E_v[m_{dj}p_{dj}(1-p_{dj})v_d].$$

The expectation of $y_{d.}^2$ is $E_{\boldsymbol{\theta}}\left[y_{d.}^2\right] = E_v\left[E_{\boldsymbol{\theta}}\left[y_{d.}^2|\boldsymbol{v}\right]\right]$, where

$$y_{d.}^2 = \sum_{j=1}^{n_d} y_{dj}^2 + \sum_{j_1 \neq j_2}^{n_d} y_{dj_1}y_{dj_2},$$

$$E_{\boldsymbol{\theta}}\left[y_{dj}^2|\boldsymbol{v}\right] = \mathrm{var}_{\boldsymbol{\theta}}[y_{dj}|\boldsymbol{v}] + E_{\boldsymbol{\theta}}^2[y_{dj}|\boldsymbol{v}] = m_{dj}p_{dj}(1-p_{dj}) + m_{dj}^2p_{dj}^2.$$

Therefore, we have

$$E_{\boldsymbol{\theta}}\left[y_{d.}^2\right] = E_v\left[\sum_{j=1}^{n_d} m_{dj}p_{dj}(1-p_{dj}) + \left(\sum_{j=1}^{n_d} m_{dj}p_{dj}\right)^2\right].$$

Let us define $\xi_d = \sum_{j=1}^{n_d} m_{dj}p_{dj}$. The derivatives of $E_{\boldsymbol{\theta}}\left[y_{d.}^2\right]$ are

$$\frac{\partial E_{\boldsymbol{\theta}}\left[y_{d.}^2\right]}{\partial \beta_k} = \sum_{j=1}^{n_d} E_v[m_{dj}p_{dj}(1-p_{dj})\{1-2(p_{dj}-\xi_d)\}x_{djk}],$$

$$\frac{\partial E_{\boldsymbol{\theta}}\left[y_{d.}^2\right]}{\partial \phi} = \sum_{j=1}^{n_d} E_v[m_{dj}p_{dj}(1-p_{dj})\{1-2(p_{dj}-\xi_d)\}v_d].$$

The elements of the matrix of first partial derivatives are

$$H_{k\ell} = \frac{\partial f_k(\boldsymbol{\theta})}{\partial \theta_\ell} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} \frac{\partial E_{\boldsymbol{\theta}}[y_{dj}]}{\partial \theta_\ell} x_{djk}, \quad k = 1, \ldots, p, \ \ell = 1, \ldots, p+1,$$

$$H_{p+1\ell} = \frac{\partial f_{p+1}(\boldsymbol{\theta})}{\partial \theta_\ell} = \sum_{d=1}^{D} \frac{\partial E_{\boldsymbol{\theta}}\left[y_{d.}^2\right]}{\partial \theta_\ell}, \quad \ell = 1, \ldots, p+1,$$

where $\theta_1 = \beta_1, \ldots, \theta_p = \beta_p$, $\theta_{p+1} = \phi$. The expectations appearing in $\boldsymbol{f}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta})$ can be approximated by Monte Carlo simulation.

A.2. Proof of Proposition 4.1

A first-order multivariate Taylor expansion of $\boldsymbol{M}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}$ yields

$$\boldsymbol{M}(\hat{\boldsymbol{\theta}}) = \boldsymbol{M}(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|).$$

Therefore

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta})) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|). \tag{A.1}$$

Let us consider a univariate second order Taylor expansion of $\boldsymbol{M}_k(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}$ and substitute (A.1) in the quadratic term, that is

$$\begin{aligned}\boldsymbol{M}_k(\hat{\boldsymbol{\theta}}) &= \boldsymbol{M}_k(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}_k(\boldsymbol{\theta})\right)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2}M_k(\boldsymbol{\theta})\right) \\ &\quad (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2) \\ &= \boldsymbol{M}_k(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}_k(\boldsymbol{\theta})\right)'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta}))'\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\right)' \\ &\quad \cdot \left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2}M_k(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta})) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2).\end{aligned}$$

The corresponding multivariate Taylor expansion of $\boldsymbol{M}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}$ is

$$\begin{aligned}\boldsymbol{M}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{M}(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}\underset{1\le k\le p+1}{\text{col}}\left((\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta}))'\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\right)'\right.\\ &\quad\left.\cdot\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2}M_k(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta}))\right) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2)\\ &= \boldsymbol{M}(\boldsymbol{\theta}) + \left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2D}\boldsymbol{q}_D + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2).\end{aligned}$$

Therefore

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\left[(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta})) - \frac{1}{2D}\boldsymbol{q}_D\right] + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2). \tag{A.2}$$

By substituting (A.2) in (4.3), we obtain

$$\begin{aligned}b_d(\boldsymbol{\theta}) &= \left(\frac{\partial}{\partial \boldsymbol{\theta}}g_d(\boldsymbol{\theta})\right)' D\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\{E[\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta})] - \frac{1}{2D}E[\boldsymbol{q}_D]\}\\ &\quad + \frac{1}{2}E\left[\sqrt{D}\left[(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta}))' - \frac{1}{2D}\boldsymbol{q}'_D\right]\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\right)'\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2}g_d(\boldsymbol{\theta})\right)\right.\\ &\quad\left.\cdot\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\sqrt{D}\left[(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta})) - \frac{1}{2D}\boldsymbol{q}_D\right]\right] + Do(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2).\end{aligned}$$

On the one hand, we substitute $\boldsymbol{M}(\hat{\boldsymbol{\theta}})$ by $\hat{\boldsymbol{M}}$, so that $E[\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta})] = E[\hat{\boldsymbol{M}} - \boldsymbol{M}(\boldsymbol{\theta})] = 0$. All the quadratic forms in the second summand containing $\boldsymbol{q}_D$ are $o(1)$. Therefore

$$\begin{aligned}b_d(\boldsymbol{\theta}) &= -\frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{\theta}}g_d(\boldsymbol{\theta})\right)'\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}E[\boldsymbol{q}_D]\\ &\quad + \frac{1}{2}E\left[\sqrt{D}(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta}))'\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\right)'\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2}g_d(\boldsymbol{\theta})\right)\right.\\ &\quad\left.\cdot\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}\sqrt{D}(\boldsymbol{M}(\hat{\boldsymbol{\theta}}) - \boldsymbol{M}(\boldsymbol{\theta}))\right] + o(1)\\ &= \frac{1}{2}\left\{E[r_{D,d}] - \left(\frac{\partial}{\partial \boldsymbol{\theta}}g_d(\boldsymbol{\theta})\right)'\left(\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{M}(\boldsymbol{\theta})\right)^{-1}E[\boldsymbol{q}_D]\right\} + o(1) = B_d(\boldsymbol{\theta}) + o(1).\end{aligned}$$

A.3. MSE Components

The MSE estimators $mse^0(\hat{\mu}_d)$ and $mse^1(\hat{\mu}_d)$ given by (4.4) have the components g_d, c_d and B_d. This appendix gives algorithms for approximating them.

The term $g_d(\hat{\boldsymbol{\theta}})$ can be approximated by $\hat{g}_d(\hat{\boldsymbol{\theta}}) = \hat{E}\big[\hat{\mu}_d^2(\hat{\boldsymbol{\theta}})\big] - \hat{E}\big[\hat{\mu}_d^2(\hat{\boldsymbol{\theta}})\big]$, where the expectations are calculated by Monte Carlo simulation. This is to say, generate $v_d^{(s)}$ to be i.i.d. $N(0,1)$ random variables and $v_d^{(S+s)} = -v_d^{(s)}$, $s = 1, \ldots, S$, and calculate

$$\hat{E}\big[\mu_d^2(\hat{\boldsymbol{\theta}})\big] = \frac{1}{2S}\sum_{s=1}^{2S}\left(\sum_{k=1}^{K} N_{dk}\frac{\exp\{z_k\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(s)}\}}{1+\exp\{z_k\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(s)}\}}\right)^2,$$

$$\hat{E}\big[\hat{\mu}_d^2(\hat{\boldsymbol{\theta}})\big] = \sum_{j=0}^{n_d}\hat{\psi}_d^2(j,\hat{\boldsymbol{\theta}})\hat{p}_d(j,\hat{\boldsymbol{\theta}}),$$

where $\hat{\psi}_d(y_{d.},\hat{\boldsymbol{\theta}}) = \sum_{k=1}^{K} N_{dk}\frac{\hat{A}_{dk}^{z}}{\hat{C}_d}$ and

$$\hat{p}_d(j,\hat{\boldsymbol{\theta}}) = \sum_{\boldsymbol{y}\in S_{n_d j}}\left\{\exp\left\{\sum_{i=1}^{n_d} y_i \boldsymbol{x}_{di}\hat{\boldsymbol{\beta}}\right\}\frac{1}{2S}\sum_{s=1}^{2S}\exp\left\{j\hat{\phi}v_d^{(s)} - \sum_{i=1}^{n_d}\log\big[1+\exp\{\boldsymbol{x}_{di}\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(s)}\}\big]\right\}\right\}.$$

The term $c_d(\hat{\boldsymbol{\theta}})$ can be approximated by

$$\hat{c}_d(\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n_d}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\psi}_d(j,\hat{\boldsymbol{\theta}})\right)'\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}})\left(\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\psi}_d(j,\hat{\boldsymbol{\theta}})\right)\hat{p}_d(j,\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}) = D\widehat{\mathrm{var}}_B(\hat{\boldsymbol{\theta}})$.

The bias correction term $B_d(\boldsymbol{\theta})$ can be approximated by parametric bootstrap, that is

1. Fit the model to the sample and calculate $\hat{\boldsymbol{\theta}}$, $\boldsymbol{R}_d(\hat{\boldsymbol{\theta}})$, $\boldsymbol{M}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{Q}_k(\hat{\boldsymbol{\theta}})$.
2. Generate bootstrap samples $\left\{y_{dj}^{(b)} : d = 1, \ldots, D, j = 1, \ldots, n_d\right\}$, $b = 1, \ldots, B$, from the fitted model.
3. For each bootstrap sample, calculate $\Delta_D^{(b)} = \sqrt{D}(\hat{\boldsymbol{M}}^{(b)} - \boldsymbol{M}(\hat{\boldsymbol{\theta}}))$, where $\hat{\boldsymbol{M}}^{(b)} = \underset{1\le k\le p+1}{\mathrm{col}}\left(\hat{M}_k^{(b)}\right)$, $\hat{M}_k^{(b)} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} y_{dj}^{(b)} x_{djk}$, $k = 1, \ldots, p$, $\hat{M}_{p+1}^{(b)} = \sum_{d=1}^{D}\left(y_{d.}^{(b)}\right)^2$, and calculate

$$r_{D,d}^{(b)} = \Delta_D^{(b)\prime}\boldsymbol{R}_d(\hat{\boldsymbol{\theta}})\Delta_D^{(b)}, \quad q_{Dk}^{(b)} = \Delta_D^{(b)\prime}\boldsymbol{Q}_k(\hat{\boldsymbol{\theta}})\Delta_D^{(b)}, \quad \boldsymbol{q}_D^{(b)} = \underset{1\le k\le p+1}{\mathrm{col}}\left(q_{Dk}^{(b)}\right).$$

4. Calculate $\hat{E}_B[r_{D,d}] = \frac{1}{B}\sum_{b=1}^{B} r_{D,d}^{(b)}$, $\hat{E}_B[\boldsymbol{q}_D] = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{q}_D^{(b)}$.
5. Output: $\hat{B}_d(\hat{\boldsymbol{\theta}}) = \frac{1}{2}\left\{\hat{E}_B[r_{D,d}] - \left(\frac{\partial}{\partial\boldsymbol{\theta}}\hat{g}_d(\hat{\boldsymbol{\theta}})\right)'\left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{M}(\hat{\boldsymbol{\theta}})\right)^{-1}\hat{E}_B[\boldsymbol{q}_D]\right\}$.

A.4. Derivatives Needed to Calculate MSE Components

To calculate the approximation of the terms c_d and particularly B_d we need to evaluate first and second order derivatives of $g_d(\boldsymbol{\theta})$ and $M_k(\boldsymbol{\theta})$ (cf. Proposition 4.1.). In this appendix we present formulas for derivatives of these and other necessary terms with respect to parameters β_r, $r = 1, \ldots, p$, in the case of first order and with respect to $\beta_s\beta_r$, $r, s = 1, \ldots, p$, in the case of second order. The derivatives with respect to ϕ, $\beta_r\phi$ and ϕ^2 can be obtained in a similar form and they are omitted here

A.4.1. Derivatives of $\boldsymbol{M}(\boldsymbol{\theta})$

We recall that $\boldsymbol{M}(\boldsymbol{\theta}) = \underset{1\le k\le p+1}{\text{col}}(M_k(\boldsymbol{\theta}))$, where

$$M_k(\boldsymbol{\theta}) = \sum_{d=1}^{D}\sum_{j=1}^{n_d} E_{\boldsymbol{\theta}}[y_{dj}]x_{djk} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} x_{djk}E_v[p_{dj}],$$

$$M_{p+1}(\boldsymbol{\theta}) = \sum_{d=1}^{D} E_{\boldsymbol{\theta}}\left[y_{d.}^2\right] = \sum_{d=1}^{D} E_v\left[\sum_{j=1}^{n_d} p_{dj}(1-p_{dj}) + \left(\sum_{j=1}^{n_d} p_{dj}\right)^2\right].$$

Let us define $\xi_d = \sum_{j=1}^{n_d} p_{dj}$ for p_{dj} given in (2.3) and note that

$$\frac{\partial p_{dj}}{\partial \beta_r} = x_{djr}p_{dj}(1-p_{dj}), \quad \frac{\partial p_{dj}(1-p_{dj})}{\partial \beta_r} = x_{djr}p_{dj}(1-p_{dj})(1-2p_{dj})$$

and

$$\frac{\partial(p_{dj}-\xi_d)}{\partial \beta_s} = x_{djs}p_{dj}(1-p_{dj}) - \sum_{i=1}^{n_d} p_{di}(1-p_{di})x_{dis}.$$

The first and second order partial derivatives of $M_k(\boldsymbol{\theta})$ are for $k, r, s = 1, \ldots, p$

$$\frac{\partial M_k}{\partial \beta_r} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} x_{djk}x_{djr}E_v[p_{dj}(1-p_{dj})],$$

$$\frac{\partial M_{p+1}}{\partial \beta_r} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} x_{djr}E_v[p_{dj}(1-p_{dj})\{1-2(p_{dj}-\xi_d)\}],$$

$$\frac{\partial^2 M_k}{\partial \beta_s \partial \beta_r} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} x_{djk}x_{djr}x_{djs}E_v[p_{dj}(1-p_{dj})(1-2p_{dj})],$$

$$\frac{\partial^2 M_{p+1}}{\partial \beta_s \partial \beta_r} = \sum_{d=1}^{D}\sum_{j=1}^{n_d} x_{djr}E_v[x_{djs}p_{dj}(1-p_{dj})(1-2p_{dj})\{1-2(p_{dj}-\xi_d)\}$$
$$- 2p_{dj}^2(1-p_{dj})^2x_{djs} + 2p_{dj}(1-p_{dj})\sum_{i=1}^{n_d} p_{di}(1-p_{di})x_{dis}].$$

A.4.2. Derivatives of $C_d(j, \boldsymbol{\theta})$ and $A_{dk}^z(j, \boldsymbol{\theta})$

Let us denote

$$R_d(\boldsymbol{\theta}, j, v_d) = R_d(j) = \exp\left\{ j\phi v_d - \sum_{i=1}^{n_d} \log[1+\exp\{\boldsymbol{x}_{di}\boldsymbol{\beta} + \phi v_d\}]\right\}.$$

Then C_d, defined in (3.1), can be written for $y_{d.} = j$ in the form

$$C_d(j, \boldsymbol{\theta}) = C_d(j) = \int_R R_d(\boldsymbol{\theta}, j, v_d)f(v_d)\,dv_d \tag{A.3}$$

and its first and second order partial derivatives are

$$\frac{\partial C_d(j)}{\partial \beta_r} = -\int_R R_d(j)\left(\sum_{i=1}^{n_d} p_{di}x_{dir}\right)f(v_d)\,dv_d,$$

$$\frac{\partial^2 C_d(j)}{\partial \beta_s \partial \beta_r} = \int_R R_d(j)\left\{\left(\sum_{i=1}^{n_d} p_{di}x_{dir}\right)\left(\sum_{i=1}^{n_d} p_{di}x_{dis}\right) - \sum_{i=1}^{n_d} p_{di}(1-p_{di})x_{dir}x_{dis}\right\}f(v_d)\,dv_d.$$

The term A_{dk}^z, defined in (3.5), can be expressed for $y_{d\cdot} = j$ and q_{dk}, given in (3.2), as

$$A_{dk}^z(j,\boldsymbol{\theta}) = A_{dk}^z(j) = \int_R q_{dk}R_d(\boldsymbol{\theta},j,v_d)f(v_d)\,dv_d$$

and its first and second order partial derivatives are

$$\frac{\partial A_{dk}^z(j)}{\partial \beta_r} = \int_R q_{dk}R_d(j)\left\{(1-q_{dk})z_{kr} - \sum_{i=1}^{n_d} p_{di}x_{dir}\right\}f(v_d)\,dv_d,$$

$$\begin{aligned}\frac{\partial^2 A_{dk}^z(j)}{\partial \beta_s \partial \beta_r} = &\int_R q_{dk}(1-q_{dk})z_{ks}R_d(j)\left\{(1-q_{dk})z_{kr} - \sum_{i=1}^{n_d} p_{di}x_{dir}\right\}f(v_d)\,dv_d \\ &- \int_R q_{dk}R_d(j)\left(\sum_{i=1}^{n_d} p_{di}x_{dis}\right)\left\{(1-q_{dk})z_{kr} - \sum_{i=1}^{n_d} p_{di}x_{dir}\right\}f(v_d)\,dv_d \\ &+ \int_R q_{dk}R_d(j)\left\{-q_{dk}(1-q_{dk})z_{kr}z_{ks} - \sum_{i=1}^{n_d} p_{di}(1-p_{di})x_{dir}x_{dis}\right\}f(v_d)\,dv_d.\end{aligned}$$

A.4.3. Derivatives of $p_d(j,\boldsymbol{\theta})$

From the definitions of $p_d(j,\boldsymbol{\theta}) = p_d(j)$ in (4.1) and of $C_d(j,\boldsymbol{\theta}) = C_d(j)$ in (A.3), it follows that

$$p_d(j,\boldsymbol{\theta}) = \sum_{\boldsymbol{y}\in S_{n_d\cdot j}}\left[\exp\left\{\sum_{i=1}^{n_d} y_i\boldsymbol{x}_{di}\boldsymbol{\beta}\right\}C_d(j,\boldsymbol{\theta})\right].$$

So that the first- and second-order partial derivatives of $p_d(j)$ are

$$\frac{\partial p_d(j)}{\partial \beta_r} = \sum_{\boldsymbol{y}\in S_{n_d\cdot j}}\exp\left\{\sum_{i=1}^{n_d} y_i\boldsymbol{x}_{di}\boldsymbol{\beta}\right\}\left[\left(\sum_{i=1}^{n_d} y_i x_{dir}\right)C_d(j) + \frac{\partial C_d(j)}{\partial \beta_r}\right],$$

$$\frac{\partial^2 p_d(j)}{\partial \beta_s \partial \beta_r} = \sum_{\boldsymbol{y} \in S_{n_d j}} \exp\left\{ \sum_{i=1}^{n_d} y_i \boldsymbol{x}_{di} \boldsymbol{\beta} \right\} \left[\left(\sum_{i=1}^{n_d} y_i x_{dir} \right) \left(\sum_{i=1}^{n_d} y_i x_{dis} \right) C_d(j) \right.$$

$$\left. + \left(\sum_{i=1}^{n_d} y_i x_{dir} \right) \frac{\partial C_d(j)}{\partial \beta_s} + \left(\sum_{i=1}^{n_d} y_i x_{dis} \right) \frac{\partial C_d(j)}{\partial \beta_r} + \frac{\partial^2 C_d(j)}{\partial \beta_s \partial \beta_r} \right].$$

A.4.4. Derivatives of $\psi_d(y_d, \boldsymbol{\theta})$

The first- and second-order partial derivatives of

$$\psi_d(j, \boldsymbol{\theta}) = \psi_d(j) = \sum_{k=1}^{K} N_{dk} \frac{A_{dk}^z(j)}{C_d(j)},$$

defined in (3.4), are

$$\frac{\partial \psi_d(j, \boldsymbol{\theta})}{\partial \theta_r} = \sum_{k=1}^{K} N_{dk} \left\{ \frac{\frac{\partial A_{dk}^z(j)}{\partial \theta_r}}{C_d(j)} - \frac{A_{dk}^z(j) \frac{\partial C_d(j)}{\partial \theta_r}}{C_d^2(j)} \right\},$$

$$\frac{\partial^2 \psi_d(j, \boldsymbol{\theta})}{\partial \theta_s \partial \theta_r} = \sum_{k=1}^{K} N_{dk} \left\{ \frac{\frac{\partial^2 A_{dk}^z(j)}{\partial \theta_s \partial \theta_r}}{C_d(j)} - \frac{\frac{\partial A_{dk}^z(j)}{\partial \theta_r} \frac{\partial C_d(j)}{\partial \theta_s} + \frac{\partial A_{dk}^z(j)}{\partial \theta_s} \frac{\partial C_d(j)}{\partial \theta_r} + A_{dk}^z(j) \frac{\partial^2 C_d(j)}{\partial \theta_s \partial \theta_r}}{C_d^2(j)} \right.$$

$$\left. + \frac{2 A_{dk}^z(j) \frac{\partial C_d(j)}{\partial \theta_s} \frac{\partial C_d(j)}{\partial \theta_r}}{C_d^3(j)} \right\}.$$

A.4.5. Derivatives of $g_d(\boldsymbol{\theta})$

We recall that

$$g_d(\boldsymbol{\theta}) = \int_R \left(\sum_{k=1}^{K} N_{dk} q_{dk}(\boldsymbol{\theta}, v_d) \right)^2 f(v_d)\, dv_d - \sum_{j=0}^{n_d} \psi_d^2(j, \boldsymbol{\theta}) p_d(j, \boldsymbol{\theta}).$$

The first order partial derivatives of $g_d(\boldsymbol{\theta})$ are

$$\frac{\partial g_d(\boldsymbol{\theta})}{\partial \beta_r} = 2 \int_R \left(\sum_{k=1}^{K} N_{dk} q_{dk} \right) \left(\sum_{k=1}^{K} N_{dk} q_{dk} (1 - q_{dk}) z_{kr} \right) f(v_d)\, dv_d$$

$$- 2 \sum_{j=0}^{n_d} \psi_d(j) \frac{\partial \psi_d(j)}{\partial \beta_r} p_d(j) - \sum_{j=0}^{n_d} \psi_d^2(j) \frac{\partial p_d(j)}{\partial \beta_r}.$$

The second order partial derivatives of $g_d(\boldsymbol{\theta})$ are

$$\frac{\partial^2 g_d(\boldsymbol{\theta})}{\partial\beta_s\partial\beta_r} = 2\int_R \left(\sum_{k=1}^{K} N_{dk}q_{dk}(1-q_{dk})z_{ks}\right)\left(\sum_{k=1}^{K} N_{dk}q_{dk}(1-q_{dk})z_{kr}\right) f(v_d)\,dv_d$$

$$+ 2\int_R \left(\sum_{k=1}^{K} N_{dk}q_{dk}\right)\left(\sum_{k=1}^{K} N_{dk}q_{dk}(1-q_{dk})(1-2q_{dk})z_{kr}z_{ks}\right) f(v_d)\,dv_d$$

$$- 2\sum_{j=0}^{n_d} \frac{\partial\psi_d(j)}{\partial\beta_s}\frac{\partial\psi_d(j)}{\partial\beta_r}p_d(j) - 2\sum_{j=0}^{n_d}\psi_d(j)\frac{\partial^2\psi_d(j)}{\partial\beta_s\partial\beta_r}p_d(j) - 2\sum_{j=0}^{n_d}\psi_d(j)\frac{\partial\psi_d(j)}{\partial\beta_r}\frac{\partial p_d(j)}{\partial\beta_s}$$

$$- 2\sum_{j=0}^{n_d}\psi_d(j)\frac{\partial\psi_d(j)}{\partial\beta_s}\frac{\partial p_d(j)}{\partial\beta_r} - \sum_{j=0}^{n_d}\psi_d^2(j)\frac{\partial^2 p_d(j)}{\partial\beta_s\partial\beta_r}.$$

A.4.6. Approximations of the Derivatives

The integrals appearing in the described derivatives can be approximated by Monte Carlo simulation. To illustrate the procedure, we present the corresponding formulas used to approximate the derivatives of the term $C_d(j)$ derived in Section A.4.2. The approximations of the derivatives of the remaining terms can be done in a similar way.

For $s = 1, \ldots, S$, let $v_d^{(s)}$ be i.i.d. $N(0,1)$ random variables and $v_d^{(S+s)} = -v_d^{(s)}$. The partial derivatives of $C_d(j)$ can be approximated as follows:

$$\frac{\partial\hat{C}_d(j)}{\partial\beta_r} = -\frac{1}{2S}\sum_{s=1}^{2S}\hat{R}_d^{(s)}(j)\left(\sum_{i=1}^{n_d}\hat{p}_{di}^{(s)}x_{dir}\right),$$

$$\frac{\partial^2\hat{C}_d(j)}{\partial\beta_s\partial\beta_r} = \frac{1}{2S}\sum_{s=1}^{2S}\hat{R}_d^{(s)}(j)\left\{\left(\sum_{i=1}^{n_d}\hat{p}_{di}^{(s)}x_{dir}\right)\left(\sum_{i=1}^{n_d}\hat{p}_{di}^{(s)}x_{dis}\right)\right.$$

$$\left. - \sum_{i=1}^{n_d}\hat{p}_{di}^{(s)}\left(1-\hat{p}_{di}^{(s)}\right)x_{dir}x_{dis}\right\},$$

where

$$\hat{R}_d^{(s)}(j) = R\left(\hat{\boldsymbol{\theta}}, j, v_d^{(s)}\right) \quad \text{and} \quad \hat{p}_{di}^{(s)} = \frac{\exp\left\{\boldsymbol{x}_{di}\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(s)}\right\}}{1+\exp\left\{\boldsymbol{x}_{di}\hat{\boldsymbol{\beta}} + \hat{\phi}v_d^{(s)}\right\}}.$$

8. References

Chambers, R., N. Salvati, and N. Tzavidis. 2012. *M-Quantile Regression for Binary Data with Application to Small Area Estimation*. Working Paper 12-12, 2012, 24. Centre for Statistical and Survey Methodology, University of Wollongong. Available at: http://ro.uow.edu.au/cssmwp/101 (accessed January 2016).

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro-level Estimation of Poverty and Inequality." *Econometrica* 71: 355–364. Doi: http://dx.doi.org/10.1111/1468-0262.00399.

Erciulescu, A.L. and W. Fuller. 2014. "Parametric Bootstrap Procedures for Small Area Prediction Variance." In Proceedings of the Joint Statistical Meeting - Survey Research Methods Section, August 6, 2014, Boston. 3307–3318. Available at: https://www. amstat.org/sections/srms/proceedings/y2014/Files/31294890280.pdf (accessed January 2016).

Esteban, M.D., D. Morales, A. Pérez, and L. Santamaría. 2012a. "Two Area-Level Time Models for Estimating Small Area Poverty Indicators." *Journal of the Indian Society of Agricultural Statistics* 66: 75–89.

Esteban, M.D., D. Morales, A. Pérez, and L. Santamaría. 2012b. "Small Area Estimation of Poverty Proportions under Area-Level Time Models." *Computational Statistics and Data Analysis* 56: 2840–2855. Doi: http://dx.doi.org/10.1016/j.csda.2011.10.015.

Farrell, P., B. MacGibbon, and T. Tomberlin. 1997. "Bootstrap Adjustments for Empirical Bayes Interval Estimates of Small Area Proportions." *Canadian Journal of Statistics* 25: 75–89. Doi: http://dx.doi.org/10.2307/3315358.

Ghosh, M. and J. Rao. 1994. "Small Area Estimation: An Appraisal." *Statistical Science* 9: 55–93.

González-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2007. "Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters under a Logistic Mixed Model." *Computational Statistics and Data Analysis* 51: 2720–2733. Doi: http://dx.doi.org/10.1016/j.csda.2006.01.012.

González-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2008a. "Bootstrap Mean Squared Error of Small-Area EBLUP." *Journal of Statistical Computation and Simulation* 78: 443–462. Doi: http://dx.doi.org/10.1080/00949650601141811.

González-Manteiga, W., M.J. Lombardía, I. Molina, D. Morales, and L. Santamaría. 2008b. "Analytic and Bootstrap Approximations of Prediction Errors under a Multivariate Fay-Herriot Model." *Computational Statistics and Data Analysis* 52: 5242–5252. Doi: http://dx.doi.org/10.1016/j.csda.2008.04.031.

Hall, P. and T. Maiti. 2006a. "Nonparametric Estimation of Mean-Squared Prediction Error in Nested-Error Regression Models." *The Annals of Statistics* 34: 1733–1750. Doi: http://dx.doi.org/10.1214/009053606000000579.

Hall, P. and T. Maiti. 2006b. "On Parametric Bootstrap Methods for Small Area Prediction." *Journal of the Royal Statistical Society, Series B* 68: 221–238. Doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00541.x.

Jiang, J. 1998. "Consistent Estimators in Generalized Linear Models." *Journal of the American Statistical Association* 93: 720–729. Doi: http://dx.doi.org/10.2307/2670122.

Jiang, J. 2003. "Empirical Best Prediction for Small-Area Inference Based on Generalized Linear Mixed Models." *Journal of Statistical Planning and Inference* 111: 117–127. Doi: http://dx.doi.org/10.1016/S0378-3758(02)00293-8.

Jiang, J. and P. Lahiri. 2001. "Empirical Best Prediction for Small Area Inference with Binary Data." *Annals of the Institute of Statistical Mathematics* 53: 217–243. Doi: http://dx.doi.org/10.1023/A:1012410420337.

Jiang, J. and P. Lahiri. 2006. "Mixed Model Prediction and Small Area Estimation." *Test* 15: 1–96. Doi: http://dx.doi.org/10.1007/BF02595419.

Malec, D., J. Sedransk, C. Moriarity, and F. LeClere. 1997. "Small Area Inference for Binary Variables in the National Health Interview Survey." *Journal of the American Statistical Association* 92: 815–826. Doi: http://dx.doi.org/10.1080/01621459.1997.10474037.

Marhuenda, Y., I. Molina, and D. Morales. 2013. "Small Area Estimation with Spatio-Temporal Fay-Herriot Models." *Computational Statistics and Data Analysis* 58: 308–325. Doi: http://dx.doi.org/10.1016/j.csda.2012.09.002.

Molina, I., B. Nandram, and J.N.K. Rao. 2014. "Small Area Estimation of General Parameters with Application to Poverty Indicators: A Hierarchical Bayes Approach." *The Annals of Applied Statistics* 8: 852–885. Doi: http://dx.doi.org/10.1214/13-AOAS702.

Molina, I. and J.N.K. Rao. 2010. "Small Area Estimation of Poverty Indicators." *The Canadian Journal of Statistics* 38: 369–385. Doi: http://dx.doi.org/10.1002/cjs.10051.

Morales, D., M.C. Pagliarella, and R. Salvatore. 2015. "Small Area Estimation of Poverty Indicators under Partitioned Area-Level Time Models." *SORT – Statistics and Operations Research Transactions* 39: 19–34.

Pfeffermann, D. 2002. "Small Area Estimation – New Developments and Directions." *International Statistical Review* 70: 125–143. Doi: http://dx.doi.org/10.2307/1403729.

Pfeffermann, D. 2013. "New Important Developments in Small Area Estimation." *Statistical Science* 28: 1–134. Doi: http://dx.doi.org/10.1214/12-STS395.

Rao, J.N.K. 1999. "Some Recent Advances in Model-Based Small Area Estimation." *Survey Methodology* 25: 175–186.

Rao, J.N.K. 2003. *Small Area Estimation*. New York: Wiley.

Rao, J.N.K. and I. Molina. 2015. *Small Area Estimation*, 2nd ed. New York: Wiley.

Received February 2015
Revised January 2016
Accepted May 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 693–718, http://dx.doi.org/10.1515/JOS-2016-0035

A Simulation Study of Weighting Methods to Improve Labour-Force Estimates of Immigrants in Ireland

Nancy Duong Nguyen[1], *Órlaith Burke*[2] *and Patrick Murphy*[3]

As immigration has become a global phenomenon in recent years, a number of European countries, including Ireland, have experienced an influx of immigrants, causing a shift in their national demographics. Therefore, it is important that the EU-LFS yield reliable labour-force estimates not only for the whole population, but also for the immigrant population.

This article uses simulation techniques to compare the effectiveness of four different weighting mechanisms in order to improve the precision of the labour-force estimates from the Irish component of the European Union Labour Force Survey (EU-LFS) called the Quarterly National Household Survey (QNHS). The four weighting methodologies for comparison include the original and the current weighting scheme of the QNHS as well as our two proposed alternative weighting schemes. The simulation results show that by modifying the current QNHS weighting mechanism, we can improve the accuracy of the labour-force estimates of the immigrant population in Ireland without affecting the estimates of the whole population and the Irish nationals.

This article highlights potential issues that other countries with new immigrant populations may face when using the EU-LFS for immigration research, and our recommendations may be useful to researchers and national statistical offices in such countries.

Key words: Quarterly National Household Survey; calibrated weights; poststratification; raking ratio; nonresponse.

1. Introduction

During the past two decades, Ireland has experienced large-scale immigration, especially following the enlargement of the European Union (EU) in 2004. Along with the United Kingdom (UK) and Sweden, Ireland was one of only three Old Member States (OMS) that allowed nationals from New Member States (NMS) to access its labour market directly. That resulted in an influx of immigrants from the accession countries to Ireland after 2004. By 2014, approximately twelve per cent of its population were foreign nationals, putting Ireland in sixth place (after Luxembourg, Latvia, Cyprus, Estonia, and Austria) among the

[1] School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland. Email: duong.nguyen@ucdconnect.ie
[2] Nuffield Department of Population Health, University of Oxford, Richard Doll Building, Old Road Campus, Oxford OX3 7LF, United Kingdom. Email: orlaith.burke@ndph.ox.ac.uk
[3] School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland. Email: patrick.murphy@ucd.ie

Acknowledgments: We would like to thank the Irish Social Science Data Archive (www.ucd.ie/issda) and the Irish Central Statistics Office (www.cso.ie) for providing us with the relevant data sets and responding to our enquiries while we work on this paper. This work is supported by the Research Demonstratorship grant from the School of Mathematics and Statistics, University College Dublin.

28 EU countries for the highest proportion of non-nationals in the population (Central Statistics Office 2015a; Eurostat 2015). Therefore, understanding Ireland's immigrants plays an important role in understanding Ireland's population as a whole.

Of all the national surveys in Ireland, the Quarterly National Household Survey (QNHS), conducted by the Central Statistics Office (CSO), is most widely used for immigration research. The QNHS is the Irish component of the EU Labour Force Survey (LFS) with the primary purpose of producing official statistics on the labour force in Ireland. Considering the significant number of foreign nationals living in Ireland and the growing literature on their assimilation into the Irish society (for example: Barrett and Duffy 2008; O'Connell and McGinnity 2008; Barrett et al. 2011; Kingston et al. 2013), it is important for the QNHS to produce reliable estimates on the labour-market participation of immigrants. This can be achieved by ensuring the representativeness of the QNHS samples not only for the whole population of Ireland, but also for the main nationality groups.

Being a voluntary sample survey, the QNHS suffers from nonresponse and other sampling and nonsampling errors, leading to unrepresentative samples. To account for this, the CSO constructs weights for the QNHS such that weighted samples match population estimates on a number of variables of interest. Since the introduction of the QNHS in 1997, its weighting scheme was modified once in the third quarter (Q3) of 2006 to reflect the change in Ireland's demographics following the EU enlargement. The effectiveness of the pre-Q3-2006 and the current (post-Q3-2006) QNHS weighting schemes for measuring the main characteristics of the immigrant population in Ireland has been examined by Nguyen and Murphy (2015). By comparing the pre-Q3-2006 weighted estimates from the QNHS with the Census 2006 figures and comparing the post-Q3-2006 weighted estimates with the Census 2011, Nguyen and Murphy (2015) come to two conclusions. First, the pre-Q3-2006 weights are not reliable for immigration research. Second, the current weighting scheme performs better than the pre-Q3-2006 scheme with regards to matching the Census figures, but the improvement in performance is minor.

A limitation to the work of Nguyen and Murphy (2015) is its inability to directly compare the efficiency of the pre-Q3-2006 weighting scheme with that of the current scheme. It is not possible to do so in that empirical study because the QNHS data sets do not come with both the pre-Q3-2006 and the post-Q3-2006 weights. Moreover, variables on strata and clusters used in the QNHS design are not available due to data confidentiality rules. Therefore, researchers are unable to calculate their own pre-Q3-2006 and post-Q3-2006 weights using a real QNHS sample. As a result, one can only compare the efficiency of these two weighting schemes using simulation.

In this article, we re-examine the performance of the pre-Q3-2006 and the current weighting scheme of the QNHS on simulated samples as well as extend the work of Nguyen and Murphy (2015) by proposing two other weighting schemes that can serve as the alternatives to the current QNHS weighting methodology. They are referred to as the modified QNHS and the raking-ratio scheme. We compare the effectiveness of the existing and the proposed QNHS weighting mechanisms for immigration research using simulation exercises.

It should be noted that this is the first time the effects of the QNHS weighting schemes have been examined using simulation and also the first time that alternative weighting

schemes have been suggested for Ireland's QNHS. Within Europe, there are studies investigating the overall effectiveness of the LFS weighting schemes in Sweden (Hörngren 1992), Finland (Djerf and Väisänen 1993; Djerf 1997), and Norway (Thomsen and Holmøy 1998), as well as their effectiveness specifically for immigration research in Norway (Villund 2010) and in Spain (Martí and Ródenas 2012). These studies are similar to ours in their objectives; however, differences in survey designs and weighting methodologies of the LFS in these countries lead to differences in the methods used in their studies and ours. In general, countries with extensive registers such as Sweden, Finland, and Norway can have more complex weighting methodologies than those without population registers (i.e Ireland). Subsequently, weighting schemes that are proposed for these register countries may not be suitable for other countries.

In summary, the aim of this article is to use simulation to compare the effectiveness of four different weighting methodologies in improving the precision of the labour-force estimates of Ireland's whole population and its main nationality groups. In Ireland, we group the nationalities into five main groups of Irish, UK, OMS, NMS, and Other Nationals. The four weighting schemes are the pre-Q3-2006, the current QNHS, the modified QNHS and the raking-ratio weighting scheme.

We begin with a brief overview of the theory of calibration and a detailed description of the existing and proposed weighting schemes. This is followed by a description of the simulation procedure, corresponding results, and conclusion.

2. Calibration Techniques

In survey sampling, calibration refers to the process of reweighting samples such that the final weighted samples are consistent with the population with regards to characteristics of interest. In this section, we will start with the general theory of calibration and its notation, then describe in detail the four weighting methods for comparison.

Suppose that we have a population U of size N and an initial sample s of size n_s selected from population U using probability sampling ($s \subset U$, $n_s \leq N$). Let π_k be the probability of selection and d_k be the design weight of the k th individual ($k \in s$) such that $d_k = 1/\pi_k$. In an ideal world without nonresponse and other sampling and nonsampling errors, the design weight would be the final weight. In reality, this is rarely the case for voluntary sample surveys. Suppose that only n_r individuals out of the initial n_s selected participants respond to the survey ($n_r \leq n_s \leq N$). Let r denote the sample of n_r respondents ($r \subset s \subset U$).

The aim of calibration is to find the final weights w_k ($k \in r$) that are "as close as possible" to the design weights d_k such that the resulting weighted samples match known population estimates for a select number of characteristics (Deville and Särndal 1992). These known population estimates, referred to as auxiliary data, are retrieved from external sources such as the Census, population registers, and other administrative sources. It is well known in survey sampling that proper use of auxiliary information at the estimation stage can reduce bias, improve the precision of variables of interest, and impose consistency with results from other sources (Zhang 2000; Särndal and Lundström 2005; Särndal 2007). In the following subsections, we will discuss two specific calibration techniques called poststratification and raking ratio and their application to the QNHS.

2.1. Poststratification

Poststratification is a classical technique used in survey sampling to adjust for nonresponse bias and improve precision of estimates of variables of interest (Thomsen 1973; Thomsen 1978; Holt and Smith 1979; Jagers 1986). Its concept is similar to that of stratification but strata (referred to as poststrata) are formed after the samples are taken, rather than at the design stage.

Poststratification is a type of calibration approach as it calculates calibrated weights under the constraint that the weighted samples match population estimates broken down by post-strata. These poststrata are formed from the cross tabulation of the auxiliary variables. For example, if we want to poststratify a sample by three age groups and sex, we obtain a cross-tabulated table of six cells. These are the six poststrata, and sex and age are the two auxiliary variables. Poststratification requires a known population count for each of these cells. It then constructs calibrated weights to ensure a perfect match between the sample weighted total and the actual population total for all the cells in the tabulated table. Hence, poststratification is commonly referred to as *calibration on known cell counts* (Deville and Särndal 1992; Deville et al. 1993).

The poststrata are H disjoint groups such that $U = \cup_{h=1}^{H} U_h$ and $r = \cup_{h=1}^{H} r_h$. The population size and the sample size of the hth poststratum are N_h and n_{r_h}, respectively. Assume that the population total N_h is known for each poststratum $h = \{1, 2, \ldots, H\}$. In poststratification, the design weight d_k for each $k \in r_h$ is adjusted by a factor of $N_h / \left(\sum_{k \in r_h} d_k \right)$, which is the ratio between the true population count and the estimated population count from the sample. The new calibrated weight has the form $w_k = d_k \left(N_h / \left(\sum_{k \in r_h} d_k \right) \right)$. When these calibrated weights w_k are used, the weighted sample will match the population totals for all poststrata.

Poststratification is straightforward to implement and widely used by National Statistical Institutes (NSIs) around the world including the CSO in Ireland.

2.1.1. The QNHS Pre-Q3-2006 Weighting Scheme

Between 1997 and Q3 2006, the CSO used simple poststratification to construct its weights based on Age, Sex, and Region. Specifically, the QNHS samples were poststratified by 18 age groups (in five year increments from 0 to 85+ years), sex, and eight NUTS3 regions (Border, Dublin, Midland, Mid-East, Mid-West, South-East, South-West, and West). This resulted in the calibration of 288 poststrata, and the weighted samples matched population estimates for all of these poststrata. In Ireland, population estimates are obtained from the latest Census adjusted for migration and vital statistics (Central Statistics Office 2014).

Within the EU, a number of countries such as Belgium, the Czech Republic, Greece, Cyprus, Luxembourg, Poland, Slovenia, Slovakia, Malta, and Germany currently use poststratification in their calculations of weights for the LFS (Eurostat 2014).

2.1.2. The Current QNHS Weighting Scheme

Since Q3 2006, the CSO has constructed weights using two different criteria. The first criterion is exactly that used in the pre-Q3-2006 weighting scheme. In the second criterion, an additional 20 cells are introduced. The QNHS samples are simultaneously poststratified by two age groups (under 15, 15+), sex, and five broad nationality groups (Irish, UK,

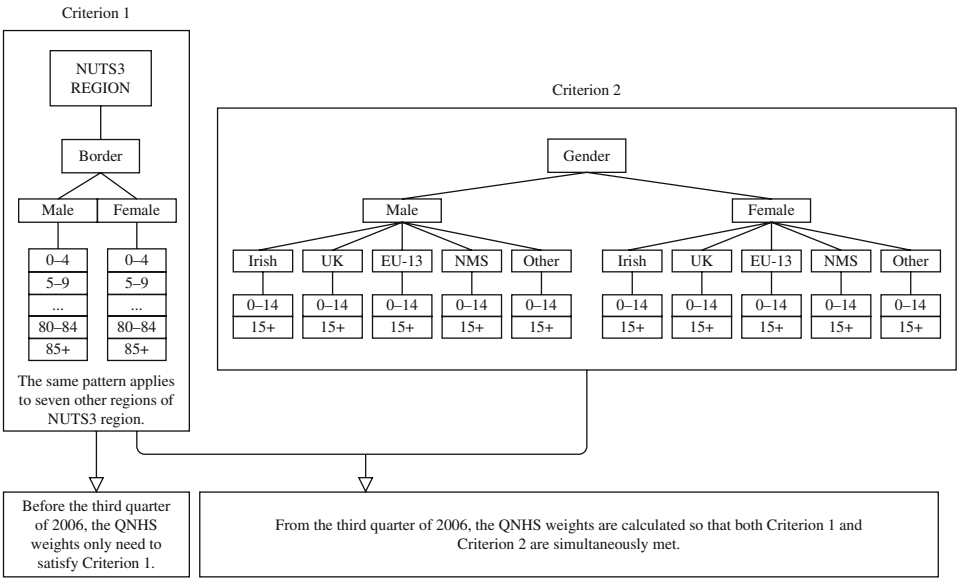


Fig. 1. Diagram of the construction of the QNHS weights (Nguyen and Murphy 2015).

OMS, NMS, and Other). The criteria used in the construction of the pre-Q3-2006 and the current QNHS weights are illustrated in Figure 1. The CALMAR 2 macro in SAS (Sautory 2003) is used to ensure that the current QNHS weights satisfy both criteria simultaneously.

Within the EU, other countries such as Bulgaria, Spain, Italy, Lithuania, the Netherlands, Portugal, Romania, and Macedonia also calibrate their LFS samples using multiple criteria similar to Ireland's current weighting scheme (Eurostat 2014).

2.1.3. The Modified QNHS Weighting Scheme

We now propose a modified version of the current QNHS weighting scheme. This new method involves an adjustment to the second criterion while making no change to the first criterion. The second criterion is extended to match population estimates by four age groups (under 15, 15–24, 25–49, 50+). The sex and nationality groups remain unchanged. The weights must now satisfy both of these criteria, that is simultaneous calibrations of 288 cells and 40 cells. As before, this is implemented using the CALMAR 2 macro in SAS. We now introduce another scheme before examining this.

2.2. Raking Ratio

While poststratification is a popular calibration technique, there are two scenarios in which it cannot be implemented. The first scenario is when a sample poststratum r_h is empty or has an extremely small sample size. The second scenario is when the population count of the poststratum N_h is unknown or not reliable. In these situations, survey statisticians may opt for a technique called raking ratio to calibrate their samples.

Formalised originally by Deming and Stephan (1940), raking ratio is a classical method of calculating survey weights when the marginal population count for each auxiliary variable is known, but not the detailed population count for each cell in the cross-tabulated

table formed by these auxiliary variables. For example, suppose we want to poststratify a sample by three age groups and sex. Assume that we do not know the population counts for all of these six cells; poststratification is therefore not possible. Suppose that from the latest Census, we know the marginal population totals (i.e the number of males and females in the population, the number of people in each of the three age brackets in the population). In this case, we can use the raking ratio method, a reliable alternative technique to poststratification, to calculate the survey weights (Deville et al. 1993). Hence, raking ratio can be referred to as *incomplete post-stratification* or *calibration on known marginal counts* (Deville and Särndal 1992; Deville et al. 1993).

Suppose that we want to calibrate a sample using two auxiliary variables with I and J number of levels, resulting in a cross-tabulated table of $I \times J$ cells. Let N_{i+} (for $i = \{1, 2, \ldots I\}$) denote the marginal population count for the ith row, and let N_{+j} (for $j = \{1, 2, \ldots J\}$) denote the marginal population count for the jth column of the cross-tabulated table. Assume that N_{i+} and N_{+j} are known. Raking ratio uses iterative steps to obtain the calibrated weights such that the final weighted marginal counts from the sample for all I rows and J columns match their corresponding marginal population counts. This procedure can be easily extended to more than two auxiliary variables (Kalton 1983).

2.2.1. Raking Ratio for the QNHS

The CSO uses poststratification to calculate the pre-Q3-2006 and the current QNHS weights. However, poststratification cannot be implemented in two scenarios: first when the poststrata are empty and second when the population counts of the poststrata are unknown or unreliable.

The first scenario can happen, but is most likely not a problem for the QNHS due to their large quarterly sample sizes of approximately 45,000 to 60,000 individuals. In our simulation study, we estimate that empty poststrata occur about one per cent of the time.

The second scenario in which poststratification is not recommended is when the population counts of the poststrata are unknown or not reliable. This was and still is potentially an issue in Ireland, where estimates of population counts are obtained from the latest Census adjusted for migration and vital statistics (Central Statistics Office 2014). The migration statistics come principally from the QNHS. It means that if the QNHS does not capture the migration flow reliably, the migration statistics are not reliable, which subsequently affects the intercensal population estimates. When the Census 2011 figures were released, they revealed that the annual migration statistics between 2006 and 2011 had been underestimated by 75 per cent or 87,000 people (Houses of the Oireachtas 2012). The CSO has since incorporated various administrative data sources to improve its measure of migration statistics, hence, intercensal population estimates. It is, however, not the aim of this article to examine the reliability of Ireland's intercensal population estimates.

When the above scenarios occur, we propose using raking ratio to calculate the QNHS weights. Specifically, raking ratio can be performed using the marginal population counts for 33 margins: 18 age groups (in five-year increments from 0 to 85+ years), two sex groups, eight NUTS3 regions, and five nationality groups (Irish, UK, OMS, NMS, and Other Nationals). We choose Age, Sex, Region, and Nationality for this weighting method because these four variables are used in the current and the proposed modified QNHS weighting schemes, thus allowing comparability.

It is noted that raking ratio also depends on reliable marginal population counts, so it faces the same issue discussed in the second scenario. However, potentially unreliable intercensal population estimates have a lesser effect on raking ratio than on poststratification because the former does not require detailed cell counts.

Within the EU, the raking-ratio method is used by Austria and Hungary for their LFS weighting methodologies (Eurostat 2014).

2.3. *Comparison of Weighting Methodologies for the QNHS*

Using the CALMAR 2 macro in SAS, we compute the calibrated weights for each of the following weighting schemes and compare the results. The four schemes are:

1. Pre-Q3-2006 QNHS weighting scheme: complete poststratification by Region (eight NUTS3 regions), Sex, and Age (18 age groups).
2. Current QNHS weighting scheme: simultaneous calibrations to allow poststratification by Region (eight NUTS3 regions), Sex, and Age (18 age groups), as well as poststratification by Sex, Age (under 15, 15+), and Nationality groups (Irish, UK, OMS, NMS, Other).
3. Modified QNHS weighting scheme: simultaneous calibrations to allow poststratification by Region (eight NUTS3 regions), Sex, and Age (18 age groups), as well as poststratification by Sex, Age (under 15, 15–24, 25–49, 50+), and Nationality groups (Irish, UK, OMS, NMS, Other).
4. Raking ratio: calibration on known marginal counts of Region (eight NUTS3 regions), Sex, Age (18 age groups), and Nationality groups (Irish, UK, OMS, NMS, Other).

We measure the performance of each method by calculating the total Mean-Squared Error (MSE) and the total Coefficient of Variation (CV) for all categories of the Principal Economic Status (PES). Initially, we also consider bias as a measure of performance. However, our simulation results show that there is no significant difference in bias across the four weighting schemes. It follows that the weighting scheme with the smallest total MSE and the smallest total CV is considered to be the best method.

It should be pointed out that the QNHS is a household survey, which means that households, not individuals, are the final sampling units. However, the pre-Q3-2006 and the current QNHS weighting schemes involve direct adjustment at individual level instead of household level. To be consistent with the existing QNHS schemes, our two proposed weighting methodologies also perform weight adjustment at individual level. This is a common practice among NSIs conducting the EU-LFS. There are only a few countries, such as Spain, Italy, Hungary, and Lithuania, that adjust the EU-LFS weights at both individual and household levels (Eurostat 2014).

3. Simulation Procedure and Measures of Performance

3.1. *Simulation Procedure*

The primary purpose of constructing calibrated weights is to attempt to account for nonresponse bias and other sampling and nonsampling errors. Therefore, we generate samples with nonresponse to evaluate the performance of the four weighting schemes.

First, 900 samples each of approximately 25,000 observations are drawn from an anonymised subset (ten per cent) of the 2011 Irish Census (Minnesota Population Center 2014). These samples are selected using the same two-stage stratified cluster sample design as the QNHS (Central Statistics Office 2011). In the first stage, Primary Sampling Units (PSUs), each containing approximately 75 households, are selected using Probability Proportional to Size Sampling. In the second stage, 15 households are selected from each PSU using Systematic Sampling. All individuals in the selected households are included in the samples.

Next, we generate nonresponse for each sample. Since the QNHS is a household survey, nonresponse is generated at the household level instead of the individual level. We consider the following six nonresponse (NR) scenarios:

- NR1: We randomly remove 20% of households from the samples. This is consistent with the general nonresponse level of the QNHS.
- NR2: We generate nonresponse based on NUTS3 regions as reported for the QNHS 2013 (Eurostat 2013). The nonresponse rates for the eight NUTS3 regions are: Border (24.10%), Midland (16.64%), West (27.30%), Dublin (26.54%), Mid-East (22.70%), Mid-West (23.22%), South-East (18.45%), and South-West (19.20%).
- NR3: Nonresponse is generated for the two NUTS2 regions reported for the QNHS 2013 (Eurostat 2013). The nonresponse rates for the Border-Mid-West region and for the South-East region are 23.67% and 22.65%, respectively.
- NR4: Nonresponse rates are generated for different household types. There are four types of households: Cohabiting partners without children, Cohabiting partners with children, Lone parents with children, and Other. Their nonresponse rates are estimated using the QNHS 2011 (Q2) and the Irish Census 2011 samples. The estimated nonresponse rates for these four types of households are 16.37%, 15.14%, 23.18%, and 17.53%, respectively.
- NR5: Nonresponse rates depend on urbanicity estimated from the EU-SILC 2011 and the Irish Census 2011 samples. The nonresponse rate for urban areas is 25%, and that for the rural areas is 13%. This is consistent with literature that shows that rural areas are more likely to participate in surveys than urban areas (United Nations 2005; King et al. 2009; Pérez-Duarte et al. 2010).
- NR6: Nonresponse rates vary for Irish households and immigrant households. We categorise a household as an immigrant household if two thirds or more than two thirds of its members are foreign nationals. We then estimate the nonresponse rates for Irish households and immigrant households using the QNHS 2011 (Q2) and the Census 2011. They are 17% and 39%, respectively.

In each of the six nonresponse scenarios, we obtain 900 final samples. For each of the 900 samples, we compute calibrated weights using the four weighting schemes described in Subsection 2.3. We then obtain the overall PES distribution and that for each of the five nationality groups (Irish, UK, OMS, NMS, and Other). In the following subsection, we describe the two measures of performance used to determine the best weighting scheme for the QNHS.

3.2. Measures of Performance

The PES indicates the status of each individual in the labour force. It has three categories: Employed, Unemployed, and Inactive. Suppose that their corresponding population percentages are p_1, p_2, and p_3. Let $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_3$ be the weighted sample estimates (in percentage) of those employed, unemployed, and inactive, respectively. Let the estimated mean over the Monte Carlo simulations for each PES category be:

$$\bar{\hat{p}}_i = \frac{1}{900}\sum_{k=1}^{900} \hat{p}_{ik} \quad \text{for} \quad i = 1, 2, 3$$

and the estimated sampling variance be:

$$\hat{V}(\hat{p}_i) = \frac{1}{899}\sum_{k=1}^{900} (\hat{p}_i - \bar{\hat{p}}_i)^2 \quad \text{for} \quad i = 1, 2, 3$$

In our study, we use the MSE and the CV as measures of performance. The MSE measures the accuracy of an estimator and is equal to the average squared distance between each sample estimate and the corresponding true population percentage. On the other hand, the CV measures the relative variability of an estimate and is equal to the ratio of the standard error of the estimate and the estimate itself. We estimate the MSE and the CV using the following formulae, with index i indicating the category of PES and k indicating the simulation index.

1. Estimated Mean-Squared Error (MSE)

$$\hat{\text{MSE}}(\hat{p}_i) = \frac{1}{900}\sum_{k=1}^{900} (\hat{p}_{ik} - p_i)^2 \tag{1}$$

$$\hat{\text{MSE}}(\text{PES}) = \sum_{i=1}^{3} \hat{\text{MSE}}(\hat{p}_i) = \sum_{i=1}^{3} \left[\frac{1}{900}\sum_{k=1}^{900} (\hat{p}_{ik} - p_i)^2\right] \tag{2}$$

2. Estimated Coefficient of Variation (CV)

$$\widehat{\text{CV}}(\hat{p}_i) = \frac{\sqrt{\hat{V}(\hat{p}_i)}}{\bar{\hat{p}}_i} \times 100\% \tag{3}$$

$$\widehat{\text{CV}}(\text{PES}) = \sum_{i=1}^{3} \widehat{\text{CV}}(\hat{p}_i) = \sum_{i=1}^{3} \left[\frac{\sqrt{\hat{V}(\hat{p}_i)}}{\bar{\hat{p}}_i}\right] \times 100\% \tag{4}$$

We consider the best weighting scheme to be the one with the smallest $\hat{\text{MSE}}(\text{PES})$ (2) and the smallest $\widehat{\text{CV}}(\text{PES})$ (4).

3.3. MSE and CV Estimation in NSIs

In this article, we use Monte Carlo simulations to estimate the MSE and the CV, which are functions of the sampling variance. In reality, NSIs around Europe estimate the sampling

variance not only based on Monte Carlo simulation, but also based on analytic or replication methods.

Variance estimation in a complex sample survey is a challenging task. It depends on the type of sampling design, the type of estimator, the type of nonresponse corrections, and the form of statistics (Eurostat 2002). With the QNHS, it is almost impossible to use exact analytic methods to calculate the sampling variance. This is due to its complex two-stage stratified cluster sample design and its complex weighting scheme. Moreover, our interest in the estimation of the PES distribution for subpopulations (i.e five nationality groups) makes the exact calculation of the sampling variance and hence the MSE and the CV even more unfeasible.

Within the EU, some common variance estimation methods employed by countries for their LFS are the Taylor linearisation, jackknife, bootstrap, balanced repeated replication, and random-groups method. Apart from the Taylor linearisation method, these are replication methods which require intensive computer power. Of these, the jackknife method for variance estimation is recommended by Eurostat's Task Force to all countries except Luxembourg (Eurostat 2002). Currently, the Irish CSO also uses the jackknife method for the QNHS (Central Statistics Office 2015b). If our proposed weighting schemes were to be adopted for the QNHS, we would suggest using the jackknife method to estimate the sampling variance and hence the MSE and the CV.

4. Results

As mentioned previously, we use the MSE and the CV as measures of performance in this article. The weighting method with the smallest $\hat{\text{MSE}}$ (PES) (2) and the smallest $\widehat{\text{CV}}$ (PES) (4) is considered the best weighting scheme for the QNHS. We will start this section by discussing the MSE, followed by the CV results.

The MSE is made up of two components, bias and sampling variance, and there is usually a trade-off between these components. In official statistics, interest often lies on obtaining point estimates of the population and subpopulations, so having a small bias is desirable. However, our simulation indicates that there is no significant difference in bias across the four methods, neither for the whole population nor any nationality group (results not shown). It is the difference in the sampling variance that contributes to the difference in the MSE across the four weighting schemes. The MSE results are presented in Table 1 to Table 6.

Table 1. $\hat{MSE}$ (PES) for the whole population.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	0.30	0.30	0.30	0.30
NR2	0.31	0.31	0.31	0.31
NR3	0.31	0.31	0.31	0.31
NR4	0.33	0.32	0.32	0.33
NR5	0.31	0.31	0.31	0.31
NR6	0.30	0.30	0.30	0.29

(Apply to all tables) Within each row, the figure(s) shaded in gray is (are) the smallest. It indicates the best weighting scheme in each nonresponse scenario.

Table 2. $\hat{MSE}$ (PES) for the Irish nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	0.36	0.36	0.35	0.36
NR2	0.38	0.38	0.37	0.38
NR3	0.37	0.37	0.37	0.37
NR4	0.43	0.41	0.41	0.40
NR5	0.38	0.37	0.35	0.38
NR6	0.49	0.36	0.34	0.36

Table 3. $\hat{MSE}$ (PES) for the UK nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	10.97	10.91	10.01	10.79
NR2	11.77	11.69	10.43	11.63
NR3	11.95	12.00	10.97	11.76
NR4	11.70	11.66	10.62	11.59
NR5	11.24	11.23	10.12	11.01
NR6	14.20	13.25	11.51	13.41

Table 4. $\hat{MSE}$ (PES) for the OMS nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	23.50	22.89	18.70	23.32
NR2	24.11	23.59	19.21	23.96
NR3	23.61	23.14	18.94	23.33
NR4	24.61	23.90	20.18	24.47
NR5	24.63	24.21	19.07	24.59
NR6	27.79	27.38	22.35	27.88

Table 5. $\hat{MSE}$ (PES) for the NMS nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	6.46	6.42	6.28	6.41
NR2	6.85	6.77	6.62	6.78
NR3	7.19	7.12	6.94	7.15
NR4	6.76	6.70	6.61	6.70
NR5	7.20	7.13	6.95	7.16
NR6	8.78	8.61	8.48	8.65

There are a number of things to note in Tables 1–6. First of all, the proposed modified QNHS weighting scheme produces the smallest $\hat{MSE}$ (PES) in 34 out of 36 scenarios presented (six nonresponse scenarios for six groups – the whole population and five nationality groups). In the remaining two scenarios (NR6 for the whole population and

Table 6. $\hat{MSE}$ (PES) for other nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	8.31	8.21	7.15	8.24
NR2	8.76	8.64	7.39	8.72
NR3	8.55	8.41	7.11	8.47
NR4	8.66	8.59	7.32	8.60
NR5	8.68	8.56	7.34	8.60
NR6	10.52	10.24	8.90	10.18

NR4 for the Irish nationals), the difference between the $\hat{MSE}$ (PES) produced by the modified QNHS weighting scheme and that of the best method in that case is not material. This result is very encouraging because by making a small change to the current QNHS weighting scheme, the modified QNHS scheme repeatedly gives the most accurate estimates.

When we examine Tables 1–6 closely, we do not perceive a material difference in the $\hat{MSE}$ (PES) among the four weighting schemes for the whole population in Table 1. In Table 2, even though the modified QNHS method produces the smallest MSE in five out of six nonresponse scenarios, the difference among the MSE figures across the four weighting mechanisms is quite small. This is not surprising since the Irish nationals make up the majority of the population, and thus their behaviour should mimic that of the population. On the other hand, the modified QNHS weighting method consistently produces a large reduction in the MSE for the four immigrant groups – UK, OMS, NMS, and Other Nationals.

Additionally, Tables 1–6 show that the current QNHS weighting method does indeed improve the accuracy of the pre-Q3-2006 scheme. This is expected because the current QNHS weighting method takes the nationality of the respondents into account, while the pre-Q3-2006 scheme does not (Nguyen and Murphy 2015). For the same reason, the raking-ratio method also performs better than the pre-Q3-2006 weighting scheme, since the former also calibrates samples on nationality. When compared with the performance of the current QNHS weighting scheme, the raking-ratio method performs relatively similarly.

A similar pattern is observed with the CV results. The $\widehat{CV}$ (PES) for the whole population and the five nationality groups can be seen in Tables 7–12. The tables show that the modified QNHS weighting scheme produces the smallest $\widehat{CV}$ (PES) across the board except for the NR6 scenario of the whole population. Overall, the CV findings agree with the MSE results that the modified QNHS weighting scheme is the best out of the four considered weighting mechanisms.

5. Discussion and Conclusions

Our simulation results have shown that the modified QNHS weighting scheme gives the best results out of the four weighting methodologies, as demonstrated by its consistently smallest MSE and CV. We also notice that the current QNHS scheme performs better than the pre-Q3-2006 one. However, as the pre-Q3-2006, the current, and the modified QNHS

Table 7. $\widehat{CV}$ *(PES) for the whole population (%).*

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	3.76	3.77	3.75	3.76
NR2	3.81	3.81	3.80	3.80
NR3	3.88	3.88	3.87	3.88
NR4	3.73	3.73	3.72	3.73
NR5	3.70	3.70	3.70	3.71
NR6	3.67	3.70	3.69	3.71

weighting schemes all use the poststratification technique, they cannot be implemented when samples contain empty poststrata or when the population counts for poststrata are unknown or unreliable. When these scenarios occur, we suggest using the raking-ratio method as an alternative weighting scheme. As we discussed in Section 4, the raking-ratio method performs better than the pre-Q3-2006 weighting scheme and similarly to the current one.

While we consider the best weighting method to be the one with the smallest $\widehat{\text{MSE}}$ (PES) (2) and the smallest $\widehat{\text{CV}}$ (PES) (4), we also provide the estimated MSE (1) and the estimated CV (3) for each of the three categories of the PES (i.e Employed, Unemployed, and Inactive) in the Appendix A (Tables A.1–A.12). Interestingly, while the modified QNHS weighting scheme outperforms other methods in most scenarios, the raking-ratio method performs better or just as well as the modified QNHS scheme for the Unemployed category of the four immigrant groups (Tables A.5–A.12).

Table 8. $\widehat{CV}$ *(PES) for the Irish nationals (%).*

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	4.22	4.22	4.20	4.20
NR2	4.27	4.28	4.24	4.26
NR3	4.30	4.30	4.26	4.29
NR4	4.20	4.20	4.18	4.21
NR5	4.17	4.15	4.11	4.17
NR6	4.12	4.11	4.07	4.12

Table 9. $\widehat{CV}$ *(PES) for the UK nationals (%).*

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	20.68	20.64	20.11	20.56
NR2	21.37	21.34	20.63	21.28
NR3	21.55	21.56	21.00	21.42
NR4	21.24	21.16	20.63	21.15
NR5	20.95	20.94	20.32	20.76
NR6	22.13	21.95	21.26	21.98

Table 10. $\widehat{CV}$ *(PES) for the OMS nationals (%).*

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	37.28	36.95	35.10	37.16
NR2	37.99	37.78	35.89	37.93
NR3	37.45	37.19	35.41	37.33
NR4	37.92	37.55	36.07	37.79
NR5	38.11	37.83	35.74	37.96
NR6	40.70	40.51	38.63	40.67

While the simulation has shown strong performances and encouraging results, it should be noted that the information on the PSU to which each person or household belongs is not available to us. Therefore, in simulating the 900 QNHS samples (Subsection 3.1), we have to generate artificial PSUs. Because of the artificial PSUs, the clustering effect in our samples is not the same as the real clustering effect.

In reality, it is well known that immigrants usually cluster together in some geographical areas (Robinson 2006; O'Boyle 2009). This means that the proportion of immigrants in some real PSUs would be higher than that in our artificial PSUs. This is because in this study we randomly allocate households among the artificial PSUs, so each artificial PSU would contain approximately the same amount of immigrants.

To understand the effect of artificial PSUs on the robustness of our proposed weighting methods in the estimation of the immigrant population, we have simulated another set of artificial PSUs under an extreme scenario. Instead of being randomly allocated to PSUs as done previously, households are now allocated to either "immigrant" PSUs or Irish PSUs based on their status. A household is classified as an "immigrant" household if two thirds or more than two thirds of their members are foreign nationals. Otherwise, it is classified as an Irish household. All "immigrant" households are randomly allocated to "immigrant" PSUs with each PSU containing approximately 75 households. Similarly, all Irish households are assigned to Irish PSUs, each of 75 households as well. This set-up represents the extreme scenario in which all PSUs are homogeneous with regards to nationality (Irish or non-Irish). When every household in the Census sample is allocated to one PSU, another 900 samples are drawn with the same procedure as described in Subsection 3.1. Of the six nonresponse scenarios considered previously, we pick the sixth nonresponse scenario (NR6) to demonstrate the results, because it is directly linked to

Table 11. $\widehat{CV}$ *(PES) for the NMS nationals (%).*

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	18.42	18.31	18.14	18.32
NR2	19.15	19.00	18.77	19.06
NR3	19.44	19.28	19.06	19.37
NR4	18.96	18.86	18.61	18.86
NR5	19.36	19.23	19.00	19.26
NR6	21.41	21.24	20.99	21.36

Table 12. $\widehat{CV}$ *(PES) for other nationals (%).*

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
NR1	17.18	17.10	16.37	17.13
NR2	17.51	17.43	16.63	17.45
NR3	17.40	17.28	16.40	17.31
NR4	17.18	17.09	16.35	17.15
NR5	17.61	17.52	16.71	17.53
NR6	18.74	18.59	17.78	18.58

immigrants' nonresponse propensity. The $\widehat{MSE}$ (PES) and the $\widehat{CV}$ (PES) for the NR6 scenario under this new "extreme" PSUs allocation can be seen in Table 13 and Table 14. The estimated MSE and CV for each category of PES in this case are provided in the Appendix B (Tables B.1–B.2).

From Tables 13–14, we see that our modified QNHS weighting scheme also performs the best out of the four weighting methods for all five nationality groups (Irish, UK, OMS, NMS, and Other Nationals) in terms of both MSE and CV. With regards to the distribution of PES for the whole population, all four weighting methods perform equally well on the MSE criterion, but the pre-Q3-2006 weighting scheme produces the smallest CV. However, the difference between the estimated CV under the pre-Q3-2006 scheme and the modified one is minor. The results show the robustness of our proposed modified QNHS weighting scheme to the clustering effect of immigrants.

In conclusion, our study has demonstrated that the proposed modified QNHS weighing scheme is the best weighting method for obtaining the labour-force estimates of the main foreign-national groups while not affecting the estimates on the population and the Irish nationals. Considering the fact that foreign nationals make up a significant portion of Ireland's population and the growing interest in understanding their characteristics, we recommend using our proposed modified QNHS weighting scheme in place of the current scheme for more reliable estimates on Ireland's labour force. In the event that poststratification is not possible as previously discussed, we recommend using the raking-ratio method, whose performance is similar to that of the current QNHS scheme, as an alternative weighting scheme.

Although our data are entirely Irish, this study highlights potential issues that other countries may face when using the EU-LFS for immigration research. In recent years,

Table 13. $\widehat{MSE}$ *(PES) for NR6 with "extreme" PSUs.*

Nationality group	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
Population	0.28	0.28	0.28	0.28
Irish	0.50	0.32	0.30	0.31
UK	14.98	13.97	11.92	14.21
OMS	27.83	27.55	22.15	27.80
NMS	8.94	9.05	8.79	9.03
Other nationals	10.91	10.78	9.23	10.70

Table 14. $\widehat{CV}$ *(PES) for NR6 with "extreme" PSUs (%).*

Nationality group	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
Population	3.71	3.74	3.74	3.74
Irish	4.14	4.05	4.00	4.03
UK	22.66	22.49	21.74	22.53
OMS	40.82	40.74	38.49	40.77
NMS	21.42	21.48	21.02	21.43
Other nationals	19.02	18.92	17.92	18.92

migration has become a global phenomenon with Europe at its centre. A number of European countries have seen an influx of immigrants from other European and non-European states. This is causing a shift in their population demographics that is similar to Ireland's following EU enlargement. As such, there is growing interest in understanding the characteristics of immigrants and their labour-market participation. With its high frequency, large sample sizes, and a certain level of harmonisation among EU countries, the LFS is a popular data source for immigration research. Even though the traditional objective of the EU-LFS is to produce official statistics on the labour force for the whole population, we believe that it is important for the EU-LFS to also produce reliable statistics for the immigrant population.

Other than for Ireland, we have not examined in detail the effectiveness of the EU-LFS weighting schemes for immigration research in other countries. However, an overview of the individual weighting schemes used in the EU-LFS raises some concerns to us. For example, countries with a large number of immigrants such as the UK and Italy, each with a foreign national population of approximately five million (Eurostat 2015), do not have Nationality included in their EU-LFS weighting schemes (Eurostat 2014). Other smaller countries such as Cyprus and Latvia, which rank second and third respectively among the 28 EU countries for the highest proportion of non-nationals in the population (Eurostat 2015), also do not use Nationality as a calibration variable (Eurostat 2014). Our study demonstrates that by making changes to the current LFS weighting schemes, we can achieve more reliable labour-force statistics not only for the whole population, but also for the immigrant one. Therefore, we recommend that other NSIs revisit their EU-LFS weighting schemes for immigration research.

A. Appendix

A.1. Whole Population

Table A.1. MSE for the whole population.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	0.13	0.13	0.13	0.13
NR2	0.14	0.14	0.14	0.14
NR3	0.14	0.14	0.14	0.14
NR4	0.15	0.15	0.15	0.15
NR5	0.14	0.14	0.14	0.14
NR6	0.13	0.13	0.13	0.13
State 2: Unemployed				
NR1	0.07	0.07	0.07	0.07
NR2	0.07	0.07	0.07	0.07
NR3	0.07	0.07	0.07	0.07
NR4	0.09	0.08	0.08	0.09
NR5	0.07	0.07	0.07	0.07
NR6	0.07	0.07	0.07	0.07
State 3: Inactive				
NR1	0.10	0.10	0.10	0.10
NR2	0.10	0.10	0.10	0.10
NR3	0.10	0.10	0.10	0.10
NR4	0.09	0.09	0.09	0.09
NR5	0.10	0.10	0.10	0.10
NR6	0.10	0.10	0.10	0.09

(Apply to all tables) Within each row, the figure(s) shaded in gray is (are) the smallest. It indicates the best weighting scheme in each nonresponse scenario.

Table A.2. CV for the whole population (%).

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	0.73	0.73	0.73	0.73
NR2	0.75	0.75	0.75	0.75
NR3	0.74	0.74	0.74	0.74
NR4	0.72	0.72	0.72	0.72
NR5	0.72	0.72	0.72	0.72
NR6	0.72	0.73	0.72	0.72
State 2: Unemployed				
NR1	2.21	2.22	2.20	2.21
NR2	2.21	2.21	2.20	2.21
NR3	2.31	2.31	2.30	2.32
NR4	2.20	2.20	2.19	2.21
NR5	2.16	2.16	2.16	2.18
NR6	2.13	2.16	2.15	2.18
State 3: Inactive				
NR1	0.82	0.82	0.82	0.82
NR2	0.85	0.85	0.85	0.84
NR3	0.83	0.83	0.83	0.82
NR4	0.81	0.81	0.81	0.80
NR5	0.82	0.82	0.82	0.81
NR6	0.82	0.81	0.82	0.81

A.2. Irish Nationals

Table A.3. MSE for the Irish nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	0.16	0.16	0.15	0.16
NR2	0.17	0.17	0.16	0.17
NR3	0.16	0.16	0.16	0.16
NR4	0.20	0.18	0.19	0.18
NR5	0.17	0.17	0.16	0.17
NR6	0.20	0.16	0.15	0.16
State 2: Unemployed				
NR1	0.08	0.08	0.08	0.08
NR2	0.08	0.08	0.08	0.08
NR3	0.08	0.08	0.08	0.08
NR4	0.10	0.10	0.10	0.10
NR5	0.07	0.07	0.07	0.08
NR6	0.08	0.07	0.07	0.07
State 3: Inactive				
NR1	0.12	0.12	0.12	0.12
NR2	0.13	0.13	0.13	0.13
NR3	0.13	0.13	0.13	0.13
NR4	0.13	0.13	0.12	0.12
NR5	0.14	0.13	0.12	0.13
NR6	0.21	0.13	0.12	0.13

Table A.4. CV for the Irish nationals (%).

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	0.80	0.80	0.79	0.80
NR2	0.83	0.83	0.81	0.83
NR3	0.82	0.82	0.80	0.81
NR4	0.80	0.80	0.79	0.80
NR5	0.81	0.81	0.79	0.81
NR6	0.80	0.80	0.78	0.80
State 2: Unemployed				
NR1	2.53	2.53	2.53	2.53
NR2	2.53	2.53	2.53	2.52
NR3	2.57	2.57	2.57	2.58
NR4	2.51	2.51	2.51	2.52
NR5	2.45	2.44	2.44	2.47
NR6	2.43	2.42	2.42	2.43
State 3: Inactive				
NR1	0.89	0.89	0.88	0.88
NR2	0.92	0.92	0.90	0.91
NR3	0.91	0.91	0.89	0.90
NR4	0.89	0.89	0.88	0.89
NR5	0.91	0.90	0.88	0.90
NR6	0.89	0.89	0.87	0.89

A.3. UK Nationals

Table A.5. MSE for the UK nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	4.52	4.49	4.13	4.45
NR2	4.93	4.87	4.32	4.86
NR3	4.93	4.96	4.52	4.87
NR4	4.86	4.82	4.38	4.81
NR5	4.58	4.56	4.07	4.48
NR6	6.09	5.62	4.77	5.69
State 2: Unemployed				
NR1	2.12	2.13	2.12	2.11
NR2	2.25	2.27	2.26	2.25
NR3	2.30	2.29	2.29	2.28
NR4	2.27	2.27	2.27	2.27
NR5	2.20	2.21	2.20	2.18
NR6	2.31	2.30	2.29	2.29
State 3: Inactive				
NR1	4.33	4.29	3.76	4.23
NR2	4.59	4.55	3.85	4.52
NR3	4.72	4.75	4.16	4.61
NR4	4.57	4.57	3.97	4.51
NR5	4.46	4.46	3.85	4.35
NR6	5.80	5.33	4.45	5.43

Table A.6. CV for the UK nationals (%).

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	4.49	4.48	4.29	4.46
NR2	4.69	4.66	4.39	4.66
NR3	4.69	4.71	4.49	4.66
NR4	4.64	4.62	4.41	4.62
NR5	4.52	4.51	4.26	4.47
NR6	4.78	4.74	4.55	4.76
State 2: Unemployed				
NR1	10.91	10.92	10.89	10.87
NR2	11.24	11.28	11.25	11.23
NR3	11.36	11.34	11.33	11.32
NR4	11.21	11.18	11.20	11.18
NR5	11.11	11.13	11.10	11.05
NR6	11.41	11.35	11.31	11.33
State 3: Inactive				
NR1	5.28	5.24	4.93	5.22
NR2	5.44	5.39	4.99	5.39
NR3	5.50	5.51	5.18	5.44
NR4	5.39	5.36	5.02	5.35
NR5	5.32	5.30	4.96	5.24
NR6	5.94	5.86	5.40	5.90

A.4. OMS Nationals

Table A.7. MSE for the OMS nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	10.43	10.15	8.16	10.33
NR2	10.77	10.53	8.51	10.67
NR3	10.37	10.18	8.21	10.23
NR4	10.87	10.56	8.78	10.78
NR5	10.91	10.74	8.27	10.93
NR6	12.54	12.40	10.01	12.61
State 2: Unemployed				
NR1	3.14	3.14	3.16	3.12
NR2	3.33	3.33	3.37	3.32
NR3	3.20	3.21	3.25	3.20
NR4	3.23	3.21	3.27	3.20
NR5	3.30	3.30	3.34	3.26
NR6	3.84	3.87	3.95	3.83
State 3: Inactive				
NR1	9.93	9.60	7.38	9.87
NR2	10.01	9.73	7.33	9.97
NR3	10.04	9.75	7.48	9.90
NR4	10.51	10.13	8.13	10.49
NR5	10.42	10.17	7.46	10.40
NR6	11.41	11.11	8.39	11.44

Table A.8. CV for the OMS nationals (%).

Scenario	*Pre-Q3-2006*	*Current QNHS weights*	*Modified QNHS weights*	*Raking ratio*
State 1: Employed				
NR1	5.06	4.99	4.47	5.03
NR2	5.14	5.09	4.57	5.12
NR3	5.05	5.00	4.49	5.01
NR4	5.17	5.09	4.64	5.14
NR5	5.18	5.13	4.50	5.18
NR6	5.57	5.52	4.96	5.57
State 2: Unemployed				
NR1	20.81	20.73	20.80	20.75
NR2	21.41	21.41	21.53	21.39
NR3	20.93	20.88	21.04	20.92
NR4	21.04	20.94	21.19	20.96
NR5	21.36	21.26	21.43	21.23
NR6	22.99	22.95	23.23	22.95
State 3: Inactive				
NR1	11.41	11.23	9.83	11.38
NR2	11.44	11.28	9.79	11.42
NR3	11.47	11.31	9.88	11.39
NR4	11.71	11.51	10.24	11.70
NR5	11.57	11.44	9.81	11.55
NR6	12.14	12.04	10.44	12.15

A.5. NMS Nationals

Table A.9. MSE for the NMS nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	3.01	3.01	2.92	2.99
NR2	3.12	3.09	3.01	3.09
NR3	3.31	3.29	3.18	3.30
NR4	3.07	3.05	3.01	3.04
NR5	3.37	3.34	3.24	3.36
NR6	4.05	3.97	3.91	3.97
State 2: Unemployed				
NR1	2.08	2.07	2.06	2.07
NR2	2.20	2.20	2.19	2.18
NR3	2.40	2.40	2.38	2.38
NR4	2.23	2.23	2.22	2.22
NR5	2.32	2.32	2.31	2.31
NR6	2.90	2.89	2.89	2.87
State 3: Inactive				
NR1	1.37	1.34	1.30	1.35
NR2	1.53	1.48	1.42	1.51
NR3	1.48	1.43	1.38	1.47
NR4	1.46	1.42	1.38	1.44
NR5	1.51	1.47	1.40	1.49
NR6	1.83	1.75	1.68	1.81

Table A.10. CV for the NMS nationals (%).

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	2.59	2.59	2.56	2.58
NR2	2.63	2.62	2.59	2.62
NR3	2.71	2.70	2.67	2.71
NR4	2.64	2.63	2.61	2.62
NR5	2.70	2.70	2.67	2.70
NR6	3.02	3.00	2.98	3.00
State 2: Unemployed				
NR1	7.33	7.32	7.29	7.31
NR2	7.54	7.55	7.52	7.51
NR3	7.88	7.88	7.85	7.84
NR4	7.59	7.59	7.58	7.57
NR5	7.75	7.75	7.73	7.73
NR6	8.63	8.61	8.61	8.58
State 3: Inactive				
NR1	8.50	8.40	8.29	8.43
NR2	8.98	8.83	8.66	8.93
NR3	8.85	8.70	8.54	8.82
NR4	8.73	8.64	8.42	8.67
NR5	8.91	8.78	8.60	8.83
NR6	9.76	9.63	9.40	9.78

A.6. Other Nationals

Table A.11. MSE for the other nationals.

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	3.25	3.21	2.74	3.22
NR2	3.61	3.55	2.94	3.59
NR3	3.37	3.30	2.69	3.35
NR4	3.51	3.47	2.93	3.49
NR5	3.56	3.52	2.89	3.53
NR6	4.16	4.07	3.55	4.08
State 2: Unemployed				
NR1	1.85	1.86	1.86	1.85
NR2	1.86	1.88	1.90	1.85
NR3	1.87	1.88	1.89	1.85
NR4	1.86	1.88	1.87	1.85
NR5	1.87	1.87	1.91	1.87
NR6	2.27	2.26	2.24	2.22
State 3: Inactive				
NR1	3.21	3.14	2.55	3.17
NR2	3.29	3.21	2.55	3.28
NR3	3.31	3.23	2.53	3.27
NR4	3.29	3.24	2.51	3.26
NR5	3.25	3.17	2.54	3.20
NR6	4.09	3.91	3.11	3.88

Table A.12. CV for the other nationals (%).

Scenario	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1: Employed				
NR1	3.84	3.81	3.53	3.82
NR2	4.05	4.01	3.66	4.04
NR3	3.91	3.87	3.49	3.90
NR4	3.99	3.96	3.65	3.98
NR5	4.03	4.00	3.62	4.00
NR6	4.36	4.30	4.00	4.31
State 2: Unemployed				
NR1	8.56	8.55	8.58	8.55
NR2	8.61	8.63	8.70	8.57
NR3	8.63	8.62	8.66	8.58
NR4	8.39	8.39	8.47	8.38
NR5	8.77	8.76	8.83	8.76
NR6	9.05	9.03	9.07	9.01
State 3: Inactive				
NR1	4.78	4.72	4.26	4.76
NR2	4.85	4.79	4.27	4.84
NR3	4.86	4.79	4.25	4.83
NR4	4.80	4.74	4.23	4.79
NR5	4.81	4.76	4.26	4.77
NR6	5.33	5.26	4.71	5.26

B. Appendix

Table B.1. MSE for the NR6 scenario with "extreme" PSUs.

Nationality group	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1:				
Employed				
Population	0.12	0.12	0.12	0.12
Irish	0.20	0.14	0.13	0.13
UK	6.37	5.96	5.02	6.03
OMS	12.58	12.47	9.92	12.62
NMS	4.25	4.32	4.21	4.32
Other nationals	4.30	4.31	3.74	4.28
State 2:				
Unemployed				
Population	0.07	0.07	0.07	0.07
Irish	0.08	0.08	0.07	0.07
UK	2.55	2.54	2.51	2.49
OMS	3.91	3.94	3.93	3.90
NMS	2.79	2.82	2.84	2.80
Other nationals	2.26	2.22	2.20	2.19
State 3:				
Inactive				
Population	0.09	0.09	0.09	0.09
Irish	0.22	0.11	0.10	0.11
UK	6.05	5.47	4.39	5.69
OMS	11.34	11.13	8.31	11.28
NMS	1.91	1.91	1.74	1.91
Other nationals	4.35	4.24	3.28	4.23

Table B.2. CV for the NR6 scenario with "extreme" PSUs.

Nationality group	Pre-Q3-2006	Current QNHS weights	Modified QNHS weights	Raking ratio
State 1:				
Employed				
Population	0.69	0.70	0.70	0.70
Irish	0.76	0.74	0.73	0.73
UK	5.06	4.99	4.74	5.02
OMS	5.56	5.53	4.94	5.57
NMS	3.11	3.12	3.08	3.12
Other nationals	4.43	4.43	4.12	4.41
State 2:				
Unemployed				
Population	2.23	2.25	2.25	2.25
Irish	2.51	2.48	2.47	2.48
UK	11.66	11.66	11.65	11.55
OMS	23.13	23.11	23.12	23.07
NMS	8.37	8.37	8.40	8.38
Other nationals	9.09	9.00	8.97	9.01
State 3:				
Inactive				
Population	0.79	0.79	0.79	0.79
Irish	0.87	0.83	0.80	0.81
UK	5.94	5.84	5.36	5.96
OMS	12.13	12.10	10.43	12.14
NMS	9.94	9.99	9.54	9.94
Other nationals	5.50	5.49	4.84	5.50

6. References

Barrett, A., A. Bergin, and E. Kelly. 2011. "Estimating the Impact of Immigration on Wages in Ireland." *The Economic and Social Review* 42: 1–26.

Barrett, A. and D. Duffy. 2008. "Are Ireland's Immigrants Integrating into Its Labor Market?" *International Migration Review* 42: 597–619. Doi: http://dx.doi.org/10.1111/j.1747-7379.2008.00139.x.

Central Statistics Office. 2011. "Quarterly National Household Survey Quarter 2 2011." Available at: http://www.cso.ie/en/media/csoie/releasespublications/documents/labourmarket/2011/qnhs_q22011.pdf (accessed 21 December 2015).

Central Statistics Office. 2014. "Population and Migration Estimates." Available at: http://www.cso.ie/en/surveysandmethodology/population/populationandmigrationestimates/ (accessed 14 January 2016).

Central Statistics Office. 2015a. "Estimated Population classified by Sex and Nationality, 2009–2015 [Table 9]." Available at: http://www.cso.ie/en/releasesandpublications/er/pme/populationandmigrationestimatesapril2015/ (accessed 19 January 2016).

Central Statistics Office. 2015b. "Standard Report on Methods and Quality for QNHS." Cork. Available at: http://www.cso.ie/en/qnhs/qnhsmethodology/ (accessed 19 January 2016).

Deming, W.E. and F.F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." *The Annals of Mathematical Statistics* 11: 427–444. Doi:http://dx.doi.org/10.1214/aoms/1177731829.

Deville, J.C. and C.-E. Särndal. 1992. "Calibration Estimation in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382.

Deville, J.C., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 1013–1020.

Djerf, K. 1997. "Effects of Post-Stratification on the Estimates of the Finnish Labour Force Survey." *Journal of Official Statistics* 13: 29–39.

Djerf, K. and P. Väisänen. 1993. "Effects of Post-Stratication on the Estimates of the Finnish Labour Force Survey." International Statistical Institute 49th Session, Contributed Papers, Book 1. Florence: ISI, 375–376.

Eurostat. 2002. "Monograph of Offcial Statistics – Variance Estimation Methods in the European Union." Luxembourg: Eurostat. Available at: http://ec.europa.eu/eurostat/ramon/statmanuals/files/KS-CR-02-001-EN.pdf (accessed 21 January 2016).

Eurostat. 2013. "Employment and Unemployment (Labour Force Survey) [Ireland]." Available at: http://ec.europa.eu/eurostat/cache/metadata/EN/employ_esqrs_ie.htm (accessed May 2015).

Eurostat. 2014. "Labour Force Survey in the EU, Candidate and EFTA countries – Main characteristics of national surveys, 2013." Eurostat Methodologies and Working Paper. Luxembourg: Eurostat.

Eurostat. 2015. "Share of Non-Nationals in the Resident Population, 1 January 2014 [Table 4]." Available at: http://ec.europa.eu/eurostatstatisticsexplained/index.php/Migration_and_migrant_population_ statistics (accessed 3 February 2016).

Holt, D. and T.M.F. Smith. 1979. "Post-Stratification." *Journal of the Royal Statistical Society Series A* 142: 33–46.

Hörngren, J. 1992. "The Use of Registers as Auxiliary Information in the Swedish Labour Force Survey." In Proceedings of the Workshop on Uses of Auxiliary Information in Surveys, 5–7 October 1992, Örebro. Statistics Sweden and the University of Örebro.

Houses of the Oireachtas. 2012. Parliamentary Debates on Migration Data, 18 September 2012 (Minister of State at the Department of the Taoiseach, Deputy Paul Kehoe). Available at: http://oireachtasdebates.oireachtas.ie/debates%20authoring/debateswebpack.nsf/takes/dail2012091800054?opendocument#WRD01250 (accessed 22 January 2016).

Jagers, P. 1986. "Post-Stratification Against Bias in Sampling." *International Statistical Review* 54: 159–167.

Kalton, G. 1983. "Compensating for Missing Survey Data." Ann Arbor, MI: Survey Research Center, Institute for Social Research, the University of Michigan.

King, S.L., B. Chopova, J. Edgar, J.M. Gonzalez, D. McGrath, and L. Tan. 2009. "Assessing Nonresponse Bias in the Consumer Expenditure Interview Survey." In Proceedings of the Section on Survey Research Methods: Joint Statistical Meetings of the American Statistical Association, 6 August 2009, Washington, DC. Available at: http://www.bls.gov/osmr/pdf/st090220.pdf (accessed June 2015).

Kingston, G., P.J. O'Connell, and E. Kelly. 2013. "Ethnicity and Nationality in the Irish Labour Market: Evidence from the QNHS Equality Module 2010." Dublin: Equality Authority/ESRI.

Martí, M. and C. Ródenas. 2012. "Measuring International Migration through Sample Surveys: Some Lessons from the Spanish Case." *Population* 67: 235–463.

Minnesota Population Center. 2014. "Integrated Public Use Microdata Series, International: Version 6.3 [Machine-readable database]." Minneapolis: University of Minnesota.

Nguyen, N.D. and P. Murphy. 2015. "To Weight or Not To Weight – A Statistical Analysis of How Weights Affect the Reliability of the Quarterly National Household Survey for Immigration Research in Ireland." *Journal of Economic and Social Reviews* 46: 567–603.

O'Boyle, N. and B. Fanning. 2009. "Immigration, Integration and Risks of Social Exclusion: The Social Policy Case for Disaggregated Data in the Republic of Ireland." *Irish Geography* 42: 145–164. Doi: http://dx.doi.org/10.1080/00750770903112795 (accessed 01 February 2016).

O'Connell, P.J. and F. McGinnity. 2008. "Immigrants at Work: Nationality and Ethnicity in the Irish Labour Market." Dublin: Equality Authority/ESRI.

Pérez-Duarte, S., C. Sánchez Muñoz, and V.-M. Törmälehto. 2010. "Reweighting to Reduce Unit Nonresponse Bias in Household Wealth Surveys: A Cross-Country Comparative Perspective Illustrated by a Case Study." In Proceedings of the Conference on European Quality in Statistics, May 4–6, 2010, Helsinki. Available at: http://www.ecb.europa.eu/home/pdf/research/hfcn/WealthSurveys.pdf (acccessed June 2015).

Robinson, D. and K. Reeve. 2006. *Neighbourhood Experiences of New Immigration: Reflections from the Evidence Base*. London: Joseph Rowntree Foundation.

Särndal, C.-E. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33: 99–119.

Särndal, C.-E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: Wiley.

Sautory, O. 2003. "CALMAR 2: A New Version of the CALMAR Calibration Adjustment Program." In Proceedings of Statistics Canada's Symposium 2003 Challenges in Survey Taking for the Next Decade. Available at: http://www.statcan.gc.ca/pub/11-522-x/2003001/session13/7713-eng.pdf (accessed June 2016).

Thomsen, I. 1973. "A Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Non-Response when Analyzing Survey Data." *Statistisk Tidskrift* 11: 278–285.

Thomsen, I. 1978. "A Second Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Non-Response when Analyzing Survey Data." *Statistisk Tidskrift* 16: 278–285.

Thomsen, I. and A.M.K. Holmøy. 1998. "Data from Surveys and Administrative Record Systems. The Norwegian Experience." *International Statistical Review* 66: 201–221.

United Nations. 2005. "Household Sample Surveys in Developing and Transition Countries." Available at: http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf (accessed June 2015).

Villund, O. 2010. "Effect of Non-Response Bias on Labour Market Statistics for Immigrants." In Proceedings of the 5th Workshop on Labour Force Survey Methodology, 15–16 May 2010, Paris. Available at: http://www.insee.fr/en/insee-statistique-publique/connaitre/colloques/wlfsm/documents/wlfsm-villund-paper.pdf (accessed June 2015).

Zhang, L.-C. 2000. "Post-Stratification and Calibration – A Synthesis." *The American Statistician* 54: 178.

Received August 2015
Revised February 2016
Accepted April 2016

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 719–732, http://dx.doi.org/10.1515/JOS-2016-0036

An Imputation Model for Dropouts in Unemployment Data

Petra Nilsson[1]

Incomplete unemployment data is a fundamental problem when evaluating labour market policies in several countries. Many unemployment spells end for unknown reasons; in the Swedish Public Employment Service's register as many as 20 percent. This leads to an ambiguity regarding destination states (employment, unemployment, retired, etc.). According to complete combined administrative data, the employment rate among dropouts was close to 50 for the years 1992 to 2006, but from 2007 the employment rate has dropped to 40 or less. This article explores an imputation approach. We investigate imputation models estimated both on survey data from 2005/2006 and on complete combined administrative data from 2005/2006 and 2011/2012. The models are evaluated in terms of their ability to make correct predictions. The models have relatively high predictive power.

Key words: Follow-up study; multiple regression imputation; register data.

1. Introduction

Active labour market policies (ALMPs) have increasingly been promoted in the Organisation for Economic Co-operation and Development (OECD) countries and transition economies as a principal means of dealing with unemployment. Evaluation of ALMPs is important for future policy making and implementation, but incomplete information in unemployment data is a fundamental problem. Many unemployment spells end for unknown reasons, which leads to an ambiguity regarding the labour market state. In the Swedish Public Employment Service's register, the percentage of exits for unknown reasons is approximately 20 percent. When evaluating ALMPs, assumptions have to be made concerning whether these unemployment spells ended because of work or not. In the agency's performance reports only known exits to employment are presented, which means that the number of exits to employment is underestimated.

Similar data problems exist in several countries, and these countries have employed different methods of dealing with incomplete information in unemployment administrative data. For example, Wilke (2009) constructs bounds for unemployment duration in UK administrative unemployment data to describe the effect of missing information on interval information (length of unemployment) and destination states (employment, unemployment, retired, etc.). Arntz et al. (2007) conduct a similar exercise with German

[1] Research and Evaluation, Swedish Public Employment Service, SE-113 99 Stockholm, Email: petra.nilsson@arbetsformedlingen.se

Acknowledgments: The author thanks Anders Harkman, who came up with the idea and helped along the way, and Linda Wänström for her valuable contribution.

data. Our interest is in the destination states, and we choose to explore an imputation approach rather than constructing bounds.

Bring and Carling (2000) use data from a small survey in 1994 to estimate an imputation model that Statistics Sweden has used to compensate for missing destination states since 1994. This article extends the Bring and Carling (2000) methodology and uses more recent data. We estimate new imputation models based on both survey and register data. We use data from a larger survey conducted in 2005 and 2006 by the Swedish Public Employment Service. We also use data from the Swedish Longitudinal Integration Database for Health Insurance and Labour Market Studies (LISA), which includes information about gainful employment as of November each year. The imputation models are evaluated and the concordance between predicted values and survey/register data is studied. The predicted power of the new models is compared to the predicted power of Bring and Carling's model as well as to random imputation.

We also present the employment rate among dropouts over time based on LISA data. Another contribution is that the new imputation models based on survey data deal with nonresponse. Bring and Carling (2000) used only survey responses in their model and did not account for survey nonresponse. The nonresponse rate in the survey used in Bring and Carling (2000) is 20 percent. This refers both to unit and item nonresponse, since there was only one question in the survey.

2. Data

Survey data and register data are used to estimate imputation models for dropouts in unemployment data. Both survey data and register data contain measurement error. In register data there is unobserved/misreported information, since not all individual employment biographies are covered by the administrative process. Bound et al. (2001) discuss the causes of measurement errors in survey reports. The longer the recall period, the more difficult the reporting task and the less salient the event the more difficult it is to retrieve the information requested. Socially undesirable events tend to go unreported, while the opposite is true for socially desirable events. See for example Pyy-Martikanen and Rendtel (2009), who used Finnish linked survey and administrative data to analyse measurement error in survey data.

2.1. *The Survey*

The survey was conducted on twelve different measurement occasions between September 2005 and August 2006. Each measurement occasion includes exits for unknown reasons during one week, where the job seeker does not return to the Public Employment Service within 14 days. An unrestricted random selection of 300 periods of registration were made each measurement week. The total sample is therefore 3,600 periods of registration. We chose to include measurement weeks from as many different periods as possible during one year to take into account any seasonal effects.

The survey can be seen as a stratified sample with measurement weeks as strata. A stratification of a finite population $U = \{1, \ldots, k, \ldots, N\}$ means a partition of U in H subsets of the population (Lundström and Särndal 2001). The number of elements in stratum h is denoted N_h and the sample size in stratum h is denoted n_h. The probability that

a given element is included in the sampling, the inclusion probability, is given by

$$\pi_k = \frac{n_h}{N_h}. \tag{1}$$

Let

$$d_k = \frac{1}{\pi_k} \tag{2}$$

denote the design weight of element k.

Note that it is the period of registration and not the individual that constitutes an element in the survey. The population consists of unique periods of registration, but not of unique individuals, as some people occur in the data set multiple times. The outcome of registration periods for the same individual are probably correlated and the observations cannot be assumed to be independent. This problem is called correlated failure-time modelling or multiple spells modelling and is studied for example in the economic literature (Lancaster 1979; Heckman and Singer 1982).

The correlation structure is ignored in this article, which might lead to an underestimation of the variability of the imputation model. A very large percentage of registration periods concern unique individuals, however. In fact, 97.4 percent of registration periods that ended with deregistration for unknown reasons during the measurement weeks concern unique persons. Of the periods that constitute the sampling frame in the survey, periods where the individuals have not returned within 14 days, 98.5 percent concern unique individuals. In the sample, 99.8 percent concern unique individuals; three individuals are found twice.

The survey was conducted in the form of computer-aided telephone interviews by the Public Employment Service's interview unit. The interviews were conducted as close to deregistration from the Public Employment Service as possible, in order for the interviewed persons to be able to recall their work situation when they ceased to have contact with the Public Employment Service. Since there is a 14-day wait in order to exclude the return of job seekers to the Public Employment Service, the interviews were conducted within two to three weeks of deregistration.

In the survey, the individuals were asked about their current work situation ("What is your work situation today?"). The response options were the following:

1. Have work (full-time)
2. Have work (part-time)
3. Studying/in training
4. Participating in a labour market programme
5. Have started my own company
6. Long-term sick leave/sick leave/on parental leave
7. Unemployed/seeking work
8. Other

Individuals responding according to option 1, 2, or 5 are defined as having found work. The interviewers did not read out the response options to the question. When the interviewed person had difficulties giving an answer that fitted the response options, the interviewer helped by interpreting the purpose of the question.

Out of the total sample of 3,600 periods, 2,443 responded, giving a response rate of 68 percent. Nonresponse in the survey is thus 32 percent. Older persons, persons born outside Europe, persons with a low level of education, and persons without unemployment insurance are overrepresented in the nonresponse group. Probably, a lower extent of these persons had found work than those who responded in the survey (see, for example, Bennmarker et al. 2007).

There is a risk that estimates using data only from respondents will be biased. We therefore impute missing values. There are various methods of imputation; we use regression imputation as described in Lundström and Särndal (2001). Since both socioeconomic and employment-related explanatory variables can be linked to the individuals that drop out, missing data may be imputed using a logistic regression model that explains which categories of individuals have the greatest probability of having found work. Rubin (1996) recommends that an imputation model contain as many relevant variables as possible and the model used to impute the nonresponse in the survey includes many socioeconomic and employment-related explanatory variables; see Appendix. Variables in Bennmarker et al. (2007) were considered.

We denote by y_k whether or not a period of registration k has ended because of work; $y_k = 1$ if period k ended because of work and $y_k = 0$ if period k did not end because of work. We assume that response or register values are obtained for the elements in a set denoted r. Regression imputation gives an imputed value for element k according to

$$\hat{y}_k = z_k' \hat{\beta} \tag{3}$$

where z_k is the value of the imputation vector for element k by $z_k = (z_{1k}, \ldots, z_{jk}, \ldots, z_{Jk})'$, a column vector with J explanatory variables, where z_{jk} is the value, for element k, of the jth explanatory variable and

$$\hat{\beta} = \left(\sum_r d_k z_k z_k' \right)^{-1} \sum_r d_k z_k y_k. \tag{4}$$

$\hat{\beta}$ is a vector of regression coefficients, resulting from the fit of a regression using the data (y_k, z_k) available for $k \in r$ and weighted with d_k.

Replicates of data containing regression-imputed values tend to have a lower degree of variance than data containing observed values y_k. We therefore add a randomly selected residual. Then the imputed value for element k is

$$\hat{y}_k = z_k' \hat{\beta} + e_k^* \tag{5}$$

where e_k^* is a randomly selected residual from the data set containing calculated residuals $\{e_k : k \in r\}$, where

$$e_k = y_k - z_k' \hat{\beta}. \tag{6}$$

Each missing value is imputed 20 times. The y_k values for the respondents are the same in all data sets, while the imputed values are different.

2.2. Register Data

In addition to survey data, we also use administrative records on gainful employment for this study when constructing imputation models. The LISA database is used to impute missing information about employment status to obtain complete combined administrative data sets. The administrative records in LISA are limited to the month of November so the combined administrative data sets are also limited to November. Information about gainful employment in LISA is used to impute missing destination states for dropouts in November each year.

The LISA database holds annual registers and includes all individuals 16 years of age and older registered in Sweden. It is available in spring the following year. The individuals are classified as employed if they are assumed to have worked for at least four hours during November. The estimation in LISA is model based where the correlation between several variables, for example information about payments from employers, is used for the classification. There are some misclassifications in the data compared to real working hours in November. The risk of misclassification is larger for persons who were working only parts of the year and for persons with a weaker connection to the labour market. Misclassifications are partly due to errors in the model but also due to incomplete information.

Imputation models for dropouts are estimated on complete combined administrative data from November 2005/2006 and from November 2011/2012. The years 2005/2006 are chosen to enable a comparison between a model based on administrative data and a model based on survey data for the same years. An imputation model is also estimated on administrative data from 2011/2012 to enable a comparison with later years.

Table 1 describes the combined administrative data sets for the years 2005/2006 and 2011/2012. Item nonresponse regarding employment status not filled in by LISA for 2005/2006 is 1.5 percent and for 2011/2012 it is two percent. Item nonresponse is more common for persons 55 years or older, for persons born outside of Europe, and for persons with a low level of education.

For the employment rate among dropouts over time we use all available data, which is November data for the years 1992 to 2012. For the years 1992 to 2006, the number of periods each November is approximately 12,000, while for the years 2007 to 2012 the number of periods each November has dropped to approximately 7,000 per year. The sharp decline from 12,000 to 7,000 is due to better administrative routines at the Swedish Public Employment Service. Fewer ended unemployment spells lack employment status.

Table 1. Description of register data November 2005/2006 and November 2011/2012 respectively.

	2005/2006 Complete data	Item nonresponse	2011/2012 Complete data	Item nonresponse
Number of observations	20,566	311	14,375	282
16–24 years (%)	45.7	15.8	43.8	20.9
55–66 years (%)	3.6	7.4	4.6	8.2
Born outside Europe (%)	17.2	24.4	29.5	44.0
Functional impairment (%)	3.6	2.9	6.6	3.6
Compulsory school (%)	26.2	52.8	29.5	44.3

Item nonresponse regarding employment in November not filled in by LISA is about one percent in the beginning of the period but close to two percent for later years.

3. The Employment Rate

Figure 1 shows the employment rate among dropouts in November each year according to combined administrative data. For the years 1992 to 2006 the employment rate is close to 50, but from 2007 the employment rate drops to 40 or less. The decline in the employment rate is probably due to the better administrative routines mentioned above.

The estimated employment rate is based on the 20 nonresponse imputed replicates of data, that is, 20 separate models are estimated. The different parameter estimates are then combined as described in Rubin (1987).

Suppose that $\hat{Q}_i$ is an estimate of a scalar quantity of interest, obtained from a data set i, $i = 1, 2, \ldots, m$ and $\hat{W}_i$ is the variance associated with $\hat{Q}_i$. The overall estimate is the average of the individual estimates from the m complete replicates of data

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i. \tag{7}$$

Assume that $\bar{W}$ is the within-imputation variance, which is the mean value of the estimates from the m complete replicates of data

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m} \hat{W}_i \tag{8}$$

and B is the between-imputation variance

$$B = \frac{1}{m-1}\sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2 \tag{9}$$

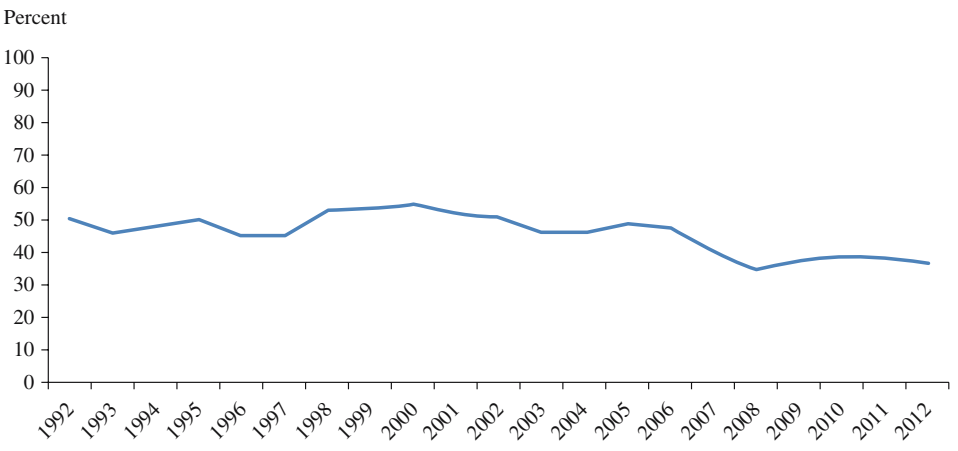


Fig. 1. Employment rate among dropouts in November 1992 to November 2012 according to combined administrative data.

Table 2. The estimated employment rate among dropouts based on the 2005/2006 survey.

	Estimated percentage	Stand. errors	95% Confidence Interval
Based both on the response group and the nonresponse group	47.3	1.1	(45.1; 49.4)
Based on the respondents	50.7	1.0	(48.6; 52.7)
Based on the nonresponse group (imputed values)	39.9	2.6	(34.8; 45.1)

then the total variance is

$$T = \bar{W} + (1 + \frac{1}{m})B. \tag{10}$$

Table 2 shows the estimated employment rate among dropouts based on the survey from 2005/2006.

The estimated employment rate is 47.3 percent. A 95% confidence interval for this percentage is of the magnitude plus/minus two percentage points. The estimated employment rate based on the survey is close to the employment rate according to the combined administrative data for November 2005 and November 2006, which is 48.1 percent. Table 2 also shows the estimated employment rate based only on those who responded to the survey and based only on the imputed values for the nonresponse group.

The estimated employment rate among dropouts in the 1994 survey is 44.7 percent (Bring and Carling 2000). The sample in the 1994 survey was drawn from the population of dropouts in January and February. In the combined administrative data for November 1994, the employment rate is 47.6 percent.

A factor that affects the comparability of survey data and the combined administrative data is that administrative data refers to the month of November while survey data refers to different periods during the year. The employment rate can be different depending on the season. Table 3 displays the estimated employment rate per measurement week in the 2005/2006 survey. Standard errors and 95% confidence intervals are also shown.

Table 3. The estimated employment rate among dropouts in the 2005/2006 survey divided into measurement week.

Measurement week	Estimated employment rate	Stand. errors	95% Confidence Interval
Week 34 2005 (Aug)	44.3	3.1	(38.1; 50.5)
Week 35 2005 (Aug-Sep)	44.8	3.1	(38.7; 50.9)
Week 36 2005 (Sep)	43.0	3.1	(36.8; 49.1)
Week 37 2005 (Sep)	47.1	3.4	(40.3; 53.9)
Week 40 2005 (Oct)	46.6	3.4	(39.9; 53.4)
Week 3 2006 (Jan)	43.9	3.3	(37.4; 50.5)
Week 5 2006 (Jan-Feb)	44.8	3.3	(38.4; 51.3)
Week 9 2006 (Feb-Mar)	48.6	3.3	(42.1; 55.0)
Week 14 2006 (Apr)	52.4	3.2	(46.1; 58.8)
Week 22 2006 (May-Jun)	56.3	3.3	(49.9; 62.8)
Week 25 2006 (Jun)	54.5	3.6	(47.3; 61.7)
Week 31 2006 (Jul-Aug)	45.6	3.2	(39.2; 51.9)

The estimated employment rate varies from 43.0 to 56.3 between measurement weeks. The employment rate is lower at the beginning of the autumn and spring semester when the exit rate to education is high, and the employment rate is higher at the beginning of the summer when the exit rate to summer jobs is high.

4. Imputation Models for Dropouts

4.1. The Models

Imputation models for dropouts are estimated on survey data from 2005/2006 and on complete combined administrative November data from 2005/2006 and 2011/2012. The models are logistic regression models and the dependent variable is whether the person is employed or not. We estimate

$$\hat{P}(Y_k = 1|z_k) = \frac{1}{1 + \exp(-z_k\hat{B}')}. \tag{11}$$

Age, country of birth, functional impairment, education, membership of an unemployment insurance fund, status prior to deregistration, and experience are used as explanatory variables. We want to include as many relevant socioeconomic and employment-related explanatory variables as possible to improve the models' predictive power, but at the same time we want to keep the models as simple as possible to use. Variables used in Bring and Carling (2000) and Bennmarker et al. (2007) have been considered and variables with p-values smaller than 0.05 when estimated on administrative data are used in the final models.

Since survey data includes samples from different periods during a year, it is possible to include month of deregistration as an explanatory variable. Two different models are estimated on survey data; one with (Model 1) and one without (Model 2) month of deregistration.

For survey data we estimate the imputation models on the 20 nonresponse imputed replicates of data, that is, 20 separate models are estimated. The different parameter estimates are then combined as described in Rubin (1987). Table 4 displays the imputation models based on survey data.

The employment rate is lower for, for example, older persons, persons with a low level of education, persons born outside of Europe, and persons with a functional impairment. Persons being registered as part-time or temporarily employed prior to deregistration have a higher employment rate than persons categorised as unemployed. Persons with many previous transitions to work and members of an unemployment insurance fund also have a higher employment rate. The alternative model (Model 2), where month of deregistration is included, shows that those deregistered in May, June, or July are employed to a higher extent.

Table 5 shows imputation models estimated on complete combined administrative data for November 2005/2006 (Model 3) and November 2011/2012 (Model 4).

The interpretation of the estimates is basically the same as for Table 4. One difference is that the estimate for the intercept and the estimate for individuals 16–24 years have

Table 4. Imputation models based on survey data 2005/2006, eleven and twelve covariates respectively.

	Model 1			Model 2		
	11 variab. Estimate	Stand. errors	*p*-value	12 variab. Estimate	Stand. errors	*p*-value
Intercept	0.14	0.14	0.33	0.07	0.14	0.61
16–24 years	−0.37	0.11	0.00	−0.37	0.11	0.00
55–66 years	−0.36	0.18	0.05	−0.37	0.18	0.05
Born outside Europe	−0.53	0.12	<.0001	−0.53	0.12	<.0001
Functional impairment	−0.37	0.31	0.23	−0.38	0.31	0.22
Compulsory school only	−0.30	0.12	0.02	−0.31	0.13	0.02
Member of an unemployment insurance fund	0.52	0.11	<.0001	0.52	0.11	<.0001
Work prior to deregistration	0.84	0.14	<.0001	0.85	0.14	<.0001
Other status prior to deregistration	−0.48	0.13	0.00	−0.50	0.13	<.0001
Number of periods of registration	−0.16	0.03	<.0001	−0.16	0.03	<.0001
Number of transitions into work	0.31	0.06	<.0001	0.31	0.06	<.0001
No experience	−0.15	0.12	0.21	−0.15	0.12	0.21
Deregistration in May, June or July				0.50	0.12	<.0001

Table 5. Imputation models based on complete combined administrative November data 2005/2006 and 2011/2012.

	Model 3			Model 4		
	2005/2006 Estimate	Stand. errors	*p*-value	2011/2012 Estimate	Stand. errors	*p*-value
Intercept	−0.50	0.05	<.0001	−0.67	0.06	<.0001
16–24 years	0.22	0.04	<.0001	0.42	0.05	<.0001
55–66 years	−0.43	0.09	<.0001	−0.48	0.10	<.0001
Born outside Europe	−0.50	0.05	<.0001	−0.20	0.05	<.0001
Functional impairment	−0.52	0.10	<.0001	−0.50	0.10	<.0001
Compulsory school only	−0.54	0.04	<.0001	−0.64	0.05	<.0001
Member of an unemployment insurance fund	1.22	0.04	<.0001	1.26	0.05	<.0001
Work prior to deregistration	1.66	0.05	<.0001	1.86	0.07	<.0001
Other status prior to deregistration	−0.24	0.05	<.0001	−0.45	0.07	<.0001
Number of periods of registration	−0.24	0.01	<.0001	−0.24	0.02	<.0001
Number of transitions into work	0.34	0.02	<.0001	0.40	0.03	<.0001
No experience	−0.41	0.04	<.0001	−0.64	0.06	<.0001

changed sign compared to the models based on survey data. The standard errors are about half of those from the models based on survey data (Table 4).

4.2. The Predictive Power of the Models

The predictive power of the imputation models can tell us whether or not the models can be used to impute missing values for those who leave the Public Employment Service for unknown reasons. By predictive power we mean the percentage of correct predictions.

We compare the predictive power of each model in Table 4 and Table 5 (Model 1–4) for both the survey data from 2005/2006 and the complete combined administrative data from November 2005/2006 and November 2011/2012. We also calculate the predictive power for random imputation and for Bring and Carling's model based on survey data from 1994. In the cross validation, the imputation model has been estimated on 60 percent of the data and evaluated against the remaining 40 percent.

The imputation models estimate a probability between 0 and 1 that the individuals have found work. Imputation then requires a threshold, that is, at which predicted values the imputed value of having found work, $\hat{y}_k = 1$, or the imputed value of not having found work, $\hat{y}_k = 0$ should be classified. For each model, we select a threshold so that the imputation produces the employment rate observed in the dataset. For survey data the employment rate is 47.3, for the combined administrative data from November 2005/2006 it is 48.1 and for November 2011/2012 it is 37.4.

Table 6 shows the predictive power of each model for the different data sets. For survey data from 2005/2006, random imputation has the lowest predictive power, 50 percent correct predictions. The imputation model based on survey data from 1994 has 54 percent correct predictions. The imputation models investigated in this article have higher predictive power, 68 percent for all models. For the combined administrative data from November 2005/2006, the models based on combined administrative data have higher predictive power than the models based on survey data. The model based on combined administrative data from the same years (2005/2006) has the highest predictive power, 76 percent. For administrative data from 2011/2012, the models based on administrative

Table 6. Percent correct predictions.

	Model Survey 1994	Model 1 Survey 2005/2006 11 variab.	Model 2 Survey 2005/2006 12 variab.	Model 3 Register 2005/2006	Model 4 Register 2011/2012	Model Random
Data: Survey 2005/2006	54	68	68	68	68	50
Data: Combined administrative November data 2005/2006	58	74	74	76	75	50
Data: Combined administrative November data 2011/2012	61	72	72	74	74	53

data again have higher predictive power than the models based on survey data. Both models based on administrative data have the same predictive power, 74 percent.

5. Conclusions and Closing Discussion

Imputation can be used in evaluations using unemployment data as a means of dealing with missing information about destination state after a period of unemployment. It also can be used in the Public Employment Service's performance reports when the number of exits to employment is presented to avoid underestimation.

Two imputation models based on survey data and two models based on combined administrative data were investigated. The four models all have similar predictive power. The models based on administrative data have slightly higher predictive power than the models based on survey data.

The two imputation models using 2005/2006 survey data are based on more data and have a higher predictive power than the imputation model suggested in Bring and Carling (2000), which is estimated on a small 1994 sample. The new imputation models based on survey data from 2005/2006 and multiple imputation deal with nonresponse in a more satisfactory way. According to the new survey, the estimated employment rate among dropouts is 47 percent for 2005 and 2006, which is consistent with administrative November data for the same years.

One difference between survey data and combined administrative data is that administrative data refers to the month of November, while survey data refers to twelve different measurement occasions during 2005 and 2006. We have no information about the predictive power of the investigated imputation models for all dropouts in unemployment data. We therefore cannot say which model is the best. Probably it does not matter a great deal which model is used. One suggestion is to use the imputation model based on the combined administrative data from November 2011/2012, which is the model based on the latest available data.

Appendix

In the model used to impute survey data the explanatory variables are:

female
16–24 years
35–44 years
45–66 years
born in the Nordic Countries
born abroad
functional impairment
compulsory school only
higher education $<= 2$ years
higher education > 2 years
experience in professions applied for
seeking only full-time work
seeking work beyond commuting distance
member of an unemployment insurance fund

participating in the activity guarantee
status work prior to deregistration
other status prior to deregistration (not work or unemployment)
forest county
other counties (not forest or major-city region)
number of periods of registration the previous five years
number of transitions into work the previous five years
Deregistration in the months May to July
16–24 years and member of an unemployment insurance fund
35–44 years and member of an unemployment insurance fund
45–66 years and member of an unemployment insurance fund
16–24 years and experience
35–44 years and experience
45–66 years and experience

6. References

Arntz, M., S. Lo, and R. Wilke. 2007. "Bounds Analysis of Competing Risks: A Nonparametric Evaluation of the Effect of Unemployment Benefits on Migration in Germany." ZEW - Centre for European Economic Research Discussion Paper No. 07-049. Doi: http://dx.doi.org/10.2139/ssrn.1010286.

Bennmarker, H., K. Carling, and A. Forslund. 2007. *Vem blir långtidsarbetslös?* Report 2007:29. Uppsala: Institute for Labour Market Policy Evaluation (IFAU). Available at: http://www.ifau.se/globalassets/pdf/se/2007/r07-20.pdf (accessed June 1, 2016).

Bound, J., C. Brown, and N. Mathiowetz. 2001. "Measurement error in survey data." In *Handbook of Econometrics*, vol. 5, edited by J. Heckman and E. Leamer, 3705–3833. Amsterdam: Elsevier. Available at: http://www.psc.isr.umich.edu/pubs/pdf/rr00-450.pdf (accessed June 1, 2016).

Bring, J. and K. Carling. 2000. "Attrition and Misclassification of Drop-outs in the Analysis of Unemployment Duration." *Journal of Official Statistics* 16: 321–330. Available at: http://www.jos.nu/Articles/abstract.asp?article=164321 (accessed June 1, 2016).

Heckman, J. and B. Singer. 1982. "Population Heterogeneity in Demographic Models." In *Multidimensional Mathematical Demography*, edited by K. Land and A. Rogers, 567–599. New York: Academic Press. Available at: http://www.popline.org/node/410098 (accessed June 1, 2016).

Lancaster, T. 1979. "Econometric Methods for the Duration of Unemployment." *Econometrica* 47: 939–956. Doi: http://dx.doi.org/10.2307/1914140.

Lundström, S. and C-E. Särndal. 2001. *Estimation in the Presence of Nonresponse and Frame Imperfections*. Örebro: SCB-Tryck.

Pyy-Martikainen, M. and U. Rendtel. 2009. "Measurement Errors in Retrospective Reports of Event Histories. A Validation Study with Finnish Register Data." *Survey Research Methods* 3: 139–155. Doi: http://dx.doi.org/10.18148/srm/2009.v3i3.2372.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Rubin, D.B. 1996. "Multiple Imputation After 18+ Years (with discussion)." *Journal of the American Statistical Association* 91: 473–489. Doi: http://dx.doi.org/10.1080/01621459.1996.10476908.

Wilke, R. 2009. "Unemployment Duration in the United Kingdom: An Incomplete Data Approach." Doi: http://dx.doi.org/10.2139/ssrn.1348019.

Received January 2015
Revised April 2016
Accepted April 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 733–745, http://dx.doi.org/10.1515/JOS-2016-0037

The Marginal Effects in Subgroup Decomposition of the Gini Index

Tomson Ogwang[1]

In this article, we derive the elasticity of the Gini index with respect to changes in subgroup incomes for subgroups that are characterized by significant income separation. The resulting elasticity, which is structurally similar to that of the empirically popular Lerman and Yitzhaki's (1985) elasticity for Gini income-source decomposition, entails easy and transparent computations. Some possible checks for income separation are described and an illustrative example using Canadian data is provided. The advantages of the proposed methodology over the Shapley value approach to Gini subgroup decomposition are stated.

Key words: Subgroup decomposition; income separation; Gini index; elasticity; pseudo-Lorenz regression curve.

1. Introduction

In a recent article, Ogwang (2014) developed a convenient method of decomposing the Gini index by population subgroups where the subgroups may be formed by gender, race, occupation, region, and so on. The advantage of Ogwang's approach is that the overall Gini index is simultaneously decomposed into the traditional within-group, between-group, and interaction (overlapping) components as well as decomposed by the contributions of the various subgroups to overall inequality. Hence, Ogwang's approach entails a two-way decomposition of the Gini index solely by population subgroups.

Ogwang's approach is able to isolate the contributions of the various subgroups to the overall Gini index, providing an alternative way of reconciling Gini subgroup decomposition with its income-source decomposition, for which the contributions of the various income sources to the overall Gini index are also isolated. It should be mentioned from the outset that other ways of reconciling subgroup decomposition of the Gini index with its income-source decomposition have also been proposed in the literature. One example of such a reconciliation is the multidimensional decomposition considered by Mussard (2004), Mussard and Richard (2012), Mussard and Savard (2012), and Mussini (2013), among others, which entails simultaneous decomposition of the Gini index by

[1] Brock University – Economics, 500 Glenridge Avenue St. Catharines Ontario L2S3A1, Canada. Email: togwang@brocku.ca

Acknowledgments: The author would like to thank D. Cho, L. Kwong, J. F. Lamarche, R. Watuwa, the Associate Editor and three anonymous referees for their valuable comments. Thanks are also extended to the seminar participants at the African Economic Research Consortium in Nairobi for their valuable insights. Research support from the Council for Research in the Social Sciences at Brock University is also gratefully acknowledged. The usual disclaimer is applicable.

population subgroups and income source. Other examples include the regression approach to inequality decompositions (e.g., Cowell and Fiorio 2011) and the Shapley decomposition (e.g., Chantreuil and Trannoy 2011; Shorrocks 2013).

From a policy perspective, the information on the contributions of the various subgroups to the overall Gini index is potentially very useful. This is because the contributions could in principle be used in the analysis of the marginal effects by revealing the elasticity of the Gini index with respect to subgroup-income changes. The elasticity statistic, which indicates the extent to which a small proportionate change in the incomes of the members of a particular population subgroup increases or decreases overall inequality, controlling for the incomes of the members of other population subgroups, aids policy makers in devising appropriate inequality-reduction strategies. For example, the statistic could be used to gauge the extent to which income changes (e.g., transfer payments) targeting members of a particular population subgroup (e.g., race) affect the overall Gini index. Hence, it is important to report the relevant elasticities in empirical studies involving subgroup decompositions of the Gini index.

One of the earliest endeavors to derive the elasticity of the Gini index with respect to changes in subgroup incomes as a function of the constituent subgroup concentration indexes was proposed by Podder (1993). Subsequently, Aaberge et al. (2005) developed a general theoretical framework for determining the elasticity of the Gini index with respect to changes in subgroup incomes from the estimated parameters of the so-called pseudo-Lorenz regression curve. Mussard and Richard (2012) and Jurkatis and Strehl (2014) have considered the marginal effects, including elasticities, in the context of multidimensional decompositions of the Gini index.

Shorrocks (2013) has articulated how the Shapley decomposition could be used in the analysis of the marginal effects in both subgroup and income decompositions of any inequality measure, including the Gini index. In principle, the Shapley decomposition could be extended to the elasticity analysis. Charpentier and Mussard (2011), Cowell and Fiorio (2011), Chantreuil and Trannoy (2011), and Shorrocks (2013), among others, provide detailed discussions of the strengths and weaknesses of the Shapley decomposition. One of the main strengths of the decomposition, which is relevant for elasticity analysis, is the ability to express overall inequality (or poverty) as the sum of the contributory factors. Three major weaknesses of the Shapley decomposition in the context of the Gini index are particularly noteworthy. First, as pointed out in Section 3 below, the Shapley subgroup decomposition of the Gini index entails loss of information contained in the interaction component of the Gini index. This is because in the decomposition the interaction component of the Gini index is absorbed into the within-group or between-group component. Second, as pointed out by Shorrocks (2013, 117), the Shapley subgroup decomposition does not solve the so-called "subgroup inconsistency" problem associated with the Gini index, where subgroup inconsistency refers to a situation where overall inequality (as measured by the Gini index) could rise even though inequality in every constituent subgroup has fallen and the subgroup mean incomes and subgroup sizes are unchanged. Third, the Shapley decomposition results are sensitive to the structure of the income inequality game (Charpentier and Mussard 2011) or the hierarchical structures associated with the game (Chantreuil and Trannoy 2011).

Although many empirical papers that have reported the elasticity of the Gini index in the context of income-source decompositions can be found in the literature (e.g., Lerman and Yitzhaki 1985; Stark et al. 1986; Leibbrandt et al. 2000; Lopéz-Feldman 2006), it is not yet commonplace for empirical researchers to report the elasticity in the context of Gini subgroup decompositions – even though this elasticity has policy significance, as has already been alluded to above. Podder (1993) and Chatterjee and Podder (2007) are rare empirical papers reporting the elasticity of the Gini index with respect to changes in subgroup incomes. The elasticities they report are functions of the concentration indexes for the constituent population subgroups. An extensive literature search reveals hardly any empirical papers that report elasticities in the context of the Shapley value approach to Gini subgroup decomposition.

The paucity of empirical papers reporting the elasticity of the Gini index with respect to changes in subgroup incomes is partly ascribed to the difficulties with the empirical implementation of the existing methods. Hence, the search for convenient methods of estimating the elasticity continues.

An important first step in the derivation of the Gini subgroup decomposition elasticities is to isolate the contributions of the various subgroups to the overall Gini index. Hence, in principle, any approach to Gini subgroup decomposition that isolates the contributions of the various subgroups to the overall Gini index could be exploited in the derivations of the relevant elasticities, albeit with varying degrees of difficulty. For example, as pointed out by a referee, the subgroup decomposition considered by Radaelli (2010), which entails a breakdown of the overall Gini index into a within-group component and a between-group component in two steps, provides the contributions of the various subgroups to the overall Gini index. However, this approach, like the Shapley decomposition, cannot be conveniently extended for purposes of elasticity analysis.

As already mentioned above, the approach developed by Ogwang (2014) also enables the isolation of the contributions of the various subgroups to the overall Gini index. In light of the computational convenience associated with Ogwang's approach, it makes sense to exploit the approach for the purposes of Gini elasticity analysis.

Therefore, the present article aims to extend Ogwang's 2014 article by deriving the elasticity of the Gini index with respect to changes in the incomes of all the income-receiving units constituting a particular population subgroup. The derivation is made under the assumption of significant subgroup-income separation. Under this assumption, small proportionate changes in the incomes of all income-receiving units in the subgroup of interest do not change the rankings of these income-receiving units relative to those of all receiving units in all population subgroups; that is, the new ranks after the income changes are identical to the corresponding ranks before the changes. As will be seen below, the resulting elasticity turns out to be structurally similar to that of the widely reported Lerman and Yitzhaki's (1985) elasticity for Gini income-source decompositions. Some plausible empirical situations that entail subgroup-income separation are mentioned below, as are possible income-separation checks.

The rest of the article is structured as follows: in Section 2, we derive the elasticity of the Gini index with respect to changes in subgroup incomes under the assumption of significant subgroup-income separation. In Section 3, some possible checks for income

separation are briefly described. An illustrative example using Canadian data is provided in Section 4. The concluding remarks are made in Section 5.

2. Deriving the Gini Subgroup Decomposition Elasticities

To facilitate the exposition of the Gini subgroup decomposition elasticities, it is helpful to provide a brief overview of the nature of the contributions of the subgroups to the overall Gini index, as presented in Ogwang's (2014) article, from which the marginal effects are then derived in this article.

Suppose that n income-receiving units are classified into k mutually exclusive and exhaustive subgroups (e.g., by gender, age, race, education, occupation, or region). The k subgroups are arranged in ascending order of their subgroup mean incomes, but it is not necessary to arrange the incomes within each subgroup in any particular order.

We shall adopt the following notations also employed in Ogwang's (2014) article:

n_j ($j = 1, 2, \ldots, k$): the number of income-receiving units in the subgroup with the jth smallest mean income.
$n = \sum_{j=1}^{k} n_j$: the total number of income-receiving units in all the k subgroups.
y_{ij} ($i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k$): the income of the ith income-receiving unit in the subgroup with the jth smallest mean income in which case $\bar{y}_j = (1/n_j)\sum_{i=1}^{n_j} y_{ij}$ is the mean income for the same subgroup.
r_{ij} ($i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k$): the rank of y_{ij} in relation to the incomes of all the $n = \sum_{j=1}^{k} n_j$ income-receiving units in all the k subgroups.
r'_{ij}($i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k$): the rank of y_{ij} in relation to the incomes of only the n_j income-receiving units in the subgroup with the jth smallest mean income
$\tilde{r}_{ij}$ ($i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k$): the rank of y_{ij} in relation to the incomes of all the $n = \sum_{j=1}^{k} n_j$ income-receiving units, assuming that $y_{ij} = \bar{y}_j$.

In the construction of the ranks r_{ij}, r'_{ij}, and $\tilde{r}_{ij}$ tied incomes are assigned the average of the ranks they would have been assigned assuming that they were not tied. For example, if under a particular ranking rule two incomes are tied and one of them would have been assigned a rank of two and the other would have been assigned a rank of three in the absence of ties, then under this ranking rule both incomes are assigned the average rank of $(2 + 3)/2 = 2.5$.

The following rank transformations are also required: $r^*_{ij} = 2r_{ij} - n - 1$ $(i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k)$, $r^{*\prime}_{ij} = 2r'_{ij} - n_j - 1 (i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, k)$; $\tilde{r}^*_{ij} = 2\tilde{r}_{ij} - n - 1 (i = 1, 2, \ldots, n_j;\ j = 1, 2, \ldots, k)$; $\hat{r}^*_{i1} = (r_{i1} - r_{i1}{}')$, $i = 1, 2, \ldots, n_1$ and $\hat{r}^*_{ij} = \left(r_{ij} - r'_{ij} - \sum_{i=1}^{j-1} n_i\right), i = 1, 2, \ldots, n_j; j = 2, \ldots, k$.

Following Ogwang (2014), the overall Gini index is given by

$$G = \sum_{j=1}^{k} p_j s_j G_{Wj} + \sum_{j=1}^{k} p_j s_j G_{Bj} + 2\sum_{j=1}^{k} p_j s_j G_{Ij} \tag{1}$$

where $p_j = n_j/n$ is the population share of the subgroup with the jth smallest mean income, $j = 1, 2, \ldots, k$; $s_j = \left(\sum_{i=1}^{n_j} y_{ij}\right)/\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}$ is the income share of the subgroup with the jth smallest mean income, $j = 1, 2, \ldots, k$; and G_{Wj}, G_{Bj}, and G_{Ij} are the within-group Gini, the between-group pseudo-Gini, and the interaction pseudo-Gini, respectively, for the subgroup with the jth smallest mean income, $j = 1, 2, \ldots, k$. The relevant formulas

for computing G_{Wj}, G_{Bj}, and G_{Ij} are as follows:

$$G_{Wj} = \frac{1}{n_j} \frac{\sum_{i=1}^{n_j} r_{ij}^{*\prime} y_{ij}}{\sum_{i=1}^{n_j} y_{ij}} \tag{2}$$

$$G_{Bj} = \frac{1}{n_j} \frac{\sum_{i=1}^{n_j} \tilde{r}_{ij}^{*} y_{ij}}{\sum_{i=1}^{n_j} y_{ij}} \tag{3}$$

$$G_{Ij} = \frac{1}{n_j} \frac{\sum_{i=1}^{n_j} \hat{r}_{ij}^{*} y_{ij}}{\sum_{i=1}^{n_j} y_{ij}} \tag{4}$$

Ogwang articulates how G_{Wj}, G_{Bj}, and G_{Ij} can be conveniently estimated by setting up artificial regressions with known heteroscedastic structures for which the variance of the error term is related to the incomes of the income-receiving units in the subgroup with the jth smallest mean income.

The first, second, and third terms on the right-hand side of Equation (1) are the within-group, between-group, and interaction components, respectively. It is also apparent from the same equation that the contribution of the subgroup with the jth smallest mean income to the overall Gini index is given by

$$G_j = p_j s_j G_{Wj} + p_j s_j G_{Bj} + 2 p_j s_j G_{Ij} \tag{5}$$

The first, second, and third terms on the right-hand side of Equation (5) represent the subgroup's contribution to the within-group, between-group, and interaction components, respectively, of the overall Gini index. As will be discussed in Section 3 below, the third term on the right-hand side of Equation (5), which defines the contribution of the subgroup with the jth smallest mean income to the interaction component, is a possible measure of the extent to which the incomes of the income-receiving units in this subgroup are separated from those of the income-receiving units in all other subgroups.

We now turn our attention to the effect of increasing the incomes of all the income-receiving units in the subgroup with the jth smallest mean income by a small proportion $\delta > 0$. Potentially, this income increase could change the rankings of the income-receiving units in this subgroup relative to those of the income-receiving units in all the subgroups, thereby complicating the analysis. To circumvent this problem, it is necessary to invoke the assumption of income separation. As suggested by a referee, the income-separation assumption may be called the "rank-preserving condition." Hence, hereafter, the terms "income-separation assumption" and "rank-preserving condition" will be used interchangeably.

Given that income separation is an important assumption in the analysis of the marginal effects, it is pertinent to elaborate on what income separation entails. Strictly speaking, income separation refers to situations for which the incomes of the members of a particular subgroup are removed from those of the members of other population subgroups. Under such a separation, small proportionate changes in the incomes of all income-receiving units in the separated population subgroup do not change the rankings of these income-receiving units relative to those of the income-receiving units in all population subgroups.

In general, income separation applies to subgroup formations with some nonoverlapping income ranges. One such formation, which commonly features in official statistics releases, is by income quantiles. This subgroup formation results in the interaction component of the overall Gini index being zero. Other subgroup formations, such as poor/rich and low income/high income, also give rise to nonoverlapping income ranges. The elasticity derived in this article applies to these subgroup formations, provided that small proportionate changes in the incomes of the income-receiving units in the first subgroup do not result in some of its members becoming richer than the income-receiving units in the second subgroup.

In some empirically likely situations, income separation may only apply to a few subgroups whose incomes are far removed from those of the other subgroups. In these situations, it is appropriate to analyze the marginal effects only for the subgroup(s) whose incomes are deemed to be separated from those of other subgroups. The issue of checking for subgroup-income separation is discussed in the next section.

In fact, owing to the complications arising from the existence of an interaction component, multidimensional decompositions of the Gini index have traditionally focused on the case of subgroups with nonoverlapping income ranges (e.g., Mussard and Richard 2012). Also, the concentration-ratio-based elasticities reported by Podder (1993) and Chatterjee and Podder (2007) are strictly valid under the assumption of significant subgroup-income separation although these researchers justified their reporting of these elasticities based on the assumption that the proportional income changes are too small to have any effect on the relative ranks.

Let $G(j,0)$ and $G(j,\delta)$ denote the Gini index before increasing and after increasing, respectively, the incomes of the members of the subgroup with the jth smallest mean income by the proportion δ. It is easy to verify that under income separation, the mean income for the subgroup is increased by the same proportion δ but the within-group Ginis $\left(G_{Wj},\, j=1,2,\ldots,k\right)$, the between-group pseudo-Ginis $\left(G_{Bj},\, j=1,2,\ldots,k\right)$, and the interaction pseudo-Ginis $\left(G_{Ij},\, j=1,2,\ldots,k\right)$ are unchanged. However, the income share for the subgroup with the jth smallest mean income increases from $s_j=\frac{\sum_{i=1}^{n_j} y_{ij}}{\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}}$ to s_j+B_j where $B_j=\frac{\delta\sum_{i=1}^{n_j} y_{ij}\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}-\delta\left(\sum_{i=1}^{n_j} y_{ij}\right)^2}{\left(\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}\right)^2+\delta\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}\sum_{i=1}^{n_j} y_{ij}}$ whereas the income share for the subgroup with the rth smallest mean income, where $r\neq j$, decreases from $s_r=\frac{\sum_{i=1}^{n_r} y_{ir}}{\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}}$ to s_r+A_r where $A_r=\frac{-\delta\sum_{i=1}^{n_r} y_{ir}\sum_{i=1}^{n_j} y_{ij}}{\left(\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}\right)^2+\delta\sum_{i=1}^{n_j}\sum_{j=1}^{k} y_{ij}\sum_{i=1}^{n_j} y_{ij}}$.

Substituting the new income shares into Equation (1) taking into account the fact that the within-group Ginis, the between-group pseudo-Ginis and the interaction pseudo-Ginis are unchanged under the separability assumption, as already alluded to above, as well as the fact that $\lim(\delta\to 0)(B_j/\delta)=s_j\,(1-s_j)$ and $\lim(\delta\to 0)(A_r/\delta)=-s_r s_j$, it follows that

$$[G(j,0)]^{-1}\lim(\delta\to 0)\left(\frac{G(j,\delta)-G(j,0)}{\delta}\right)$$

$$=[G(j,0)]^{-1}\left[G_j-\sum_{i=1}^{k} s_j G_i\right]\quad=[G(j,0)]^{-1}\left[G_j-s_j\sum_{i=1}^{k} G_i\right]\qquad(6)$$

Since $\sum_{i=1}^{k} G_i = G(j,0)$, i.e., the sum of the contributions of the k population subgroups is equal to the overall Gini index, it follows that the elasticity of the Gini index with respect to changes in the incomes of all income-receiving units in the subgroup with the jth smallest mean income component is given by

$$\eta_j = \left[\{G_j/G(j,0)\} - s_j\right] \tag{7}$$

To the best of our knowledge, the elasticity formula given by Equation (7), which is clearly very simple, has not previously been featured in the literature on subgroup decomposition of the Gini index (detailed derivations of the results are available from the author).

Five important features of Equation (7) are worthy of mention. First, the elasticity is structurally similar to that of the empirically popular Lerman and Yitzhaki's (1985) elasticity for Gini income-source decomposition. Specifically, the elasticity of the Gini index with respect to small proportional changes in subgroup incomes is the difference between the ratio of the contribution of the subgroup to the overall Gini index prior to the income change (i.e., $G_j/G(j,0)$) and the income share of that subgroup prior to the change (i.e., s_j). Likewise, Lerman and Yitzhaki's elasticity with respect to small proportional changes in the incomes of all receiving units from a particular source is the difference between the ratio of the contribution of that income source to the overall Gini index prior to the income change and the share of that income source in total income prior to the change. Second, given that the ranks are unchanged under subgroup-income separability, the resulting elasticity given by Equation (7) solely reflects the share effects. Third, as the illustrative example presented in Section 4 below also indicates, the elasticity could be positive, zero, or negative depending on the magnitude of G_j, since $G(j,0)$ and s_j cannot be negative. Fourth, $\sum_{j=1}^{k} \eta_j = 0$. The zero sum of all the elasticities is an artifact of the scale-independence property of the Gini index, which guarantees that increasing the incomes of all the receiving units in all population subgroups by a constant proportion does not affect the overall Gini index. Fifth, the computations entailed are transparent and straightforward.

3. Checking for Income Separation

It is repeatedly stressed in this article that income separation is necessary for the elasticity formula as given by Equation (7) to be strictly valid. Hence, it is important to check each subgroup for income separation to determine whether it is appropriate to compute the elasticity for that subgroup. In cases where the subgroups are formed by income quantiles, as is the case with the illustrative example presented in Section 4 below, the elasticities can be reported for all quantiles since all of them would satisfy income separation.

In other cases, however, it is necessary to check individual subgroups for income separation and only report the elasticities for those subgroups that are deemed to satisfy the income-separation (or rank-preserving) condition. In this regard, the contribution of the interaction component to the overall Gini index, as defined by the third term on the right-hand side of Equation (5), provides a simple check of how segregated the incomes of the income-receiving units in this subgroup are from those of the income-receiving units of all other subgroups. Specifically, if the incomes in the subgroup with the jth smallest mean

income are characterized by significant income separation, then the third term on the right-hand side of Equation (5) should be (very close to) zero. One positive aspect of this simple income-separation check is that subgroup size, an important consideration in the development of appropriate segregation indexes, is automatically taken into account via the population shares. Clearly, the contribution of each subgroup to the interaction component for the overall Gini index is beneficial in that it provides a simple but valid check for income separation.

Yitzhaki and Lerman (1991) proposed a stratification index that indicates the extent to which the incomes in a particular subgroup are separated from those in all other subgroups. Further details about Yitzhaki and Lerman's stratification index can be found in their paper. Yitzhaki (1994) and Allanson (2014), among others, propose indexes of overall income stratification that provide useful insights into the overall extent of income separation.

Regardless of which method is used to check for income separation, the elasticities should be reported only for the individual subgroups that are deemed to satisfy the appropriate income-separation (or rank-preserving) condition.

Before presenting an illustrative example using Canadian data, it is important to state the advantages of the marginal analysis of the Gini subgroup decompositions as developed in this article over the marginal analysis based on the Shapley decomposition. First, the fact that the Shapley approach results in the interaction term for the Gini index being absorbed into the within-group or the between-group component results in loss of information on the contribution of the subgroups to the interaction component of the Gini index, information which could be used to measure income separation as described in this section. Second, from a practical perspective the elasticities developed in this article are computationally more convenient, involving simple analytical formulas as opposed to the complex algorithms entailed in the Shapley decompositions.

4. Illustrative Example Using Canadian Data

To demonstrate the computation and interpretation of the elasticity of the Gini index with respect to changes in subgroup incomes, we applied the methodology described in Section 2 to the data on the total pretax post-transfer incomes, in Canadian dollars, of a random sample of 4,883 persons, derived from the Canadian Census 2006 Public Use Microdata Files. The sample data are available from the author upon request. To ensure that the appropriate conditions for the empirical validity of the elasticity are met, we first created nonoverlapping subgroups by categorizing the incomes into quintiles.

As indicated above, an important first step in the computation of the elasticities of the Gini index with respect to changes in subgroup incomes is the computation of the contributions of the various subgroups to the overall Gini index. These contributions are reported in the second last column of Table 1. The corresponding elasticities, which are computed using Equation (7), are reported in the last column of the same table.

It is apparent from the entries in the last column of Table 1 that the contribution of the income-receiving units in each of the first two quintiles to the overall Gini index is negative, whereas the contributions of those in the other quintiles are positive. These results raise the issue of the conditions under which the contributions of a particular

Table 1. Elasticities of the Gini index with respect to subgroup-income changes.*

Quintile	Within-group Gini	Between-group and interaction pseudo-Ginis		Pop. share	Income share	Contributions of the various components to overall inequality				
						Within-group	Between-group	Interaction	Total	Elasticity
j	$\hat{G}_{Wj}$	$\hat{G}_{Bj}$	$\hat{G}_{Ij}$	p_j	s_j	$p_j s_j \hat{G}_{Wj}$	$p_j s_j \hat{G}_{Bj}$	$2p_j s_j \hat{G}_{Ij}$	$\hat{G}_j$	$\hat{\eta}_j$
First ($j = 1$)	0.5333	−4.0029	0.0030	0.2	0.0158	0.0017	−0.0126	0.0000	−0.0109	−0.0368
Second ($j = 2$)	0.1257	−2.0000	0.0001	0.2	0.0749	0.0019	−0.0300	0.0000	−0.0281	−0.1289
Third ($j = 3$)	0.0954	0.0000	0.004	0.2	0.1421	0.0027	0.0000	0.0000	0.0027	−0.1368
Fourth ($j = 4$)	0.0760	2.0000	−0.0003	0.2	0.2315	0.0035	0.0926	−0.0000	0.0961	−0.0466
Fifth ($j = 5$)	0.2928	4.0031	−0.0001	0.2	0.5357	0.0314	0.4286	−0.0000	0.4600	0.3492
Total	-	-	-	1.0	1.0000	0.0412	0.4787	0.0000	0.5198	0.0000

*All the notations are as described in Section 2. $\hat{G}_j = p_j s_j \hat{G}_{Wj} + p_j s_j \hat{G}_{Bj} + 2p_j s_j \hat{G}_{Ij}$, j = 1, 2, . . ., 5 (see Equation (2)); $\hat{G}(j,0) = \sum_{j=1}^{5} \hat{G}_j$; $\hat{\eta}_j = \left(\hat{G}_j / \hat{G}(j,0)\right) - s_j$ j = 1, 2, . . ., 5 (see Equation (7)); $\sum_{j=1}^{5} p_j = \sum_{j=1}^{5} s_j = 1$; and n = 4,883.

subgroup could be negative. To see how the contributions to the overall Gini index of the income-receiving units in the lower quintiles could be negative, we refer to Ogwang's (2014) analytical results, which indicate that the within-group Ginis are always positive whereas the between-group and interaction pseudo-Ginis can be positive or negative. Since the income shares and population shares cannot be negative, it follows from Equation (5) that the net contribution of a particular quintile could be positive or negative. In the present empirical example, it turns out that for the first two quintiles the negative effects dominate the positive effects associated with the within-group component. Podder (1993) and Chatterjee and Podder (2007) also uncovered negative contributions of some population subgroups to the overall Gini index when they applied the concentration-index-based measures to their datasets.

Our experience with this dataset and several other datasets reveals that when most of the incomes in a particular subgroup fall below the median income, then the between-group pseudo-Gini associated with that subgroup will be negative. In fact, the corresponding concentration index for this subgroup formation also turns out to be negative, as pointed out by Chatterjee and Podder (2007).

As can be expected from the traditional subgroup decompositions of the Gini index, the overall Gini index and its within-group, between-group, and interaction components are all non-negative. It also turns out that the between-group inequality accounts for the largest proportion of the overall inequality and the interaction component accounts for the least proportion.

The entries in the third from last column of Table 1 also indicate that the sum of the contributions of the five quintiles to the interaction component is zero. This zero sum of the contributions to the interaction component can be expected, given that quintile formation ensures that there is no income overlapping.

The elasticities for the first four quintiles, reported in the last column of Table 1, are negative, whereas that for the top quintile is positive. The reported elasticities indicate that a one-percent increase in the incomes of the income-receiving units in the first, second, third, and fourth quintile, controlling for the incomes of the income-receiving units in all other quintiles in each case, will decrease the overall Gini index by 0.0368 percent, 0.1289 percent, 0.1368 percent, and 0.0466 percent, respectively. Also, a one-percent increase in the incomes of the income-receiving units in the top quintile, controlling for the incomes of the income-receiving units in all other quintiles, will increase the overall Gini index by 0.3492 percent. As anticipated, the sum of the elasticities for all the five quintiles is zero.

It is apparent from the results for the third and fourth quintiles that although a particular subgroup may initially make a positive contribution to the overall Gini index, small proportionate changes in the incomes of the members of this subgroup may result in a decrease in overall income inequality. This raises the interesting issue of whether empirical researchers should be focusing more on the contributions of the various subgroups to overall inequality or on the elasticities. With respect to this issue, Jurkatis and Strehl (2014) argue in favor of focusing more strongly on the elasticities, given the policy significance. The fact that the elasticities for the lower-income quintiles are negative and that for the top-income quintile is positive should not be surprising, given that increasing the incomes of the income-receiving units in a lower quintile, controlling for

other incomes, would potentially lead to a closure of the income gap. However, in general empirical situations with subgroup-income interactions, the link between proportionate marginal subgroup-income changes and overall income inequality is convoluted.

5. Concluding Remarks

The elasticity of the Gini index with respect to changes in subgroup incomes is a statistic that aids policy makers in devising appropriate inequality-reduction strategies. Presenting data on the elasticities may also help the various stakeholders to gain valuable insights from the decomposition.

The trick in deriving this elasticity is to first isolate the contributions of the various subgroups to the overall Gini index from which the marginal effects are then established. Since the marginal income changes could potentially result in relative rank changes, which complicate the analysis, it becomes necessary to invoke the rank-preserving condition. This condition applies if there is a high degree of income separation in that the incomes of the members comprising of a particular subgroup are very different from the incomes of the members of other subgroups (e.g., they are too high or too low). Under income separation, it is apparent from Equation (7) that the elasticity is simply the difference between the ratio of the contribution of that subgroup to the overall Gini index prior to the income change and the income share of the same subgroup prior to the change.

It is also possible in principle to account for sampling variability in the elasticity statistic by computing its bootstrap or jackknife standard error, which is valuable for hypothesis-testing purposes. In this regard, accuracy considerations necessitate the re-ranking of the incomes in each bootstrap/jackknife subsample in accordance with the ranking rules described in Section 2, which is computationally intensive but should not be overly difficult if the empirical researcher has access to an appropriate ranking subroutine. Shao and Tu (1995) explain the bootstrap and jackknife methodologies in general and Ogwang (2014) explains the calculation of jackknife standard errors in the subgroup decomposition of the Gini index.

It has been mentioned in the literature (e.g., Podder 1993) that the problem of the inability to neatly decompose the Gini index into only two components (i.e., within-group component and between-group component) owing to the existence of a third component, the interaction component, is circumvented by focusing on the contributions of the various subgroups to the overall Gini index. However, it is apparent from the discussion in this article that this problem in fact resurfaces in the analysis of the elasticities unless the subgroup under consideration is characterized by significant income separation.

Finally, we hope that the results presented in this article will help to popularize the reporting of the elasticities of the Gini index with respect to changes in subgroup incomes in empirical investigations involving population subgroups that are characterized by significant income separation.

6. References

Aaberge, R., S. Bjerve, and K. Doksum. 2005. "Decomposition of Rank-Dependent Measures of Inequality by Subgroups." *Metron-International Journal of Statistics*

LXIII: 493–503. Available at: ftp://metron.sta.uniroma1.it/RePEc/articoli/si11.pdf. (accessed June 30, 2015).

Allanson, P. 2014. "Income Stratification and Between-group Inequality." *Economics Letters* 124: 227–230. Doi: http://dx.doi.org/10.1016/j.econlet.2014.05.025.

Chantreuil, F. and A. Trannoy. 2011. "Inequality Decomposition Values." *Annals of Economics and Statistics* 101/102: 13–36. Available at: http://www.jstor.org/stable/41615472 (accessed June 30, 2015).

Charpentier, A. and S. Mussard. 2011. "Income Inequality Games." *Journal of Economic Inequality* 9: 529–554. Doi: http://dx.doi.org/10.1007/s10888-011-9184-1.

Chatterjee, S. and N. Podder. 2007. "Some Ethnic Dimensions of Income Distribution from Pre- to Post-Reform New Zealand, 1984–1998." *The Economic Record* 83: 275–287. Doi: http://dx.doi.org/10.1111/j.1475-4932.2007.00414.x.

Cowell, F.A. and C.V. Fiorio. 2011. "Inequality Decompositions – A Reconciliation." *Journal of Economic Inequality* 9: 509–528. Doi: http://dx.doi.org/10.1007/s10888-011-9176-1.

Jurkatis, S. and W. Strehl. 2014. "Gini Decompositions and Gini Elasticities: On Measuring the Importance of Income Sources and Population Subgroups for Income Inequality." Discussion Papers 2014/22, Freie Universität Berlin, School of Business and Economics. Available at: http://econstor.eu/bitstream/10419/102730/1/798627875.pdf (accessed June 30, 2015).

Leibbrandt, M., C. Woolard, and I. Woolard. 2000. "The Contribution of Income Components to Income Inequality in the Rural Former Homelands of South Africa: A Decomposable Gini Analysis." *Journal of African Economies* 9: 77–99. Doi: http://dx.doi.org/10.1093/jae/9.1.79.

Lerman, R.I. and S. Yitzhaki. 1985. "Income Inequality Effects by Income Source: A New Approach and Applications to the United States." *Review of Economics and Statistics* 67: 151–156. Doi: http://dx.doi.org/10.2307/1928447.

Lopéz-Feldman, A. 2006. "Decomposing Inequality and Obtaining Marginal Effects." *The Stata Journal* 6: 106–111. Available at: http://www.stata-journal.com/article.html?article=st0100 (accessed June 30, 2015).

Mussard, S. 2004. "The Bidimensional Decomposition of the Gini Ratio. A Case Study: Italy." *Applied Economics Letters* 11: 503–505. Doi: http://dx.doi.org/10.1080/1350485042000244530.

Mussard, S. and P. Richard. 2012. "Linking Yitzhaki's and Dagum's Gini Decompositions." *Applied Economics* 44: 2997–3010. Doi: http://dx.doi.org/10.1080/00036846.2011.568410.

Mussard, S. and L. Savard. 2012. "The Gini Multi-Decomposition and the Role of Gini's Transvariation: Application to Partial Trade Liberalization in the Philippines." *Applied Economics* 44: 1235–1249. Doi: http://dx.doi.org/10.1080/00036846.2010.539540.

Mussini, M. 2013. "A Matrix Approach to the Gini Index Decomposition by Subgroup and Income Source." *Applied Economics* 13: 2457–2468. Doi: http://dx.doi.org/10.1080/00036846.2012.667553.

Ogwang, T. 2014. "A Convenient Method of Decomposing the Gini Index by Population Subgroups." *Journal of Official Statistics* 30: 91–105. Doi: http://dx.doi.org/10.2478/jos-2014-0005.

Podder, N. 1993. "A New Decomposition of the Gini Coefficient among Groups and its Interpretations with Applications to Australia." *Sankhya: The Indian Journal of Statistics* 55: 262–271. Available at: http://www.jstor.org/stable/25052790 (accessed June 30, 2015).

Radaelli, P. 2010. "On the Decomposition by Subgroups of the Gini Index and Zenga's Uniformity and Inequality Indexes." *International Statistical Review* 78: 81–101. Doi: http://dx.doi.org/10.1111/j.1751-5823.2010.00100.x.

Shao, J. and D. Tu. 1995. *Jackknife and Bootstrap*. New York: Springer.

Shorrocks, A.F. 2013. "Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value." *Journal of Economic Inequality* 11: 99–126. Doi: http://dx.doi.org/10.1007/s10888-011-9214-z.

Stark, O., J.E. Taylor, and S. Yitzhaki. 1986. "Remittances and Inequality." *Economic Journal* 96: 722–740. Available at: http://www.jstor.org/stable/2232987 (accessed June 30, 2015).

Yitzhaki, S. 1994. "Economic Distance and Overlapping of Distributions." *Journal of Econometrics* 61: 147–159. Doi: http://dx.doi.org/10.1016/0304-4076(94)90081-7.

Yitzhaki, S. and R.I. Lerman. 1991. "Income Stratification and Income Inequality." *Review of Income and Wealth* 37: 313–329. Doi: http://dx.doi.org/10.1111/j.1475-4991.1991.tb00374.x.

Received February 2015
Revised July 2015
Accepted September 2015

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 747–768, http://dx.doi.org/10.1515/JOS-2016-0038

Multivariate Beta Regression with Application in Small Area Estimation

Debora F. Souza[1] *and Fernando A. S. Moura*[2]

Multivariate beta regression models for jointly modelling two or more variables whose values belong in the (0,1) interval, such as indexes, rates or proportions, are proposed for making small area predictions. The multivariate model can help the estimation process by borrowing strength between units and obtaining more precise estimates, especially for small samples. Each response variable is assumed to have a beta distribution so the models could accommodate multivariate asymmetric data. Copula functions are used to construct the joint distribution of the dependent variables; all the marginal distributions are fixed as beta. A hierarchical beta regression model is additionally proposed with correlated random effects. We present an illustration of the proposed approach by estimating two indexes of educational attainment at school level in a Brazilian state. Our predictions are compared with separate univariate beta regressions. The inference process was conducted using a full Bayesian approach.

Key words: Bayesian inference; copula function; small domain; education evaluation.

1. Introduction

In recent years, numerous applications of the beta distribution have been developed due to the distribution's suitability for modelling rates or proportions. Its properties include being defined on the range (0,1), allowing for asymmetry present in these types of variables, and assuming different forms depending on its parameters. The beta regression additionally allows heteroscedastic observations.

Ferrari and Cribari-Neto (2004) proposed a univariate beta regression for modelling rates or proportions and used a classic approach to estimate the model parameters. A Bayesian version of the static beta regression was proposed by Branscum et al. (2007). More recently, Da-Silva et al. (2011) proposed a method for beta time series data, in which the model parameters that are related to the means follow a dynamic model. However, the most frequently proposed use of the beta distribution in the context of regression has been restricted to cases where there is only one dependent variable.

[1] Coordenação de Métodos e Qualidade, Instituto Brasileiro de Geografia e Estatística (IBGE). Rio de Janeiro, Brazil. Email: debora.souza@ibge.gov.br
[2] IM-UFRJ – Statistics Department, Rio de Janeiro, Rio de Janeiro, Brazil. Email: fmoura@im.ufrj.br
Acknowledgments: This work is part of the PhD dissertation of Debora F. S., under the supervision of Fernando Moura, in the Graduate Program of Statistics of Universidade Federal do Rio de Janeiro (UFRJ). The authors would like to thank the associate editor and the anonymous referees for their very thoughtful and constructive comments.

We propose a new approach to jointly modelling indexes, rates or proportions, commonly estimated with low accuracy in small samples. Examples of variables that are measured in the range (0,1) and are related to each other are the proportion of poor people, the mortality rate and the ratio of food expenditures to total expenditures. The models proposed in this article can also be employed in estimating correlated poverty indexes for small domains. While the motivation for this work has been the estimation of rates or proportions in small areas (or domains), the strategy used to achieve this goal can be applied to a more general context. Multivariate models are developed for modelling rates or proportions, offering the possibility of jointly managing related quantities in one single model and enjoying the benefits that this joint approach offers. Borrowing strength across the response variables in the multivariate models proposed here can provide more precise estimates of the quantities of interest.

Cepeda-Cuervo et al. (2014) apply a bivariate strategy using the Farlie-Gumbel-Morgenstern (FGM) copula, modeling the dispersion parameter of the beta regression as proposed in Smithson and Verkuilen (2006) and Simas et al. (2010). However, their approach does not account for any hierarchical structure of the population and no extension to the multivariate case is discussed. Melo et al. (2009), Fabrizi et al. (2011) and Murteira and Ramalho (2014) propose and apply multivariate models for dealing with fractional data.

This article develops multivariate regression models where the dependent variables marginally follow a beta distribution. These models address data fitting in general contexts, and the models are especially advantageous for small area estimation. The beta marginal distributions were reparametrised by the mean and the dispersion, as in Ferrari and Cribari-Neto (2004). The associations between the response variables are considered as a copula function applied to the marginal densities. Copulas are useful tools for building multivariate distributions where the marginal distributions are given or known, allowing individual models be analysed together. Additionally, copula functions allow the representation of various types of dependence between variables. The use of copulas allows flexibility in handling nonlinear relationships between the response variables and is therefore a more general setup than the multivariate normal distribution, which allows only linear relationships. For a complete study on the copula function and its utilities in statistics, see Nelsen (2006).

Two types of multivariate models with beta responses are proposed: a beta regression model, where the marginal densities are connected by a copula function, and a hierarchical beta model with correlation between their means. In a small area estimation context where auxiliary variables and data from multiple characteristics are available, these models can improve the prediction of observations and target-population parameters. Several authors argue that this approach provides better estimates than fitting separate univariate models, because a multivariate model considers the correlations between the response variables after conditioning on the auxiliary variables. Fay (1987) modelled the joint behaviour of the median income in households of three, four and five dwellers. Datta et al. (1999) applied a multivariate mixed linear model and concluded from a simulation study that the multivariate approach provides better results than setting a separate model for each variable. The methods most commonly employed are based on borrowing information from neighbouring or related areas. The models proposed in this article have a direct

application to the small area estimation problem by additionally allowing strength to be borrowed between the response variables.

The article is organised as follows. In Section 2, we propose a multivariate beta regression model by employing copula functions. In Section 3, we apply our proposed models to the small area estimation problems, presenting an illustration with Brazilian education data. Section 4 offers some conclusions and suggestions for further research.

2. Multivariate Beta Regression Model Based on Copulas

The structure of dependence between two or more related response variables can be defined in terms of their joint distribution. One way of obtaining a multivariate beta distribution is to join the univariate beta using copula functions, which is one of the most useful tools when the marginal distributions are given or known. The use of copula functions enables the representation of various types of dependence between variables. In practice, this function implies a more flexible assumption about the form of the joint distribution than that given in Olkin and Liu (2003), which assumes that the marginal distributions have the same parameter. Nelsen (2006) defines a copula as a joint distribution function

$$C(u_1, \ldots, u_K) = P(U_1 \leq u_1, \ldots, U_K \leq u_K), \quad 0 \leq u_j \leq 1,$$

where $U_j, j = 1, \ldots, K$ are uniformly distributed on the interval (0,1).

Sklar's theorem, stated here in Theorem 1, shows how to obtain a joint distribution using a copula.

Theorem 1 *Let H be a K-dimensional distribution function with marginal distribution functions $F_1, \ldots, F_K$. Then, there is a K-dimensional copula C such that for all $(y_1, \ldots, y_K) \in [-\infty, \infty]^K$,*

$$H(y_1, \ldots, y_K) = C(F_1(y_1), \ldots, F_k(y_K)). \tag{1}$$

Conversely, if C is an n-dimensional copula and $F_1, \ldots, F_K$ are cumulative distribution functions, then the function H defined by (1) is a distribution function with marginal distributions $F_1, \ldots, F_K$. Moreover, if all marginal distributions are continuous, C is unique. Otherwise, the copula C is uniquely determined in $Im(F_1) \times \ldots \times Im(F_K)$, where $Im(\cdot)$ represents the image of $(\cdot)$.

Let $\mathbf{y} = ((y_{11}, \ldots, y_{1K}), \ldots, (y_{n1}, \ldots, y_{nK}))$ be a random sample of size n from a continuous joint distribution with marginal densities $f_1, \ldots, f_K$. Thus, the likelihood function is given by:

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} c(F_1(y_{i1}|\boldsymbol{\Psi}), \ldots, F_K(y_{iK}|\boldsymbol{\Psi}))f_1(y_{i1}|\boldsymbol{\Psi}) \ldots f_K(y_{iK}|\boldsymbol{\Psi}) \tag{2}$$

where $\boldsymbol{\Psi}$ denotes the set of parameters that define the distribution functions F_k, the densities f_k, and the copula-density function $c(\cdot)$, $k = 1, \ldots, K$.

In (2), we assume that each response variable k is beta distributed, such that:

$$Y_{ik}|\mu_{ik}, \phi_k \sim Beta(\mu_{ik}, \phi_k), \quad i = 1, \ldots, n, \quad k = 1, \ldots, K$$

$$g(\mu_{ik}) = \eta_{ik} = \sum_{j=1}^{p_k} x_{ij}\beta_{jk}$$

where $Beta(\mu_{ik}, \phi_k)$ denotes that Y_{ik} is beta distributed with mean μ_{ik} and variance $\frac{\mu_{ik}(1-\mu_{ik})}{1+\phi_k}$, $g(\cdot)$ is the link function and p_k is the number of covariates for the response variable k.

Denote by $BetaM(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\theta})$ the multivariate beta distribution obtained by using K marginally beta-distributed variables with parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \phi_K)^T$ and a copula function with a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T$. Thus, the structure of dependence between the K beta responses is defined by their joint distribution, which is obtained by applying a copula function, resulting in the likelihood function given by (2). Under the Bayesian approach, the specification of the model is completed by assigning a prior distribution to $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$, the parameter $\boldsymbol{\beta} = \{\beta_{jk} : j = 1, \ldots, p_k; k = 1, \ldots, K\}$ and the parameters that define the copula family. Souza (2011) developed and fitted Model (2) using different copulas to predict missing response values. It was also carried out a simulation study to compare bivariate and univariate beta models under different scenarios.

2.1. *Multivariate Hierarchical Beta Regression Model*

In the multivariate beta regression model presented in the previous section, the marginal beta regression coefficients were fixed. However, there are situations in which some or all of the coefficients are assumed to be random. In these cases, the coefficients of each observation have a common average, suffering from the influence of nonobservable effects. Such models are often called mixed-effects models and have applications in several areas. Jiang (2007) discusses linear mixed models and some inference procedures for estimating their parameters. Rao and Molina (2015) shows some use of mixed-effects models in small area estimation.

In this section, we propose a generalisation of the multivariate regression model presented in Section 2 by assuming that some or all of the coefficients associated with the linear predictor of each response variable can be random and correlated.

Let y_{idk} be the observed value of the i^{th} microunit within the d^{th} macrounit for the k^{th} response variable, $i = 1, \ldots, n_d$, $d = 1, \ldots, D$ and $k = 1, \ldots, K$. Furthermore, let us assume that y_{idk} and $y_{i'dk}$ are conditionally independent, $\forall i \neq i'$. The multivariate hierarchical beta regression model is defined as

$$\mathbf{y}_{id}|\boldsymbol{\mu}_{id}, \boldsymbol{\phi}_{id}, \boldsymbol{\theta} \sim BetaM(\boldsymbol{\mu}_{id}, \boldsymbol{\phi}_{id}, \boldsymbol{\theta}), \quad i = 1, \ldots, n_d, \quad d = 1, \ldots, D \tag{3}$$

$$y_{idk}|\mu_{idk}, \phi_{idk} \sim Beta(\mu_{idk}, \phi_{idk}), \quad k = 1, \ldots, K \tag{4}$$

$$g(\mu_{idk}) = \sum_{j=1}^{p_k} x_{idjk}(\beta_{jk} + \nu_{djk}) \tag{5}$$

$$\nu_{djk} \sim N\left(0, \sigma_{jk}^2\right), \quad j = 1, \ldots, p_k \quad \text{and} \quad k = 1, \ldots, K \tag{6}$$

where: p_k is the number of covariates for the response variable k; $BetaM(\boldsymbol{\mu}_{id}, \boldsymbol{\phi}_{id}, \boldsymbol{\theta})$ denotes a multivariate beta distribution using a copula function with parameter $\boldsymbol{\theta}$ and the beta marginal distributions for the i^{th} microunit belonging to the macrounit d; $\mathbf{y}_{id} = (y_{id1}, \ldots, y_{idK})^T$; $\boldsymbol{\mu}_{id} = (\mu_{id1}, \ldots, \mu_{idK})$; $\boldsymbol{\phi}_{id} = (\phi_{id1}, \ldots, \phi_{idK})$; $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{p_k k})$; $\mathbf{x}_{idk} = (x_{id1k}, \ldots, x_{idp_k k})^T$ and

$$\mathbf{x}_{dk}^T = \begin{pmatrix} x_{1d1k} & \cdots & x_{1dp_k k} \\ x_{2d1k} & \cdots & x_{2dp_k k} \\ \vdots & \cdots & \vdots \\ x_{N_d d1k} & \cdots & x_{N_d dp_k k} \end{pmatrix}.$$

Thus, microunits belonging to the same macrounit have the same coefficient and the coefficients are different between macrounits. Each response variable can have its own set of regressors and these are not necessarily the same.

As generally described in Equations (3) and (5), the model allows all regression coefficients to be random. However, in many applications of hierarchical models, only some coefficients are assumed to be random, specifically the intercept term. To allow fixed and random coefficients, Equation (5) can be changed to

$$g(\boldsymbol{\mu}_{idk}) = \sum_{j=1}^{p_k} x_{idjk}\beta_{jk} + \sum_{j=1}^{p_k} z_{idjk}\nu_{djk} = \mathbf{x}_{idk}^T\boldsymbol{\beta}_k + \mathbf{z}_{idk}^T\boldsymbol{\nu}_{dk},$$

with $\mathbf{z}_{idk} = (z_{id1k}, \ldots, z_{idp_k k})^T$ and $\boldsymbol{\nu}_{dk} = (\nu_{d1k}, \ldots, \nu_{dp_k k})^T$. If $z_{idjk} = x_{idjk}$, the j^{th} coefficient is random and if $z_{idjk} = 0$, the correspondent coefficient is fixed.

In the model described in Equations (3)–(6) all random effects in $\boldsymbol{\nu}$ could be considered independent, and only the correlations across the response variables would be modelled. However, to allow the averages of the responses to borrow strength across themselves for a given macrolevel d, all random coefficients for a same covariate j can be assumed to be correlated. For example, if all covariates are the same for all response models, we have $\boldsymbol{\nu}_{dj} = (\nu_{dj1}, \ldots, \nu_{djK})^T \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}_j)$, $j = 1, \ldots, p$ where

$$\boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma_{j1}^2 & \sigma_{j12} & \cdots & \sigma_{j1K} \\ \sigma_{j12} & \sigma_{j2}^2 & \cdots & \sigma_{j2K} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{j1K} & \sigma_{j2K} & \cdots & \sigma_{jK}^2 \end{pmatrix}.$$

A special case very often used in practice is to assume that only the intercepts are correlated, i.e., $\boldsymbol{\nu}_{d1} = (\nu_{d11}, \ldots, \nu_{d1K})^T \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}_1)$.

The dependence of the response variables is evident on two levels: the observations and the linear predictors. This dependence can be favourable for this model with respect to the small area estimation problem because it allows strength to be borrowed across the means, which are interpreted as the true values of indexes, rates or proportions of interest. The logistic link function was used in all applications. The model stated in Equations (3)–(6)

assumes that information about K response variables and D macrounits with n_d microunits $d = 1, \ldots, D$ are available.

Equation (5) relates the averages of the response variables in each d^{th} macrounit, and considers specific macrounits' effects. Thus, the mean μ_{idk} and $\mu_{idk'}$ additionally borrow strength among themselves because they are correlated. This is particularly important in the small area estimation problem, in which μ_{idk} is interpreted as the true value of the rate or proportion of interest and information from related quantities can produce more accurate estimators.

The vector parameter $\boldsymbol{\phi}_{id}$ is modelled as presented in the next section. The way it is modelled depends on the specific application considered and it might be subject to restrictions.

3. An Example of Small Area Estimation

The models defined in Section 2 were developed for general applications where there are K related variables, measured in the range (0,1), which can be explained by covariates. Here, we present an example of small area estimation.

The researcher may be interested in estimating functions of the response variables for small domains or for some domains with no sample at all. The multivariate models proposed in the previous section can be applied to make predictions on the nonsampled domains and to produce more accurate estimates for the small domains. Auxiliary information (covariates) must be known for all units at the level being predicted. The information can be obtained from a census or administrative records. We have not considered the case of missing values in explanatory variables.

3.1. Brazilian Educational Data

The Brazilian evaluation of basic education is conducted by the Brazilian National Institute of Education Research (INEP). The evaluation measures the performance of students of the 4th and the 8th series of elementary school. The tests are performed every two years in urban state schools with more than 20 students. The evaluation of Brazilian education combines performances in the Portuguese language and mathematics tests with socioeconomic information.

The hierarchical structure of the data, organised into municipalities and schools, suggested the use of hierarchical modelling. Only schools with students in the 4th series in Rio de Janeiro State were considered in our application.

We considered the whole data of Rio de Janeiro State as our population and in each municipality selected a two-stage simple random sample of schools and students. In fact, we know the score values of all students for all schools. However, we pretend that we only know the sample-school means and sample-school variances for the selected schools and their respective sample sizes. This is not a unrealistic illustration, because information at individual level is not usually available due to issues of confidentiality.

The response variables are respectively the averages of proportions of correct answers in Portuguese and mathematics estimated at school level. In this application, these averages of proportions in both disciplines for each selected school are direct estimates based on a sample of students in each selected school.

It is important to note that although the number of correct answers for each student can be assumed to be binomially distributed, the school total cannot. Therefore a logistic model is not feasible here, since we are supposing that data at student level are not available. We further assume that the proportion in each school can be approximated by a beta distribution. This is not a strong assumption because the number of students in each school is not too small.

The main aim is to estimate these indexes for the nonsampled schools and to reduce the errors for the sampled schools. A two-part, multivariate hierarchical beta model was applied. One part relates the direct estimates of schools' proficiency to model parameters, and the other part relates these parameters to the auxiliary variables. The schools' indexes are in the (0,1) interval because the school averages are neither zero nor one for both tests. It is assumed that there is information for all schools, selected or not, on the following chosen covariates: existence of a program to avoid school dropout (x_2); lack of books for students (x_3); the percentage of teachers who teach less than 60% of the program of their disciplines (x_4); proportion of teachers in the school with lower wages (x_5); and lack of a library in the school (x_6). The variable (x_1) refers to the intercept.

The information available about the characteristics of schools is provided by the questionnaires given to school directors and teachers. Schools where there were no answers for at least one of these questionnaires were excluded from the analysis. Municipalities where there was only one state school, after the first mentioned dropouts were additionally eliminated, leaving 82 municipalities. For each one of these 82 municipalities, a random sample of 20% of the schools was selected. In eleven municipalities, all schools were selected. From the total 1,787 schools in the 82 municipalities, only 421 were selected. Within each selected school, a sample of 20% of the students was selected.

The response variables contain sampling error that may be related to the school sample size. To consider this feature, a modification in the multivariate hierarchical model is proposed in the equation of the observations. This modification was proposed by Liu et al. (2014) for a univariate beta model. Because it is natural to assume that the variance of the estimate increases when the sample size decreases, the following two-level model (3)–(6) is proposed:

$$y_{idk} \sim Beta(\mu_{idk}, \phi_{idk}),$$

where y_{idk} is the direct estimate (based on the sampling design) of the expected index of proficiency of the discipline k, of the i^{th} school in the d^{th} municipality for $i = 1, \ldots, n_d$, $d = 1, \ldots, D$, where n_d is the number of selected schools for the d^{th} municipality.

We assume that the parameter ϕ_{idk} can be different for each sampled school, and its value depends on the sample size through the following function: $\phi_{idk} = \gamma_k n_{id} - 1$, where γ_k is a unknown fixed parameter which may vary with the k^{th} component of the response vector, $k = 1, \ldots, K$ and n_{id} is the sample size of the i^{th} school in the d^{th} municipality. This assumption for ϕ_{idk} is valid only for the sampled schools. For the nonsampled ones we constructed the estimator after inferring about the parameter; see Subsubsection 3.1.2 for details.

For the condition $\phi_{idk} > 0$ to be satisfied, we must have

$$\gamma_k > max\{1/n_{id}, \forall (i,d) \in s\} \quad (7)$$

where $max\{1/n_{id}, \forall(i,d) \in s\}$ denotes the maximum of the inverses of all school sample sizes. Note that γ_k^{-1} can be interpreted as the design effect (*deff*) with respect to the variance of the sample proportion obtained in a simple random sampling with negligible sampling fraction. Therefore, if we have a previous estimate or guess of the *deff* for each response $k = 1, \ldots, K$, we can use it to set the γ_ks. However, even if this information is not available, we can still obtain estimates of γ_ks through the model.

Taking into account the inequality (7), one should impose the following constraints on the range of γ_ks prior, based on one's prior knowledge about the signal of the intraclass correlation ρ_k for each variable of interest $k = 1, \ldots, K$:

a) if $\rho_k > 0 \rightarrow max\{1/n_{id}, \forall(i,d) \in s\} < \gamma_k < 1$;
b) if $\rho_k < 0 \rightarrow \gamma_k > 1$;
c) if one is not sure about the sign of ρ_k then $\gamma_k > max\{1/n_{id}, \forall(i,d) \in s\}$.

A simple type of prior that can be assigned to the γ_k, $k = 1, \ldots, K$ are independent uniform priors, with ranges obtained as advised above.

The following models were considered in our analysis of the school data:

Model A

$$\mathbf{y}_{idk}|\mu_{idk}, \phi_{idk} \sim Beta(\mu_{idk}, \phi_{idk}), \quad i = 1, \ldots, n_d, \quad d = 1, \ldots, D$$

$$g(\mu_{id1}) = \beta_{11} + x_{id2}\beta_{21} + x_{id3}\beta_{31} + x_{id4}\beta_{41} + x_{id5}\beta_{51} + x_{id6}\beta_{61} + \nu_{d11}$$

$$g(\mu_{id2}) = \beta_{12} + x_{id2}\beta_{22} + x_{id3}\beta_{32} + x_{id4}\beta_{42} + x_{id5}\beta_{52} + \nu_{d12}$$

$$\nu_{d1} = (\nu_{d11}, \nu_{d12})^T \sim N_2(\mathbf{0}, \mathbf{\Sigma}),$$

Model B

$$\mathbf{y}_{id}|\boldsymbol{\mu}_{id}, \boldsymbol{\phi}_{id}, \theta \sim BetaM(\boldsymbol{\mu}_{id}, \boldsymbol{\phi}_{id}, \theta), \quad i = 1, \ldots, n_d, \quad d = 1, \ldots, D$$

$$\mathbf{y}_{idk}|\mu_{idk}, \phi_{idk} \sim Beta(\mu_{idk}, \phi_{idk}),$$

$$g(\mu_{id1}) = \beta_{11} + x_{id2}\beta_{21} + x_{id3}\beta_{31} + x_{id4}\beta_{41} + x_{id5}\beta_{51} + x_{id6}\beta_{61} + \nu_{d11}$$

$$g(\mu_{id2}) = \beta_{12} + x_{id2}\beta_{22} + x_{id3}\beta_{32} + x_{id4}\beta_{42} + x_{id5}\beta_{52} + \nu_{d12}$$

$$\nu_{d1} = (\nu_{d11}, \nu_{d12})^T \sim N_2(\mathbf{0}, \mathbf{\Sigma}),$$

where only the intercepts are assumed to be random.

A preliminary analysis showed that the covariate x_6 ("lack of library in the school") is not statistically significant as a predictor of the index of proficiency in mathematics in the presence of the other covariates. Therefore, we did not use it as a predictor of the second response variable in all models.

Note that Model A generates conditional correlation between the dependent variables given municipality d, as long as Σ is not diagonal. However, Model B is much more general and useful for small area estimation purposes than Model A, since it allows the dependent variables to be correlated, conditional on the true small area parameters $\boldsymbol{\mu}_{id}$, $\boldsymbol{\phi}_{id}$ and θ. This is equivalent to assuming that the sampling errors of the respective direct estimators are correlated. Furthermore, at first we would think that the use of a suitable copula function makes it possible to assume Σ diagonal in Model B; however, a drawback of adopting this strategy is that this does not allow the municipality random effects to be correlated across the dependent variables.

In the small area context, Models A and B can be regarded neither as a unit-level model, because the response variables are direct estimators, nor as an area-level model, because the municipality random effects are not of the same level as the domains of interest (schools). Since our model can be considered a two-level generalised hierarchical model, the only input response variables required to estimate its model parameters are the design-based direct estimates. Nevertheless, an extension of the model proposed here should include the designed-based variance-covariance matrix as additional information.

Because the scores of all the students are available in the Brazilian microdata test, it is possible to calculate the true observed proportions of the selected schools and to compare them with the direct estimates and the estimates provided by the models.

It is possible to obtain various types of dependence with copula functions. However, there is a wide variety of copula functions. The question thus arises of which copula to use. It makes sense to use the copula that is most appropriate for the data. Silva and Lopes (2008) and Huard et al. (2006) presented proposals for the selection of copulas and models. The criterion proposed by Huard et al. (2006) seeks the most appropriate copula for the data within a previously established set of copulas. Silva and Lopes (2008) implemented the Deviance Information Criterion (*DIC*) found in Spiegelhalter et al. (2002) and others. This criterion examines the model globally, providing not only the choice of the copula, but also the regressors and the marginal distributions of the response variables. The criteria Akaike Information Criterion (*AIC*), in Akaike (1973) and Bayesian Information Criterion (*BIC*), in Schwarz (1978), play a similar role.

Let $L(\mathbf{y}|\boldsymbol{\Psi}_j, M_j)$ be the likelihood function for the model M_j, where $\boldsymbol{\Psi}_j$ contains the copula parameters and those related to the marginal distributions. Define $D(\boldsymbol{\Psi}_j) = -2\log L(\mathbf{y}|\boldsymbol{\Psi}_j, M_j)$. The *AIC*, *BIC* and *DIC* are given by:

$$AIC(M_j) = D(E[\boldsymbol{\Psi}_j|\mathbf{y}, M_j]) + 2q_j;$$

$$BIC(M_j) = D(E[\boldsymbol{\Psi}_j|\mathbf{y}, M_j]) + \log(n)q_j;$$

$$DIC(M_j) = 2E[D(\boldsymbol{\Psi}_j)|\mathbf{y}, M_j] - D(E[\boldsymbol{\Psi}_j|\mathbf{y}, M_j])$$

where q_j denotes the number of parameters of the model M_j.

Let $\left\{\boldsymbol{\Psi}_j^{(1)}, \ldots, \boldsymbol{\Psi}_j^{(T)}\right\}$ be a sample from the posterior distribution obtained via MCMC. Then we have the following Monte Carlo approximations:

$$E[D(\boldsymbol{\Psi}_j)|\mathbf{y}, M_j] \approx T^{-1}\sum_{t=1}^{S} D\left(\boldsymbol{\Psi}_j^{(t)}\right) \quad \text{and} \quad E[\boldsymbol{\Psi}_j|\mathbf{y}, M_j] \approx T^{-1}\sum_{t=1}^{T} \boldsymbol{\Psi}_j^{(t)}.$$

The linear correlation coefficient is not suitable for measuring the dependence between variables in a model involving copulas since it is not invariant under monotone nonlinear transformation. A further appropriate measure, which can be found in Nelsen (2006), is the Kendall's τ statistic, given by

$$\tau = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1.$$

The parameter θ has different interpretations, as well as different ranges depending on the copula, as can be seen in Table 1. As the parameter θ can be written in terms of the Kendall's τ, it is possible to compare the correlations given by different copulas. Thus, the FGM copula is useful for weak association levels between $-2/9$ and 2/9. On the other hand, the Clayton copula considers only positive correlations. Another advantage of considering τ is that this facilitates the task of assigning a prior to θ.

In the following, we focus on the bivariate case. We use the copulas described in Table 1, where the ranges of variation of copula parameters θ and the measures of dependence Kendall's τ are presented.

In the following section, the inference process on the parameters of Model B and the indirect estimators of the sampled and nonsampled areas are presented. The inference process and the estimators are analogous for Model A.

3.1.1. Inference

We assume that sample selection bias is absent from both models, that is, the sampling scheme is noninformative, see Pfeffermann et al. (2006) for further details. Let $\mathbf{y}_s$ be the matrix of the response variables for the sampled schools and $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$. The posterior density for Model B of all unknown quantities is given by:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, \boldsymbol{\nu}, \mathbf{W}|\mathbf{y}_s) \propto p(\mathbf{y}_s|\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, \boldsymbol{\nu}, \mathbf{W}) \times p(\boldsymbol{\nu}|\mathbf{W})p(\boldsymbol{\beta})p(\boldsymbol{\gamma})p(\theta)p(\mathbf{W}).$$

Assuming independent priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, θ and $\mathbf{W}$, we have:

$$p(\mathbf{y}_s|\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta, \boldsymbol{\nu}, \mathbf{W}) = \prod_{d=1}^{D}\prod_{i=1}^{n_d} c(F_1(y_{id1}), \ldots, F_K(y_{idK})|\boldsymbol{\nu}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$$

$$\times \prod_{k=1}^{K} f_k(y_{idk}|\boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \nu_{d1k})$$

Table 1. Copula Functions used in this article.

Copula	$C(u,v\|\theta)$	θ	τ
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$(0,\infty)$	$[0,1]\backslash\{0\}$
FGM	$uv[1+\theta(1-u)(1-v)]$	$[-1,1]$	$[-2/9, 2/9]$
Frank	$-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$(-\infty,\infty)\backslash\{0\}$	$[-1,1]\backslash\{0\}$
Gaussian	$\int_{-\infty}^{\Phi^{-1}(u)}\int_{-\infty}^{\Phi^{-1}(v)}\frac{1}{2\pi\sqrt{1-\theta^2}}\exp\{\frac{2\theta st - s^2 - t^2}{2(1-\theta^2)}dsdt\}$	$[-1,1]$	$\frac{2}{\pi}\text{arcsen}\,\theta$
Gumbel	$\exp\{-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}\}$	$[1,\infty)$	$[0,1]$

and

$$p(\boldsymbol{\nu}|\mathbf{W}) = \prod_{d=1}^{D} p(\boldsymbol{\nu}_{d1}|\mathbf{W}) \propto \prod_{d=1}^{D} |\mathbf{W}|^{1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{\nu}_{d1}^T \mathbf{W} \boldsymbol{\nu}_{d1} \right\}$$

$$\propto |\mathbf{W}|^{D/2} \exp\left\{ -\frac{1}{2} \sum_{d=1}^{D} tr\left(\boldsymbol{\nu}_{d1}^T \boldsymbol{\nu}_{d1} \mathbf{W} \right) \right\}$$

$$\propto |\mathbf{W}|^{D/2} \exp\left\{ -\frac{1}{2} tr\left[\left(\sum_{d=1}^{D} \boldsymbol{\nu}_{d1}^T \boldsymbol{\nu}_{d1} \right) \mathbf{W} \right] \right\},$$

with $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, $\boldsymbol{\nu} = (\nu_{111}, \nu_{112}, \ldots, \nu_{D11}, \nu_{D12})$, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}, \beta_{51}, \beta_{61})$ and $\boldsymbol{\beta}_2 = (\beta_{12}, \beta_{22}, \beta_{32}, \beta_{42}, \beta_{52})$. The cumulative distribution function and the density of the beta distribution for the response variable k is represented by F_k and f_k, respectively. In addition, $c(.)$ is the density of the copula function.

The posterior distribution of all unknown parameters has no closed form, and thus a Monte Carlo Markov Chain (MCMC) simulation can be applied. Assigning a Wishart prior to $\mathbf{W}$ and a normal one to the intercepts (β_{11}, β_{12}) provides a full conditional with known forms for them. Therefore, we can use Gibbs to sample from these parameters. The other parameters are sampled via the Metropolis-Hastings algorithm (Gamerman and Lopes 2006). Samples of the posterior distribution of τ are obtained directly from samples of the posterior of θ, since τ is a function of θ.

Souza (2011) discussed different strategies for sampling from the posterior when the random-effect model described in (3)-(6) is fitted, including slice sampling (Neal, 2003). The importance of the posterior parametrisation to the convergence of MCMC algorithm when this model is fitted is shown.

To illustrate the convergence process, Souza (2011) simulated data from the model

$$\begin{aligned}
\mathbf{y}_{id} &\sim BetaM(\boldsymbol{\mu}_{id}, \boldsymbol{\phi}, \theta), \quad i = 1, \ldots, n_d, \quad d = 1, \ldots, D \\
\boldsymbol{y}_{idk} &\sim Beta(\mu_{idk}, \phi_k) \\
g(\mu_{idk}) &= \beta_{1k} + x_{id2k}\beta_{2k} + \nu_{d1k} \\
\boldsymbol{\nu}_{d1} &= (\nu_{d11}, \nu_{d12}) \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})
\end{aligned}$$

where *BetaM* represents the distribution generated by the Farlie-Gumbel-Morgenstern (FGM) copula with beta marginals. Souza (2011) fixed $(K = 2)$ response variables, with $D = 100$ domains and $n_d = 20$ units in each one. The following priors were considered: $\theta \sim U(-1, 1)$; $\beta_{jk} \sim N(0, 10^{-6})$, $j = 1, 2$; $\phi_k \sim Gamma(0.001, 0.001)$, $k = 1, 2$; and $\mathbf{W} = \boldsymbol{\Sigma}^{-1} \sim Wishart(2, \mathbf{I}_2)$, where $\mathbf{I}_2$ is the identity matrix of order 2. Souza (2011) used the same covariate for both responses. Note that $\lambda_{djk} = \beta_{1k} + \nu_{d1k}$, which is equivalent to $\lambda_{djk} \sim N(\beta_{1k}, \sigma_{1k}^2)$. A simulation study carried out under both ways of parametrisation showed that for the same number of iterations, the convergence is reached faster when the centre parametrisation is considered, that is $\lambda_{djk} \sim N(\beta_{1k}, \sigma_{1k}^2)$. For further theoretical

discussion of how to create strategies for improving MCMC convergence, see Gilks and Roberts (1996).

Assigning a Wishart prior to $\boldsymbol{\Sigma}$ is convenient because the full conditional distribution of $\boldsymbol{\Sigma}$ is known and has close form, which allows the Gibbs sampling algorithm to be employed to sample from it. However, other parameterisations of the matrix $\boldsymbol{\Sigma}$ can be considered. One simple way of decomposing $\boldsymbol{\Sigma}$ is as follows:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

Another well-known parameterisation is the spectral decomposition. These decompositions of $\boldsymbol{\Sigma}$ facilitate the elicitation of the prior. The disadvantage is that the full conditional of the parameters no longer have any close form. Souza (2011) analysed two different ways of assigning vague prior distributions to the variances parameters: $\sigma_k^{-2} \sim Gamma(\epsilon, \epsilon)$, for $k = 1, \ldots, K$, setting ϵ small; and one of the approaches proposed by Gelman (2006), $\sigma_k \sim U(0, M)$ for $k = 1, \ldots, K$, fixing M large. For ρ_{12}, a uniform prior on the interval $(-1, 1)$ is assigned. Souza (2011) showed that the slice-sampling algorithm is efficient for sampling from the full conditional of the σ_ks, but very slow for sampling from the full conditional of the random effects when a uniform distribution is assigned to the σ_ks. The Metropolis-Hastings algorithm was employed when the gamma prior was adopted for the variances. The simulation studies showed that the posterior distribution of all model parameters does not depend much on the three different ways of assigning prior distributions to $\boldsymbol{\Sigma}$.

3.1.2. Small Area Estimation

The posterior mean of μ_{idk} and its posterior variance can be empirically evaluated by calculating the mean and the variance of the T iterations of the MCMC algorithm. These values are obtained by jointly simulating the pairs $\left(\mu_{id1}^{(t)}, \mu_{id2}^{(t)}\right)$ using the fitted model, where: $\mu_{idk}^{(t)} = g^{-1}\left(\beta_{1k}^{(t)} + \nu_{d1k}^{(t)} + \sum_{j=2}^{p_k} \beta_{jk}^{(l)} x_{idjk}\right)$ for $t = 1, .., T$, $k = 1, 2$, $i = 1, \ldots, n_d$ and $d = 1, \ldots, D$.

In addition, it is necessary to define the estimators for the nonsampled schools. Because there is information on the auxiliary variables for these schools, the estimate of the expected index in each nonselected school at each (t) sample point of the posterior distribution is given by $\mu_{idk}^{(t)} = g^{-1}\left(\beta_{1k}^{(t)} + \nu_{d1k}^{(t)} + \sum_{j=2}^{p_k} \beta_{jk}^{(t)} x_{idjk}\right)$; for $t = 1, \ldots, T$, $k = 1, 2$, $i = n_d + 1, \ldots, N_d$ and $d = 1, \ldots, D$, where N_d is the population size in d^{th} domain. Because we have T sample points from the posterior distribution of μ_{idk}, we can obtain credibility intervals for the quantities of interest μ_{idk} for sampled and non-sampled schools.

It is also possible to make inference about the students' score means at municipality level. Let us assume that N_{id}, the number of students in each school $i = 1, \ldots, N_d$ for each municipality $d = 1, \ldots, D$, is known. Then we can generate MCMC samples from

the posterior distribution of the municipality score means, (μ_{d1}, μ_{d2}), as

$$\mu_{dk}^{(t)} = \frac{\sum_{i=1}^{N_d} N_{id}\mu_{idk}^{(t)}}{\sum_{i=1}^{N_d} N_{id}}; \quad t = 1, \ldots, T, \quad k = 1, 2, \quad \textit{and} \quad d = 1, \ldots, D \tag{8}$$

Posterior means and their respective posterior variances for each municipality can be obtained easily by calculating the means and the variances of T MCMC samples.

3.1.3. Some Results: Study 1

We fitted Models A and B, using the copulas listed in Table 1, to the data. We assigned relatively vague priors to all parameters of the two fitted models. In particular, with respect to the γ_ks parameters, we assigned independent uniform priors with the lower limit of their intervals given by inequality (7) and the upper limit equal to 10^5, i.e $\gamma_k \sim U(0.5, 10^5)$, $k = 1, 2$. For the other parameters, we set $\boldsymbol{\beta}_{jk} \sim N(0, 10^6)$, $j = 1, \ldots, 6$, $\mathbf{W} \sim \textit{Wishart}(2, \boldsymbol{I}_2)$, where $\boldsymbol{I}_2$ is the identity matrix of order 2. For the parameter θ the respective prior was set according to the copula as follows: $\textit{Gamma}(0.001, 0.001)$ for Clayton; $U(-1, 1)$ for FGM; $N(0, 10^4)$ for Frank; $U(-1, 1)$ for Gaussian; and $U(1, 10000)$ for Gumbel.

In all cases, two parallel chains were generated, each one with 200,000 iterations and a burn-in of 100,000. For all models, the chains were obtained from developing a special code in *Ox* version 5.0 (Doornik 2007). The corresponding two-univariate hierarchical beta model, simply denoted as "Separated" in Table 2, was also adjusted to investigate the benefits of the multivariate framework. Table 2 presents the model selection criteria results for all fitted models. The lower the DIC, AIC and BIC values, the better is the model. As noted in Table 2, the lowest values of the criteria are obtained when Model B is fitted using the Frank and Gaussian copulas. The values are slightly lower for the Frank copula. It should be noted that these copulas allow the widest range for the correlation between the indexes.

Table 3 shows the model parameter estimates for the Frank and Gaussian copulas. The posterior mean for the parameter ρ, which represents the correlation between the random effects, is approximately 0.60 for both models, as well as the Kendall's τ coefficient. These values indicate that a multivariate approach should be considered in the analysis of the students' performances.

Table 2. Model selection criteria.

Models	pD	DIC	AIC	BIC	Log-likelihood
Model A	104.03	− 1809.37	− 1989.42	− 1932.83	956.70
Model B1 – Clayton	117.78	− 2137.79	− 2345.34	− 2288.74	1127.78
Model B2 – Fgm	108.85	− 2041.43	− 2231.14	− 2174.54	1075.14
Model B3 – Frank	119.48	− 2237.28	− 2448.24	− 2391.65	1178.38
Model B4 – Gaussian	112.94	− 2239.19	− 2437.07	− 2380.47	1176.06
Model B5 – Gumbel	117.60	− 2213.52	− 2420.73	− 2364.13	1165.56
Separated	69.08	− 1802.51	− 1910.68	− 1850.04	935.80

Table 3. Summary of the model parameters' posterior distribution for the Gaussian and Frank copulas.

	Frank					Gaussian				
Par.	2.50%	50%	97.50%	Mean	Dev.	2.50%	50%	97.50%	Mean	Dev.
β_{11}	−0.098	−0.016	0.080	−0.013	0.046	−0.153	−0.064	0.023	−0.064	0.046
β_{21}	−0.313	−0.254	−0.191	−0.254	0.031	−0.298	−0.236	−0.171	−0.236	0.032
β_{31}	0.000	0.057	0.111	0.056	0.028	0.025	0.085	0.145	0.085	0.030
β_{41}	0.137	0.222	0.302	0.221	0.043	0.134	0.215	0.299	0.215	0.042
β_{51}	−0.242	−0.161	−0.079	−0.161	0.042	−0.240	−0.147	−0.056	−0.147	0.048
β_{61}	−0.002	0.016	0.033	0.017	0.008	−0.001	0.022	0.045	0.023	0.011
γ_1	3.091	3.576	4.113	3.581	0.256	3.359	3.857	4.389	3.858	0.267
σ_1^2	0.037	0.063	0.101	0.065	0.017	0.037	0.055	0.083	0.056	0.012
σ_1	0.194	0.252	0.317	0.253	0.032	0.192	0.234	0.289	0.236	0.025
β_{12}	−0.256	−0.150	−0.059	−0.152	0.051	−0.305	−0.223	−0.122	−0.221	0.045
β_{22}	−0.276	−0.221	−0.167	−0.221	0.029	−0.268	−0.208	−0.148	−0.208	0.031
β_{32}	−0.008	0.048	0.103	0.048	0.028	0.026	0.082	0.138	0.082	0.028
β_{42}	0.120	0.198	0.276	0.197	0.040	0.125	0.206	0.286	0.205	0.041
β_{52}	−0.219	−0.142	−0.061	−0.141	0.041	−0.209	−0.113	−0.034	−0.115	0.044
γ_2	3.424	3.955	4.551	3.965	0.284	3.665	4.230	4.850	4.232	0.297
σ_2^2	0.048	0.076	0.121	0.079	0.019	0.051	0.077	0.117	0.079	0.017
σ_2	0.220	0.276	0.349	0.278	0.034	0.225	0.278	0.343	0.279	0.030
σ_{12}	0.018	0.042	0.079	0.044	0.016	0.020	0.039	0.067	0.040	0.012
ρ	0.383	0.619	0.773	0.608	0.102	0.393	0.605	0.748	0.596	0.091
θ	6.805	7.815	8.892	7.827	0.544	0.710	0.757	0.797	0.756	0.022
τ	0.554	0.596	0.633	0.595	0.021	0.502	0.546	0.587	0.546	0.022

According to the model comparisons used, the Frank copula seems to fit the Brazilian educational data somewhat better than the others. Therefore, we applied the Frank copula to compare the performance of the small area estimates obtained from Model B with its competitors.

The main goals of modelling the indexes of proficiency are to reduce the variability of the direct estimates derived from the sampling design and to obtain accurate estimates for nonsampled schools since the direct estimators can only be obtained for the selected ones. The multivariate model provides estimates for all schools, but we need to evaluate its adequacy. The 95% credible intervals of the predictive proportions by the replica $y_{idk}^{(t)}$, for $i \in s$, contain 98.1% and 97.8%, respectively, of the observed values for the disciplines of Portuguese and mathematics.

The reduction of the variability of the direct estimates by the application of the model can be assessed by estimating the coefficients of variation (CV) of the direct estimators obtained under design-based approach, that is, $\hat{CV}_D(y_{idk}) = \sqrt{(n_{id}-1)^{-1} y_{idk}(1-y_{idk})\left(1-n_{id}N_{id}^{-1}\right)} / y_{idk}$ and by calculating the coefficients of variation obtained by employing the models, that is, $CV_p(\mu_{idk}) = \sqrt{V_p(\mu_{idk})}/E_p(\mu_{idk})$, where the symbols $E_p(.)$ and $V_p(.)$ denote the posterior mean and the posterior variance under the assumed model, respectively. Figure 1 summarises the distribution of the CVs obtained from the estimators. Clearly, the CVs generated by the models are much lower than those obtained through the direct estimation.

We also assessed the relative differences between the small area prediction for each subject provided by the approach employed and the respective true value. The same measure was calculated for the predictions obtained when the two univariate independent hierarchical beta regressions are fitted. Figure 2 shows the box plots for the approaches. It can be seen from Figure 2 that there is some gain in using either Model A or Model B compared to the univariate separated model for both subjects. Model B performs a bit better than Model A in all scenarios. However, Figure 2 shows that model-based estimates

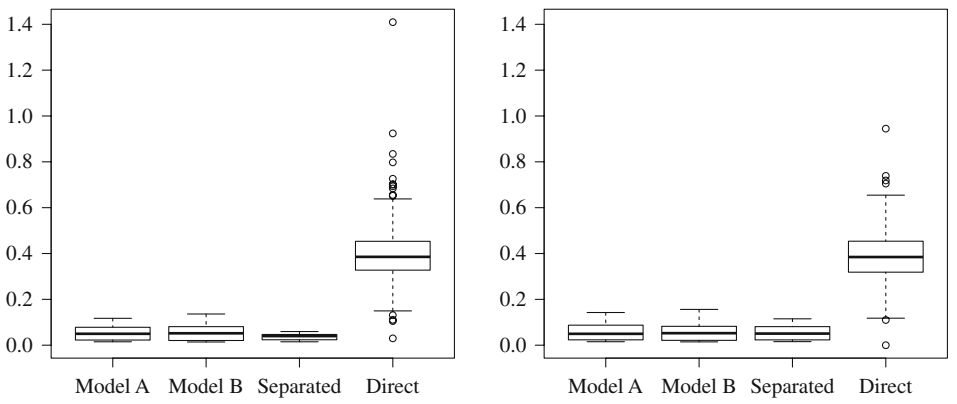
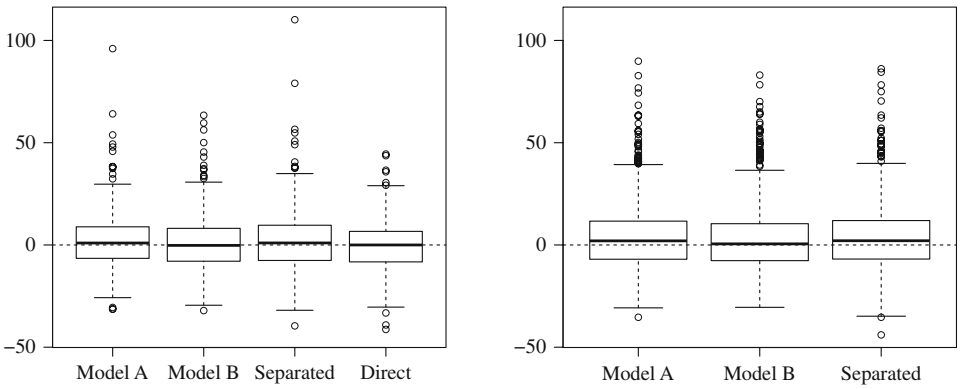
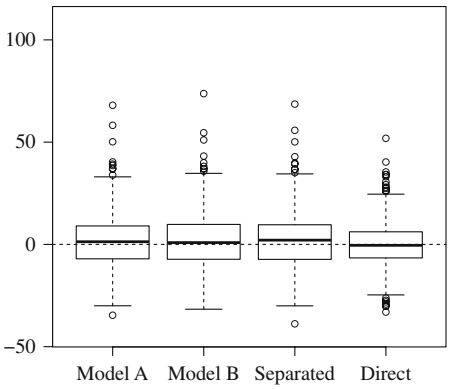


Fig. 1. Box plots of the coefficients of variation of the model-based estimators and the direct estimator for the sampled schools: Portuguese (a) and Mathematics (b).

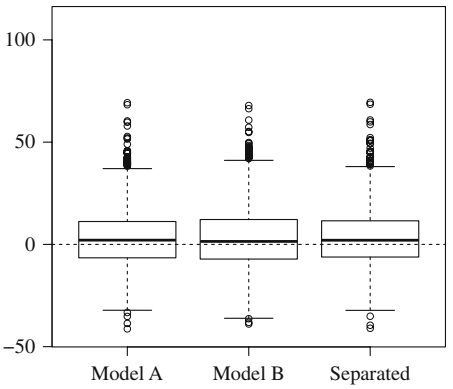


(a) Portuguese test: sampled schools

(b) Portuguese test: nonsampled schools



(c) Mathematics test: sampled schools



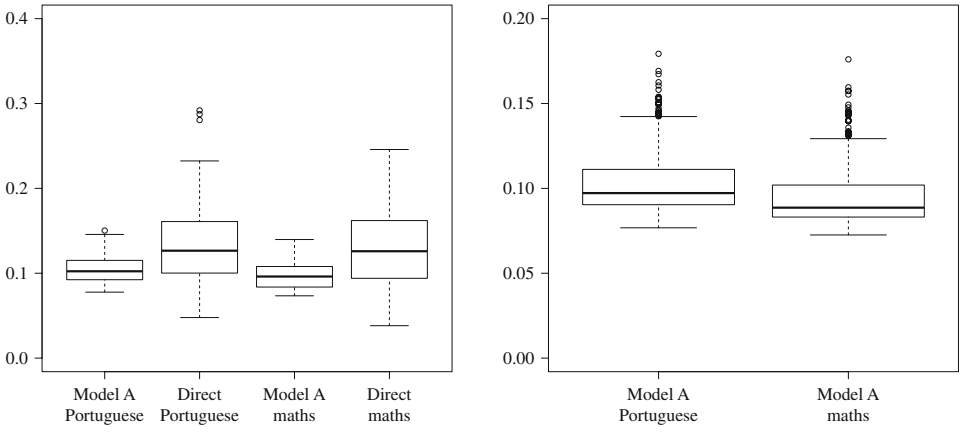
(d) Mathematics test: nonsampled schools

Fig. 2. Box plots of the relative differences in (%) between the small area prediction for each subject provided by the approach employed and the respective true value, carried out separately for schools in and out of the sample.

do not seem to substantially reduce the true relative errors of the direct estimates. This might be due to the fact that both sample sizes of schools and students for many municipalities and schools are not small enough to achieve considerable improvement of model-based estimates over designed-based estimates. This issue is investigated further in Subsubsection 3.1.4.

3.1.4. Some Results: Study 2

We conducted a second study to investigate the effect of reducing school and student sample sizes on the model-based estimates' improvement over the designed-based estimates. In this second study, the population consists of schools with at least 50 students who had taken both tests. Municipalities that had only one school after these exclusions were also discarded from the population, leaving 96,941 students. Then a simple random sample without replacement of ten percent of schools was selected, ensuring that at least two schools and at most seven schools would be selected per municipality. In the second

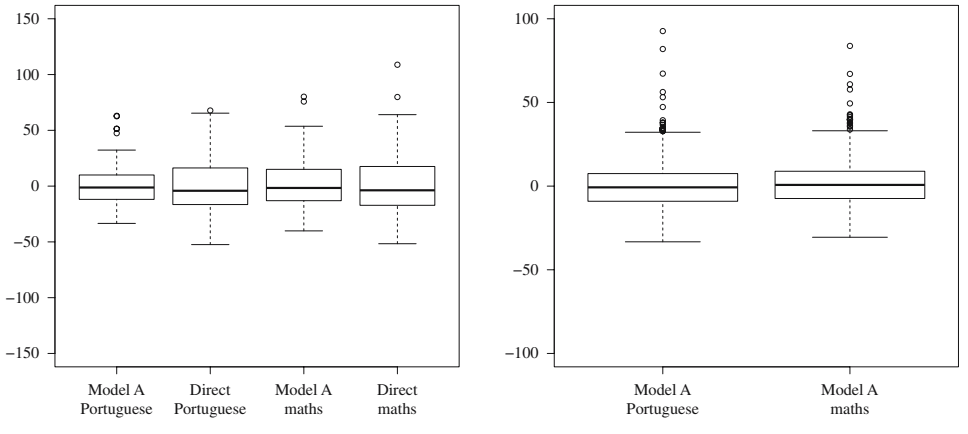


(a) Sampled schools (b) Nonsampled schools

Fig. 3. Coefficients of variation of the model-based A estimator and the direct estimator for the schools in sample (a) and the coefficients of variation of the model-based A estimator for the schools out of sample (b).

stage, a simple random sample without replacement of ten percent of students in each school were selected, imposing a restriction of a maximum of five students per school. The population consists of 32 municipalities in which 87 schools have been selected out of 1,062 schools, making a total of 719 students in the sample. In this study, we only compared the Model A estimates to the direct estimates.

We fitted Model A to the sample using the same priors described in Study 1. Figure 3 summarises the distribution of the CVs obtained from the Model A estimator and the direct estimator for both subjects. As expected, the CVs obtained by the Model A are still much lower than those obtained by the direct estimation for sampled schools.



(a) Sampled schools (b) Nonsampled schools

Fig. 4. Box plots of the relative differences (in %) with respect to the true value for the model-based A estimator and the direct estimator for the schools in sample (a); Coefficients of variation of the model-based A estimator for the schools out of sample (b).

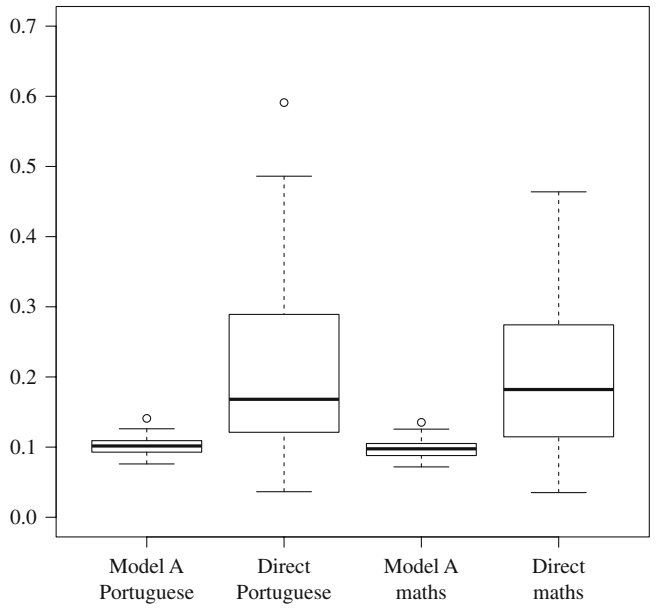


Fig. 5. Coefficients of variation of the model-based A estimator and the direct estimator calculated at municipality level for the subjects of Portuguese and mathematics.

Figure 4 shows the box plots of the relative differences between the school mean prediction for each subject provided by model-based A and designed-based approaches with respect to true value. It can be seen from Figure 4 that the reduction of the prediction errors of the proposed model-based A estimates with respect to the direct

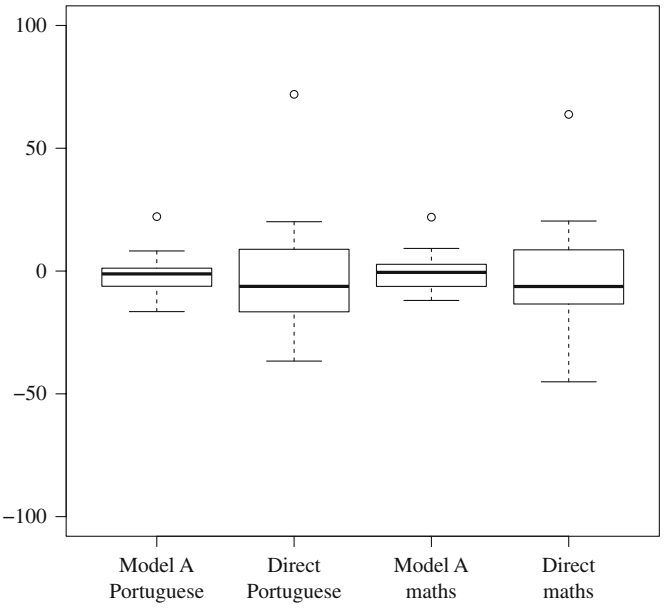


Fig. 6. Box plots of the relative differences in (%) with respect to the true value for the model-based A estimator and the direct estimator calculated at municipality level.

estimator is more appreciable than the respective result obtained from Study 1 for both subjects.

We also obtained the posterior means and the posterior variances, and consequently the coefficients of variation under the assumed model, for all 32 municipalities using formulas in Subsubsection 3.1.2. Estimates of the coefficients of variation of the direct estimators were obtained by using standard formulae to estimate the means and the variances in two-stage cluster sampling without replacement with equal probability of selection in both stages (schools and students), see, for example Lohr (1999, 147) for formulae.

Figures 5 and Figure 6 respectively present box plots of the coefficients of variation and the relative differences with respect to the true value of the Model A estimates and the direct estimates at municipality level for the subjects of Portuguese and mathematics. As expected, the gain in precision of the model-based estimates with respect to the direct estimates at municipality level are much higher than at school level for both subjects.

4. Concluding Remarks and Suggestions for Future Work

The proposed models have the advantage of keeping the response variables at their original scale. Another advantage is the use of copulas, which are marginal free, that is, the degree of the association of the variables is preserved regardless of the marginal distributions. Thus, if two indexes are correlated, whatever marginal is adopted, the measure of dependence is the same. The use of copula functions in beta marginal regressions allows the joint analysis of the response variables by taking advantage of their dependency structure. The application of multivariate models with beta responses is an appealing alternative to models that require transforming the original variables. The choice between the proposed models and their competitors in the literature should be guided by the goals of the researcher, who must observe the models' predictive power and goodness of fit. The disadvantage of models that use copulas is that they are time consuming when simulating samples from the posterior distributions of the model parameters or functions of them.

In Section 2, we propose a multivariate hierarchical model with two levels. The variables are correlated on the first level with the aid of a copula function. Despite being applicable in general situations, this model has been developed especially for the small area estimation problem to allow strength to be borrowed across the areas or small domains of interest. The random effects of the same area are assumed to be correlated, and the random effects of different areas have the same variance-covariance matrix. In the illustration presented, the multivariate hierarchical model estimated the expected indexes of proficiency for nonsampled schools and additionally presented a significant reduction of the coefficients of variation compared to the direct estimates at school as well as at municipality level.

Sample household surveys are important sources of potential applications of the models proposed in this work. Examples of variables measured in the range (0,1) are the unemployment rate and the poverty gap, the latter of which measures, on average, the distance between the poor and the poverty line. These variables are important measures both for planning and in the knowledge of the population conditions, but are rarely available for small geographic levels or population subgroups for intercensus periods. Prediction of these poverty indexes could be performed using the models proposed in this article.

It is important to note that this work focuses on building multivariate regression models in which the marginal distributions are beta. This work notes the advantages of these models over corresponding univariate models and proposes a strategy for estimating their parameters. However, the theory of copula functions can be applied to any multivariate model that can be built for any known marginal distributions, provided that the distributions of response variables are different. We can even have continuous and discrete variables in the same model. To build a model for other distributions is straightforward, but each model has a peculiar and practical feature. In the specific case of the beta model, the mean and the dispersion have been adopted as the model parameters, where the latter parameter controls the variance. Other parametrisations are possible but could lead to additional difficulties. Various strategies can be defined by the researcher according to the available data. Some important strategies are first to fix the marginal and then obtain the appropriate copulas or to estimate models with different copulas and marginal densities and decide which is the "best" model by applying a model-comparison approach.

As can be seen in a simulation study in Souza (2011), when responses share exactly the same set of regressors, the results of univariate and multivariate approaches show little difference. In the abovementioned study and applications with real data, the model selection criteria were unable to show which approach was preferable. However, in the application presented in this article where the explanatory variables are not the same for both responses, we could see a better performance of the multivariate model. Similar findings were reported for other models and can be seen in Bartels and Fiebig (1991), Gueorguieva and Agresti (2001) and Teixeira-Pinto and Normand (2009).

It should be noted that the CV of the direct estimates were calculated under a design-based approach and the measure of precision of the proposed model under a model-based (Bayesian model-based) approach. Therefore, although we can interpret γ_k^{-1} as the deff, the calculation is based on model-based premises, and thus the ratio between the model-based variance estimate and its respective estimate of the direct estimator variance could be different from the design-based deff for each school. A possible extension of our proposed approach should allow the values of deff to vary with schools.

Another point is that in practical situations where the response variables can have values of zero or one, the beta distribution will not be adequate. One possible way to circumvent this problem is to use a mixture of distributions so that the zeros and ones can be accommodated. Ospina and Ferrari (2012) propose a general class of inflated regression models to fit data with such features. We have not considered omitted values in the explanatory variables in our model formulation, which could be another possible extension of the models proposed here.

5. References

Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory*, edited by B.N. Petrov and F. Csaki, 267–281. Budapest: Akademiai Kiado.

Bartels, R. and D.G. Fiebig. 1991. "A Simple Characterization of Seemingly Unrelated Regressions Models in Which OLS is Blue." *The American Statistician* 45: 137–140. Doi: http://dx.doi.org/10.2307/2684378.

Branscum, A.J., W.O. Johnson, and M.C. Thurmond. 2007. "Bayesian Beta Regression: Applications to Household Expenditure Data and Genetic Distance Between Foot-and-Mouth Disease Viruses." *Australian & New Zealand Journal of Statistics* 49: 287–301. Doi: http://dx.doi.org/10.1111/j.1467842X.2007.00481.x.

Cepeda-Cuervo, E., J.A. Achcar, and L.G. Lopera. 2014. "Bivariate Beta Regression Models: Joint Modeling of the Mean, Dispersion and Association Parameters." *Journal of Applied Statistics* 41: 677–687. Doi: http://dx.doi.org/10.1080/02664763.2013.847071.

Da-Silva, C.Q., H.S. Migon, and L.T. Correia. 2011. "Dynamic Bayesian Beta Models." *Computational Statistics and Data Analysis* 55: 2074–2089. Doi: http://dx.doi.org/10.1016/j.csda.2010.12.011.

Datta, G.S., B. Day, and I. Basawa. 1999. "Empirical Best Linear Unbiased and Empirical Bayes Prediction in Multivariate Small Area Estimation." *Journal of Statistical Planning and Inference* 75: 269–279. Doi: http://dx.doi.org/10.1016/S0378-3758(98)00147-5.

Doornik, J.A. 2007. *Object-Oriented Matrix Programming Using Ox*, 3 ed. London: Timberlake Consultants Press.

Fabrizi, E., M.R. Ferrante, S. Pacei, and C. Trivisano. 2011. "Hierarchical Bayes Multivariate Estimation of Poverty Rates Based on Increasing Thresholds for Small Domains." *Computational Statistics and Data Analysis* 55: 1736–1747. Doi: http://dx.doi.org/10.1016/j.csda.2010.11.001.

Fay, R.E. 1987. "Application of Multivariate Regression to Small Domain Estimation." In *Small Area Statistics*, edited by R. Platek, J. Rao, C. Särndal, and M. Singh, 91–102. New York: Wiley.

Ferrari, S.L.P. and F. Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31: 799–815. Doi: http://dx.doi.org/10.1080/0266476042000214501.

Gamerman, D. and H.F. Lopes. 2006. *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*, 2 ed. London: Chapman & Hall.

Gelman, A. (2006). "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1: 515–534. Doi: http://dx.doi.org/10.1214/06-BA117A.

Gilks, W. and G. Roberts. 1996. "Strategies for Improving mcmc." In *Markov Chain Monte Carlo in Practice*, edited by S.R.W. Gilks and D. Spiegelhalter, 89–114. London: Chapman & Hall.

Gueorguieva, R.V. and A. Agresti. 2001. "A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous Responses." *Journal of the American Statistical Association* 96: 1102–1112. Doi: http://dx.doi.org/10.1198/016214501753208762.

Huard, D., G. Évin, and A.-C. Favre. 2006. "Bayesian Copula Selection." *Computational Statistics and Data Analysis* 51: 809–822. Doi: http://dx.doi.org/10.1016/j.csda.2005.08.010.

Jiang, J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics. New York: Springer.

Liu, B., P. Lahiri, and G. Kalton. 2014. "Hierarchical Bayes Modeling of Survey-Weighted Small Area Proportions." *Survey Methodology* 40: 1–13. Available at: http://www.statcan.gc.ca/pub/12-001-x/2014001/article/14030-eng.pdf (accessed 1 December 2016).

Lohr, S.L. 1999. *Sampling: design and analysis*, 1st ed. Place of publication: Brooks/Cole Publishing Company, California, USA.

Melo, T.F., K.L. Vasconcellos, and A.J. Lemonte. 2009. "Some Restriction Tests in a New Class of Regression Models for Proportions." *Computational Statistics and Data Analysis* 53: 3972–3979. Doi: http://dx.doi.org/10.1016/j.csda.2009.06.005.

Murteira, J.M.R. and J.J.S. Ramalho. 2014. "Regression Analysis of Multivariate Fractional Data." *Econometric Reviews* 0: 1–38. Doi: http://dx.doi.org/10.1080/07474938.2013.806849.

Neal, R.M. (2003). "Slice Sampling." *The Annals of Statistics* 31: 705–767. Doi: http://dx.doi.org/10.1214/aos/1056562461.

Nelsen, R.B. 2006. *An Introduction to Copulas*, 2 ed. New York: Springer.

Olkin, I. and R. Liu. 2003. "A Bivariate Beta Distribution." *Statistics & Probability Letters* 62: 407–412. Doi: http://dx.doi.org/10.1016/S0167-7152(03)00048-8.

Ospina, R. and S.L. Ferrari. 2012. "A General Class of Zero-or-One Inflated Beta Regression Models." *Computational Statistics and Data Analysis* 56: 1609–1623. Doi: http://dx.doi.org/10.1016/j.csda.2011.10.005.

Pfeffermann, D., F.A. da Silva Moura, and P.L. do Nascimento Silva. 2006. "Multi-Level Modelling Under Informative Sampling." *Biometrika* 93: 943–959. Doi: http://dx.doi.org/10.1093/biomet/93.4.943.

Rao, J.N.K. and I. Molina. 2015. *Small area estimation*, 2nd ed. Hoboken, New Jersey: Wiley.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics* 6: 461–464. Doi: http://dx.doi.org/10.1214/aos/1176344136.

Silva, R.S. and H.F. Lopes. 2008. "Copula, Marginal Distributions and Model Selection: a Bayesian Note." *Statistics and Computing* 18: 313–320. Doi: http://dx.doi.org/10.1007/s11222-008-9058-y.

Simas, A.B., W. Barreto-Souza, and A.V. Rocha. 2010. "Improved Estimators for a General Class of Beta Regression Models." *Computational Statistics and Data Analysis* 54: 348–366. Doi: http://dx.doi.org/10.1016/j.csda.2009.08.017.

Smithson, M. and J. Verkuilen. 2006. "A Better Lemon-Squeezer? Maximum Likelihood Regression With Beta-Distributed Dependent Variables." *Psychological Methods* 11: 54–71. Doi: http://dx.doi.org/10.1037/1082-989X.11.1.54.

Souza, D.F. 2011. "Regressão Beta Multivariada com Aplicações em Pequenas Áreas." Ph.D. thesis, Instituto de Matemática da Universidade Federal do Rio de Janeiro. Available at: http://www.pg.im.ufrj.br/teses/Estatistica/Doutorado/018.pdf (accessed 1 December 2016).

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A.V.D. Linde. 2002. "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64: 583–639. Doi: http://dx.doi.org/10.1111/1467-9868.00353.

Teixeira-Pinto, A. and S.-L. T. Normand. 2009. "Correlated Bivariate Continuous and Binary Outcomes: Issues and Applications." *Statistics in Medicine* 28: 1753–1773. Doi: http://dx.doi.org/10.1002/sim.3588.

Received February 2015
Revised January 2016
Accepted February 2016

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 769–785, http://dx.doi.org/10.1515/JOS-2016-0039

Nonrespondent Subsample Multiple Imputation in Two-Phase Sampling for Nonresponse

Nanhua Zhang[1], *Henian Chen*[2], *and Michael R. Elliott*[3]

Nonresponse is very common in epidemiologic surveys and clinical trials. Common methods for dealing with missing data (e.g., complete-case analysis, ignorable-likelihood methods, and nonignorable modeling methods) rely on untestable assumptions. Nonresponse two-phase sampling (NTS), which takes a random sample of initial nonrespondents for follow-up data collection, provides a means to reduce nonresponse bias. However, traditional weighting methods to analyze data from NTS do not make full use of auxiliary variables. This article proposes a method called nonrespondent subsample multiple imputation (NSMI), where multiple imputation (Rubin 1987) is performed within the subsample of nonrespondents in Phase I using additional data collected in Phase II. The properties of the proposed methods by simulation are illustrated and the methods applied to a quality of life study. The simulation study shows that the gains from using the NTS scheme can be substantial, even if NTS sampling only collects data from a small proportion of the initial nonrespondents.

Key words: Double sampling; maximum likelihood; missing data; nonignorable missing-data mechanism; quality of life; weighting.

1. Introduction

Nonresponse is very common in population surveys and clinical trials. Complete-case analysis (CC), which discards the incomplete cases, can lead to a substantial loss of information or biased estimation of the key parameters. Since the publication of Rubin's seminal paper on missing data (Rubin 1976), a number of ignorable-likelihood (IL) methods have been developed, including ignorable maximum likelihood, Bayesian inference, and multiple imputation (Dempster et al. 1977; Rubin 1987; Heitjan and Rubin 1991; Little and Zhang 2011). IL methods provide valid inference when missingness does not depend on the underlying missing values after conditioning on available data, a state termed missing at random (MAR) (Rubin 1976; Little and Rubin 2002). When MAR holds, inference can be based on the observed-data likelihood, and thus does not require modeling assumptions about the missingness indicators. When the missingness could depend on the missing values (missing not at random (MNAR) mechanism), nonignorable models (NIM) are developed based on the joint distribution of the variables and the

[1] Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, OH 45229, U.S.A. Email: nanhua.zhang@cchmc.org (corresponding author)
[2] Department of Epidemiology & Biostatistics, College of Public Health, University of South Florida, Tampa, FL 33612-3085, U.S.A. Email: hchen1@health.usf.edu
[3] Department of Biostatistics, School of Public Health, University of Michigan and Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48019, U.S.A. Email: mrelliot@umich.edu

missing-data indicators (Heckman 1976; Amemiya 1984; Little 1993, 1994; Nandram and Choi 2002, 2010).

Both IL and NIM methods make use of all available data, but they rely on assumptions about the missing-data mechanism. IL methods are vulnerable to failures of the ignorable missingness assumption; NIM methods are vulnerable to misspecification of the missing-data mechanism and suffer from problems with identifying parameters. The assumptions about the missing-data mechanism are untestable without knowing the underlying values of the missing data. The choice may be aided by learning more about the missing-data mechanism; for example, by recording reasons why particular values are missing. The difficulty in identifying parameters in NIM may be alleviated in some special cases, such as small-area estimation (Nandram and Choi 2002). In cases where the assumptions about the missing-data mechanism cannot be determined, an alternative strategy is to perform a sensitivity analysis to see whether key results are robust to alternative methods and assumptions.

Yet another alternative is to use a study design to relax to some degree the assumptions required under IL and NIM. One such design is two-phase sampling, in which a subsample of nonrespondents to the original survey (Phase I) is selected for further interview attempts (Phase II). This method is called nonresponse two-phase sampling (NTS). It was first proposed by Hansen and Hurwitz (1946) to reduce the nonresponse bias in mail questionnaires by carrying out personal interviews with a fraction of the nonrespondents. Discussions of sample-size selection and estimation of the population mean/total can be found in Hansen and Hurwitz (1946) and Srinath (1971). Some examples of using two-phase sampling to mitigate the effects of nonresponse include the National Comorbidity Survey (Elliott et al. 2000), the 2003 Survey of Small Business Finances (Harter et al. 2007) and the 2011 Canadian National Household Survey (Statistics Canada 2011).

Previous research mainly relies on using case weights developed from the two-phase sample, rather than auxiliary variables, to reduce bias in estimating population means or totals (Hansen and Hurwitz 1946; Srinath 1971; Harter et al. 2007). This article proposes nonrespondent subsample multiple imputation (NSMI), where multiple imputation (Rubin 1987) is performed within the subsample of nonrespondents in Phase I, using additional data collected in Phase II. The rationale of NSMI is that the MAR assumption, which the multiple-imputation method is based on, is valid within the nonrespondent subsample in Phase I, but may be invalid if extended to the whole sample. This is true when the missingness in Phase I is MNAR and the NSMI reduces the nonresponse bias; when the missingness in Phase I is ignorable, the NSMI is still a valid method, although there is some loss of efficiency compared with multiple imputation using all cases.

Section 2 presents a motivating application based on data from a quality of life (QOL) study. In this application, 147 out of the 750 participants did not reply to the initial QOL survey. In Phase II, all 147 nonrespondents were recontacted and 39 provided answers to an abridged version of the QOL instrument. The NSMI method consists of multiple imputation of the missing QOL outcomes within the subsample of nonrespondents in Phase I, that is, using the partial information of the 39 respondents in Phase II to impute the missing QOL data.

Section 3 introduces the framework of NTS and the necessary notation. Section 4 reviews the methods for analyzing data from NTS and proposes NSMI. Section 5 presents

simulations that illustrate the properties of NSMI, while Section 6 applies the method to the motivating data. Section 7 concludes with a discussion.

2. Motivating Problem: A Quality of Life Study

To illustrate the methods, data from 750 participants in a community-based study—the Children in the Community study (CIC) (Cohen et al. 2005)—are considered. The sample was based on a random residence-based cohort, originally drawn from 100 neighborhoods in two upstate New York counties in 1975. Additional information regarding the study is available from Cohen et al. (2005). From 1991 to 1994 (T1), 750 youths (mean age 22 years and SD 2.8 years) were interviewed in their homes by trained interviewers. QOL data were collected as part of the survey. QOL was assessed by the young adult quality of life instrument (YAQOL) (Chen et al. 2004). In 2001–2004 (T2) at mean age 33.0 years (SD = 2.8), the same group of participants was surveyed via the web using the same QOL instrument. Of the 750 subjects assessed for QOL at T1, 603 (80.4%) completed the QOL survey at T2; 147 did not respond to the follow-up survey. For these 147 subjects, an abridged version of the QOL instrument was mailed to their home address. Upon return of the completed surveys, subjects were paid for their participation. Of the 147 eligible subjects, 39 (26.5%) returned their QOL questionnaire. The resources scale used here is taken from the abridged version and identical to that employed at T1.

The goals of the QOL analysis included estimating the mean resources score and determining whether the resources scores are related to major demographic variables—age, gender, race and education. CC analysis suffers from inefficiency and potential bias if the missingness of QOL is MNAR. IL analyses make use of the partial information in the incomplete cases, but assume the missing data are MAR. NSMI is proposed for this problem, which is shown to be valid if the conditions of the Phase II sampling are met, regardless of the missing-data mechanism in Phase I.

3. Continuing Data Collection for Nonresponse

Data with the structure in Table 1 are considered. Let $\{\boldsymbol{y}_i, i = 1, \ldots n\}$ denote n independent observations on a (possibly multivariate) outcome variable $\boldsymbol{Y}$, where $\boldsymbol{Y}$ has missing values. $\boldsymbol{Y}_{obs,1}$ is used to represent the data observed in Phase I, $\boldsymbol{Y}_{obs,2}$ to represent the data missing in Phase I, but observed in Phase II, and $\boldsymbol{Y}_{mis}$ to represent the data missing after Phase II sampling. Let $\boldsymbol{Y}_{obs} = (\boldsymbol{Y}_{obs,1}, \boldsymbol{Y}_{obs,2})$ and $\boldsymbol{Y}_{mis,1} = (\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis})$ represent the observed data after Phase II and the missing data from Phase I, respectively. The vector of

Table 1. Two-phase sampling for nonresponse and general missing-data structure for Section 3.

Pattern	Observation, i	$\boldsymbol{y}_i$	$R_{1,i}$	$S_{2\cdot1,i}$	$R_{2\cdot1,i}$	$R_{2,i}$
1	$i = 1, \ldots, m$	√	1	–	–	1
2	$i = m+1, \ldots, m+r$	x	0	1	1	1
3	$i = m+r+1, \ldots, m+s$	?	0	1	0	0
4	$i = m+s+1, \ldots, n$	?	0	0	0	0

Key: √ denotes observed; ? denotes at least one entry missing; x denotes at least one entry missing in Phase I, but observed in Phase II.

covariates, $\boldsymbol{z}_i$, is assumed to be fully observed. Interest concerns the parameters $\boldsymbol{\phi}$, which govern the conditional distribution of $\boldsymbol{y}_i$ on $\boldsymbol{z}_i, p(\boldsymbol{y}_i|\boldsymbol{\phi}, \boldsymbol{z}_i)$.

In Phase I, $\boldsymbol{y}_i$s are observed for $i = 1, \ldots, m$, but contain missing values for $i = m + 1, \ldots, n$. The response indicator for Phase I is denoted as $R_{1,i}$, equal to 1 if $\boldsymbol{y}_i$ is observed and 0 otherwise. In Phase II, s subjects were sampled from the nonrespondents in Phase I and r subjects responded. $S_{2\cdot1,i}$ is used to denote whether a subject was sampled among the nonrespondents in Phase I. Let π_i denote the Phase II sampling probability among nonrespondents in Phase I,

$$\pi_i = \Pr(S_{2\cdot1,i} = 1|R_{1,i} = 0; \boldsymbol{z}_i, \boldsymbol{y}_i). \tag{1}$$

After Phase II sampling, data on r additional subjects were collected. $R_{2\cdot1,i}$ is used to denote the Phase II response indicator among the nonrespondents in Phase I. The overall response indicator after completion of Phase II is denoted as $R_{2,i}$. Depending on the context, the second-stage sampling may be a simple random, stratified or other probability sampling scheme. In certain settings such as this example, all nonrespondents may be contacted, with $\pi_i = 1$ for all i, so that $m + s = n$ and the fourth row in Table 1 is empty.

The rows of Table 1 divide the cases into four patterns. Pattern 1 ($i = 1, \ldots, m$) consists of subjects for whom $\boldsymbol{y}_i$ is fully observed after first-phase data collection. Pattern 2 consists of cases that were missing in Phase I, but subsequently observed in Phase II sampling. Pattern 3 consists of cases that were sampled in Phase II, but did not respond, and Pattern 4 were those Phase I nonrespondents were not sampled in Phase II.

4. A Comparison of Methods for Analyzing the Data

4.1. *Ignorable Likelihood Using Multiple Imputation (MI)*

In this subsection, data with the structure in Table 2 are considered. $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ are used to denote the observed and missing component of the data $\boldsymbol{Y}$, respectively. R_i is used to denote the response indicator, equal to 1 if $\boldsymbol{y}_i$ is observed and 0 otherwise. $\boldsymbol{Z}$ denotes the covariates that are fully observed. When the data contain missing values, the full model to describe the data is the joint distribution of $\boldsymbol{Y}_{obs}$, $\boldsymbol{Y}_{mis}$ and $\boldsymbol{R}$ conditional on $\boldsymbol{Z}, P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z})$, where $\boldsymbol{\xi}$ is the parameter associated with the distribution of the response indicator $\boldsymbol{R}$. The observed likelihood can be written as:

$$L(\boldsymbol{\phi}, \boldsymbol{\xi}|\boldsymbol{Y}_{obs}, \boldsymbol{R}; \boldsymbol{Z}) \propto P(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) \tag{2}$$

Table 2. General missing-data structure for Subsection 4.1.

Pattern	Observation, i	$\boldsymbol{y}_i$	$\boldsymbol{R}_i$
1	$i = 1, \ldots, m$	$\surd$	1
2	$i = m + 1, \ldots, n$	?	0

Key: $\surd$ denotes observed; ? denotes at least one entry missing.

where

$$P(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) d\boldsymbol{Y}_{mis}$$
$$= \int P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \boldsymbol{\xi}, \boldsymbol{Z}) d\boldsymbol{Y}_{mis}. \quad (3)$$

When the missing-data mechanism is missing completely at random (MCAR) or MAR (Little and Rubin 2002), (3) becomes

$$P(\boldsymbol{Y}_{obs}, \boldsymbol{R}|\boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) = \begin{cases} P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}|\boldsymbol{\xi}; \boldsymbol{Z}) & \text{if MCAR} \\ P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{\xi}; \boldsymbol{Z}) & \text{if MAR.} \end{cases} \quad (4)$$

Under the further condition that the parameter spaces of $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are distinct, the likelihood-based inference on ϕ can be conducted based on $P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z})$, ignoring the missing-data mechanism:

$$L(\boldsymbol{\phi}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) \propto P(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}). \quad (5)$$

Likelihood-based methods that ignore the missing-data mechanism are called ignorable likelihood (Little and Zhang 2011). Options for IL are maximum-likelihood estimation, Bayesian inference, and multiple imputation. Bayesian inference is based on the posterior distribution of $\boldsymbol{\phi}$ given by:

$$P(\boldsymbol{\phi}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) \propto L(\boldsymbol{Y}_{obs}|\boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{\phi}), \quad (6)$$

where $P(\boldsymbol{\phi})$ is the prior distribution of $\boldsymbol{\phi}$.

Another option of IL is multiple imputation, that is, to impute the missing data $\boldsymbol{Y}_{mis}$, and then apply complete-data-based methods to the imputed data to make inference on the parameters $\boldsymbol{\phi}$. Multiple imputation is closely related to Bayesian inference. The imputation of $\boldsymbol{Y}_{mis}$ is based on the posterior predictive distribution of $\boldsymbol{Y}_{mis}$ given $\boldsymbol{Y}_{obs}$, which is the conditional predictive distribution, $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\phi}; \boldsymbol{Z})$, averaged over the posterior distribution of $\boldsymbol{\phi}$, that is,

$$P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{\phi}|\boldsymbol{Y}_{obs}; \boldsymbol{Z}) d\boldsymbol{\phi}. \quad (7)$$

In order to generate M sets of imputations given $\boldsymbol{Y}_{obs}$, M values of $\boldsymbol{\phi}$ are independently drawn from the posterior distribution, say $\tilde{\boldsymbol{\phi}}^{(t)} (t = 1, \ldots, M)$. For each $\tilde{\boldsymbol{\phi}}^{(t)}$, one set of imputed values of $\boldsymbol{Y}_{mis}$ is obtained by taking a random draw of $\boldsymbol{Y}_{mis}$ from the corresponding posterior predictive distribution $P\left(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \tilde{\boldsymbol{\phi}}^{(t)}; \boldsymbol{Z}\right)$. Rubin (1987) showed that when the proper imputation method is followed (i.e., an imputation method that accounts for the uncertainty in the model parameters), the resulting inference based on the multiply imputed datasets is valid. The M imputed datasets are then analyzed as if each of them is a complete dataset. The analysis results from M imputed datasets are combined following the multiple-imputation combining rules (Rubin 1987).

Unlike IL methods, CC analysis discards all cases that contain missing values and is based on the following likelihood:

$$L_{\text{cc}}(\boldsymbol{\phi}) = \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i | \boldsymbol{R} = 1; \boldsymbol{\phi}, \boldsymbol{Z}). \tag{8}$$

The estimation of $\boldsymbol{\phi}$ is obtained through maximizing $L_{\propto}(\boldsymbol{\phi})$; CC analysis is the default method in most statistical packages.

4.2. *Complete-Case and Ignorable-Likelihood Methods*

In this section, data with the structure in Table 1 are considered. The notation is the same as in Section 3. Depending on whether the additional data, $\boldsymbol{Y}_{obs,2}$ are used, there are two versions of complete-case analysis and ignorable-likelihood method (e.g., multiple imputation). The ignorable-likelihood methods that use data in Phase II (IL2) can be written as:

$$\begin{aligned} L_{\text{ign},2}(\boldsymbol{\phi}) &= P(\boldsymbol{Y}_{obs} | \boldsymbol{\phi}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\phi}; \boldsymbol{Z}) d\boldsymbol{Y}_{mis} \\ &= \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i | R_{1,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}) \times \prod_{i=m+1}^{m+r} p(\boldsymbol{y}_i | R_{1,i} = 0, R_{2,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}) \\ &\times \int \cdots \int \prod_{i=m+r+1}^{n} p(\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis} | R_{2,i} = 0; \boldsymbol{\phi}, \boldsymbol{Z}) d\boldsymbol{y}_{m+r+1,mis} \cdots d\boldsymbol{y}_{n,mis} \end{aligned} \tag{9}$$

where $\boldsymbol{y}_{i,obs}$ consists of the fully observed components of $\boldsymbol{y}_i$.

Rubin's (1976) theory shows that a sufficient condition for valid inference based on (9) is that MAR holds in the Phase II data, that is:

$$P(\boldsymbol{R}_2 | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\xi}, \boldsymbol{Z}) = P(\boldsymbol{R}_2 | \boldsymbol{Y}_{obs}; \boldsymbol{\xi}, \boldsymbol{Z}). \tag{10}$$

A complete-case analysis using Phase II data (CC2) bases inferences for $\boldsymbol{\phi}$ on the complete observations in Patterns 1 and 2. In a likelihood context, the method bases inference on the conditional likelihood corresponding to the complete cases after Phase II sampling, namely:

$$L_{\text{cc},2}(\boldsymbol{\phi}) = \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i | R_{1,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}) \times \prod_{i=m+1}^{m+r} p(\boldsymbol{y}_i | R_{1,i} = 0, R_{2,i} = 1; \boldsymbol{\phi}, \boldsymbol{Z}). \tag{11}$$

Note that the first part of (9) is exactly the same as (11), and that the second part explains how (9) uses the partially observed component of the outcome $\boldsymbol{y}_i$ (possibly multivariate) for $i = m+r+1, \ldots, n$. The key assumption under which inference based on $L_{\propto,2}(\boldsymbol{\phi})$ is valid is that the missingness after Phase II is MCAR,

$$P(\boldsymbol{R}_2 | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}; \boldsymbol{\xi}, \boldsymbol{Z}) = P(\boldsymbol{R}_2 | \boldsymbol{\xi}, \boldsymbol{Z}). \tag{12}$$

Note that (12) is a special case of (10); when the missingness after Phase II is MCAR, the IL2 method is also valid and more efficient than CC2 because CC2 removes from the analysis all cases that were not observed after Phase II sampling, and fails to use the information in the partially observed data.

The ignorable-likelihood methods that use only the Phase I data (IL1) are based on the following likelihood:

$$\begin{aligned} L_{\text{ign},1}(\boldsymbol{\phi}) = P(\boldsymbol{Y}_{obs,1}|\boldsymbol{\phi},\boldsymbol{Z}) &= \int\int P(\boldsymbol{Y}_{obs,1},\boldsymbol{Y}_{obs,2},\boldsymbol{Y}_{mis}|\boldsymbol{\phi},\boldsymbol{Z})d\boldsymbol{Y}_{obs,2}d\boldsymbol{Y}_{mis} \\ &= \text{const.} \times \prod_{i=1}^{m} p(\boldsymbol{y}_i|R_{1,i}=1;\boldsymbol{\phi},\boldsymbol{Z}) \\ &\times \int\cdots\int \prod_{i=m+1}^{m+r} p(\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis}|R_{1,i}=0,R_{2,i}=1;\boldsymbol{\phi},\boldsymbol{Z})d\boldsymbol{y}_{m+1,mis}\cdots d\boldsymbol{y}_{m+r,mis} \\ &\times \int\ldots\int \prod_{i=m+r+1}^{n} p(\boldsymbol{y}_{i,obs},\boldsymbol{y}_{i,mis}|R_{2,i}=0;\boldsymbol{\phi},\boldsymbol{Z})d\boldsymbol{y}_{m+r+1,mis}\cdots d\boldsymbol{y}_{n,mis} \end{aligned} \tag{13}$$

They are valid if:

$$P(\boldsymbol{R}_1|\boldsymbol{Y}_{obs,1},\boldsymbol{Y}_{obs,2},\boldsymbol{Y}_{mis};\boldsymbol{\xi},\boldsymbol{Z}) = P(\boldsymbol{R}_1|\boldsymbol{Y}_{obs,1};\boldsymbol{\xi},\boldsymbol{Z}). \tag{14}$$

Likewise, the CC analysis based only on Phase I uses cases in Pattern 1 (CC1) in Table 1, and is equivalent to (8) when no resampling has occurred. Here, the likelihood can be written as:

$$L_{\text{cc},1}(\boldsymbol{\phi}) = \text{const.} \times \prod_{i=1}^{m} p\big(\boldsymbol{y}_i|\boldsymbol{R}_1=1;\boldsymbol{\phi},\boldsymbol{Z}\big). \tag{15}$$

The CC1 analysis is valid if the corresponding missing-data mechanism is MCAR:

$$P(\boldsymbol{R}_1|\boldsymbol{Y}_{obs},\boldsymbol{Y}_{mis};\boldsymbol{\xi},\boldsymbol{Z}) = P(\boldsymbol{R}_1|\boldsymbol{\xi},\boldsymbol{Z}). \tag{16}$$

In both cases with or without using data from Phase II, the ignorable-likelihood methods are more efficient than the CC analysis if the corresponding missing-data mechanisms are ignorable (MAR or MCAR). The choice between IL1 and IL2 relies on whether (10) or (14) is a more reasonable assumption, that is, whether the MAR assumption holds among second-wave nonrespondents regardless of the first-wave missingness mechanism (suggesting IL2), or whether MAR holds among first-wave nonrespondents, but missingness is nonignorable at Phase 2 (suggesting IL1).

4.3. Nonrespondent Subsample Multiple Imputation (NSMI)

In this section, data with the structure in Table 1 are also considered. NSMI is proposed, which applies the multiple-imputation method to the cases in Patterns 2, 3, and 4. In the NSMI method, we leave out the subjects in Pattern 1 when the missing values in Pattern 3 and 4 were imputed, and then the imputed datasets from Patterns 2, 3, 4 are combined with data from Pattern 1 for statistical analyses. The method is valid if within the nonrespondents in Phase I (Patterns 2, 3, and 4), the missingness after Phase II sampling is MAR, namely,

$$(\boldsymbol{R}_{2\cdot1}=1|\boldsymbol{R}_1=0,\boldsymbol{Y}_{obs,2},\boldsymbol{Y}_{mis};\boldsymbol{\xi},\boldsymbol{Z}) = \text{P}(\boldsymbol{R}_{2\cdot1}=1|\boldsymbol{R}_1=0,\boldsymbol{Y}_{obs,2};\boldsymbol{\xi},\boldsymbol{Z}) \tag{17}$$

This missingness mechanism is called nonrespondent subsample missing at random (NS-MAR). Conditioning on $\boldsymbol{R}_1 = 0$, the joint distribution of $\boldsymbol{Y}_{obs,2}$, $\boldsymbol{Y}_{mis}$ and $\boldsymbol{R}_{2\cdot1}$ can be written as:

$$\begin{aligned} &P(\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}, \boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) \\ &\quad = P(\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}; \boldsymbol{Z}) P(\boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}, \boldsymbol{\xi}; \boldsymbol{Z}) \end{aligned} \tag{18}$$

The joint distribution of $\boldsymbol{Y}_{obs,2}$ and $\boldsymbol{R}_{2\cdot1}$ conditional on $\boldsymbol{R}_1 = 0$ is obtained by integrating out $\boldsymbol{Y}_{mis}$ (Little and Rubin 2002):

$$P(\boldsymbol{Y}_{obs,2}, \boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) = \int P(\boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}, \boldsymbol{R}_{2\cdot1}|\boldsymbol{R}_1 = 0, \boldsymbol{\phi}, \boldsymbol{\xi}; \boldsymbol{Z}) d\boldsymbol{Y}_{mis} \tag{19}$$

The key assumption for the NSMI methods is NS-MAR, which ensures that the imputed values are from the predictive distribution of $\boldsymbol{Y}_{mis}$.

It should be noted that the assumption in (17) does not confine the missing-data mechanisms in the whole sample ($\boldsymbol{R}_2$) or the missing-data mechanism in Phase I ($\boldsymbol{R}_1$) to a certain missing-data mechanism, and therefore NSMI may be applied even under the MNAR missingness mechanism in Phase I or Phase I/II combined data as long as Phase II is MCAR or MAR. In contrast, the IL2 assumptions are violated, since under NS-MAR we have:

$$P(\boldsymbol{R}_1 = 1|\cdot) = f_1(\boldsymbol{X}, \boldsymbol{Y}_{obs,1}, \boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}), \mathrm{P}(\boldsymbol{R}_{2\cdot1} = 1|\boldsymbol{R}_1 = 0, \cdot) = f_2(\boldsymbol{X}, \boldsymbol{Y}_{obs,2})$$

where f_1 and f_2 are arbitrary functions, and thus:

$$\begin{aligned} P(\boldsymbol{R}_2 = 1|\cdot) &= P(\boldsymbol{R}_1 = 1|\cdot) P(\boldsymbol{R}_2 = 1|\boldsymbol{R}_1 = 1, \cdot) + P(\boldsymbol{R}_1 = 0|\cdot) P(\boldsymbol{R}_2 = 1|\boldsymbol{R}_1 = 0, \cdot) \\ &= P(\boldsymbol{R}_1 = 1|\cdot) + P(\boldsymbol{R}_1 = 0|\cdot) P(\boldsymbol{R}_2 = 1|\boldsymbol{R}_1 = 0|\cdot) \\ &= f_1 + (1 - f_1) f_2 \end{aligned}$$

Since f_1 involves missing values, the distribution of $\boldsymbol{R}_2$ depends on underlying missing values, and therefore the assumption for IL2 is violated.

5. Simulation Studies

This section illustrates the properties of the NSMI method using simulation studies and compares the performance of NSMI to other methods under different missing-data mechanisms in Phases I and II. For each simulation study, six methods are applied to estimate the mean of the outcome Y and the regression coefficient of Y on scalar covariates Z and X:

1. *BD:* estimates using the data before deletion (BD), that is, the full data generated from simulation before missing values are created, as a benchmark method.
2. *CC1:* complete-case analysis using respondents from Phase I.
3. *CC2:* complete-case analysis using respondents from both Phases I and II.
4. *IL1:* multiple imputation using data from Phase I.

5. *IL2:* multiple imputation using data from both Phases I and II.
6. *NSMI:* multiple imputation in the nonrespondent subsample in Phase I using only additional data from Phase II.

The first three methods (BD, CC1, CC2) were implemented using standard maximum-likelihood estimation procedures in the software package R version 2.15.0 (R Development Core Team 2012). Methods 4–6 were implemented in the R package mice (multiple imputation through chained equations, Van Buuren and Groothuis-Oudshoom 2011); the number of imputed datasets was ten and the default was used for other options.

This article compares the performance of each of the methods using empirical bias, root mean square errors (RMSE), and the coverage probabilities of the 95% confidence intervals.

The first set of simulations generates $(z, x)_i$ from the normal distribution with mean 0, and covariance matrix $\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$, for $i = 1,2,\ldots,1{,}000$. $\boldsymbol{Y}$ is related to $\boldsymbol{Z}$ and $\boldsymbol{X}$ by the linear model:

$$y_i = 1 + z_i + x_i + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0,1).$$

The response $\boldsymbol{Y}$ is subject to missingness, while $\boldsymbol{Z}$ and $\boldsymbol{X}$ are fully observed. Two covariates are used to allow the response mechanisms in Phases I and II to depend on different covariates (depending on z in Phase I and on x in Phase II). Let R_i denote the response indicator for y_i in Phase I. Phase I missing values in $\boldsymbol{Y}$ are generated based on the following three missing-data mechanisms:

(I) MCAR: $\Pr(R_i = 0|z_i, x_i, y_i) = \text{expit}(-1)$;
(II) MAR: $\Pr(R_i = 0|z_i, x_i, y_i) = \text{expit}(-1 + z_i)$;
(III) MNAR: $\Pr(R_i = 0|z_i, x_i, y_i) = \text{expit}(-y_i)$;

where $\text{expit}(\cdot)$ is the inverse logit function, $\text{expit}(\cdot) = \exp(\cdot)/[1 + \exp(\cdot)]$. R_i is then generated from a Bernoulli distribution with probability $\Pr(R_i = 0|z_i, x_i, y_i)$. Each missing-data generation scheme results in approximately 27% of the values of $\boldsymbol{Y}$ being missing in Phase I.

Let $R_{2\cdot 1,i}$ denote the response indicator in the subsample of nonrespondents in Phase I. Phase II responses in $\boldsymbol{Y}$ are generated under an MCAR mechanism:

$$\Pr(R_{2\cdot 1,i} = 1|R_i = 0; z_i, x_i, \boldsymbol{Y}_i) = 0.25.$$

The biases, root RMSE, and coverage probabilities of the 95% confidence intervals from each of the six methods are reported in Table 3. Results are based on 1,000 repetitions for each simulated condition.

For the MCAR missing-data mechanism in Phase I, all methods yield approximately unbiased estimates of both the mean of $\boldsymbol{Y}$ and the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$. IL2 has the smallest RMSE for the population mean since it makes full use of the data. For the regression parameters, CC2 and IL2 give comparable estimates since the incomplete cases do not contain additional information for the regression of $\boldsymbol{Y}$ on the covariates for this missing-data mechanism (Little and Zhang 2011). The NSMI method has moderately

Table 3. Empirical bias, root mean square error, and 95% CI coverage probabilities under three missing-data mechanisms in Phase I and MCAR in Phase II (1,000 replications).

		MCAR				MAR				MNAR			
		μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z
Bias ($\times 10^4$)	BD	10	−13	1	15	−5	−18	7	−16	16	4	−9	−16
	CC1	15	−14	8	6	−6035	−10	18	−14	7961	2827	−1081	−1082
	CC2	16	−17	9	9	−4038	−14	7	−10	5286	1661	−517	−516
	IL1	8	−15	8	3	2	−11	15	−12	2840	2829	−1082	−1081
	IL2	4	−19	12	10	1	−12	7	−9	1673	1662	−516	−520
	NSMI	−2	−25	17	14	−1	−14	−13	6	26	15	−8	−20
RMSE ($\times 10^4$)	BD	578	311	327	331	602	318	332	326	607	307	341	337
	CC1	673	375	380	397	6071	417	442	429	7984	2854	1161	1161
	CC2	647	357	364	375	4091	373	397	399	5324	1700	651	652
	IL1	614	380	389	400	682	431	450	438	2903	2857	1165	1163
	IL2	606	361	371	381	644	379	404	404	1784	1701	653	660
	NSMI	663	440	454	468	699	477	519	496	681	428	483	487
Coverage	BD	96.5	94.7	95.4	95.7	94.8	93.9	95.0	96.3	94.6	95.8	94.1	94.5
	CC1	95.7	94.4	95.6	93.9	0.0	95.5	94.4	95.9	0.0	0.0	24.7	24.1
	CC2	95.1	94.4	94.9	94.7	0.0	95.1	94.6	95.6	0.0	0.2	73.7	72.5
	IL1	95.8	94.6	95.0	93.1	94.0	95.1	94.2	95.7	0.3	0.0	26.8	28.8
	IL2	96.3	93.7	95.0	93.6	93.7	95.2	93.5	95.3	23.8	0.3	73.5	73.6
	NSMI	96.1	94.8	94.9	95.0	94.0	93.5	93.5	94.9	95.4	96.1	94.7	94.8

larger RMSEs because of increased variability in the imputed values, which uses subjects from Patterns 2, 3, and 4, but not subjects from Pattern 1.

For the MAR missing-data mechanism in Phase I, all methods give approximately unbiased estimates for the regression coefficients, but the CC1 and CC2 methods show significant biases in estimating the mean of $\boldsymbol{Y}$. This is not surprising because the missingness of $\boldsymbol{Y}$ depends on $\boldsymbol{X}$, which is conditioned on in the estimation of the regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$ and $\boldsymbol{X}$, but not in the estimation of the (marginal) mean of $\boldsymbol{Y}$. As in the Phase I MCAR case, the NSMI has a somewhat greater RMSE than the IL methods, because information about the respondents in Phase I was not used in the imputation.

For MNAR missing-data generation in Phase I, the NSMI method is the only method that provides unbiased estimates of the mean of $\boldsymbol{Y}$ and the regression coefficients of $\boldsymbol{Y}$ on $\boldsymbol{Z}$ and $\boldsymbol{X}$. All other methods show significant biases because the MCAR or MAR assumptions are violated.

When the Phase II missingness mechanism is MAR, the results are similar to the results when the Phase II missingness mechanism is MCAR. Please refer to the online supplementary material for related results found at www.dx.doi.org/10.1515/jos-2016-0039.

The second set of simulations uses the same setup as in the MNAR scenario in the previous simulations, but vary the probability of being sampled in Phase II, that is, π is 0.05, 0.15, 0.25 or 0.50. The same six methods are applied on the simulated data. The bias, RMSE, and coverage probabilities are reported in Table 4.

For the simulated MNAR data in Phase I, only NSMI gives approximately unbiased estimates of the mean of $\boldsymbol{Y}$ and the regression coefficients. The precision increases as the sampling proportion in Phase II increases. Note that even randomly sampling five percent of the nonrespondents in Phase I is enough to distinguish the NSMI results from other competing methods. However, if data are collected on a small percentage of nonrespondents in Phase I, the NSMI yields estimates with large variance, and hence increased average lengths of the 95% confidence intervals. Please refer to the online supplementary material for additional simulation studies to examine how the performance of the NSMI method depends on the proportion of missingness.

The performance of different methods when the Phase II missing-data mechanism is MNAR is presented in the online supplementary material and is examined now. When the missing-data mechanism in Phase I is MAR or MCAR, IL1 is the only method that gives approximately unbiased estimates; in this case, both methods utilizing additional data from Phase II (IL2, and NSMI) are biased, because the missingness mechanism in Phase II is MNAR. When both Phases I and II's missingness mechanisms are MNAR, no method gives unbiased estimators for any of the parameters of interest. Please refer to the online supplementary material for additional studies comparing NSMI with alternative methods.

6. Application to Motivating Example

The proposed method will now be applied to the QOL dataset. For illustration purposes, the results for the resources subscale are presented. This is to estimate the mean resources and the regression of resources on gender (male versus female), age (in years),

Table 4. Empirical bias, root mean square error, and 95% CI coverage probabilities under MNAR in Phase I and with different Phase II sampling proportions (MCAR) (1,000 replications).

		$\pi = .05$				$\pi = .15$				$\pi = .25$				$\pi = .50$			
		μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z	μ	β_0	β_x	β_z
Bias ($\times 10^4$)	BD	−32	−11	−5	0	39	21	−18	8	16	4	−9	−16	16	10	−6	4
	CC1	7933	2816	−1083	−1089	7977	2849	−1098	−1055	7961	2827	−1081	−1082	7967	2836	−1078	−1081
	CC2	7336	2522	−929	−933	6289	2066	−705	−665	5286	1661	−517	−516	3150	935	−231	−227
	IL1	2795	2814	−1083	−1085	2866	2850	−1095	−1061	2840	2829	−1082	−1081	2834	2828	−1074	−1073
	IL2	2503	2522	−930	−932	2083	2067	−706	−663	1673	1662	−516	−520	940	935	−232	−230
	NSMI	−65	−45	−21	10	6	−11	−20	22	26	15	−8	−20	16	10	−1	−1
RMSE ($\times 10^4$)	BD	600	323	321	333	604	307	329	333	607	307	341	337	600	325	338	336
	CC1	7958	2845	1158	1169	8001	2874	1171	1135	7984	2854	1161	1161	7991	2866	1154	1159
	CC2	7363	2553	1013	1023	6320	2097	807	776	5324	1700	651	652	3212	1002	440	439
	IL1	2860	2844	1161	1168	2928	2877	1173	1144	2903	2847	1165	1163	2899	2859	1153	1155
	IL2	2576	2555	1017	1025	2167	2099	812	777	1784	1701	653	660	1121	1003	444	445
	NSMI	1171	1058	1059	1145	733	526	592	584	681	428	483	487	624	373	398	393
Coverage	BD	95.1	94.5	95.2	94.2	94.7	95.2	94.9	95.2	94.6	95.8	94.1	94.5	94.7	94.8	94.4	94.2
	CC1	0.0	0.0	23.8	24.0	0.0	0.0	23.6	27.5	0.0	0.0	24.7	24.1	0.0	0.0	26.1	23.0
	CC2	0.0	0.0	36.4	36.7	0.0	0.0	55.3	59.7	0.0	0.2	73.7	72.5	0.1	25.1	89.2	89.8
	IL1	0.6	0.0	27.8	27.8	0.5	0.0	27.8	30.5	0.3	0.0	26.8	28.8	0.2	0.0	28.8	26.0
	IL2	1.6	0.0	39.5	39.7	7.3	0.0	56.8	61.9	23.8	0.3	73.5	73.6	67.3	24.8	89.2	89.9
	NSMI	94.3	94.5	95.5	93.7	95.3	95.9	94.5	93.9	95.4	96.1	94.7	94.8	95.4	94.8	95.0	94.1

race, and education. Race was dichotomized as white versus nonwhite, and education was dichotomized as high school and above versus education below high-school level.

All covariates are fully observed, whereas 147 out of 750 subjects have missing values in resources in Phase I. In Phase II, 39 out of the 147 nonrespondents provided data. Since the Phase II data collection was done within three months of Phase I, it was assumed that resources remained unchanged from Phase I. In implementing the NSMI method, we make the assumption that, among the 147 nonrespondents, the missingness after Phase II is MAR, thus meeting the conditions for NSMI. The validity of other competing methods rests on the mechanism that generates the missing data in Phase I. For instance, if the missing-data mechanism in Phase I is MCAR, then both CC and IL methods provide valid estimates. However, if the missingness in Phase I is MNAR (as suggested by Bonetti et al. 1999 and Fielding et al. 2009), then CC and IL methods will fail to give an unbiased estimation.

For all imputation methods, the fully observed resources measured at the mean age of 22 years are used in the imputation model, but not in the analysis model; this is because the resource scale measured at that age serves as a good predictor for the resources at the mean age of 33, but is not of direct interest in the analysis model (Meng 1994; van Buuren et al. 1999). The results from five methods are shown in Table 5. With respect to the modeling of resources as a function of gender, age, race, and education, NSMI shows a weaker negative association of race with resources compared with the other four methods. In particular, the NSMI method did not reveal a statistically significant association of race with resources, in contrast to the other methods, where whites had significantly greater resources. Age also had a somewhat weaker positive association with resources, although this relationship was not significant in any of the approaches. Those with higher levels of education and females had higher levels of resources, although these relationships were not statistically significant in any of the methods, nor did they differ systematically across the methods.

7. Discussion

Two-phase sampling has been proposed and used in surveys with nonresponse for five decades. However, little research has been done to show the benefit of nonresponse subsampling; traditional methods (i.e., weighting) also fail to make full use of the additional data collected from two-phase sampling. This article proposes an NSMI method to analyze data from NTS. The proposed method yields valid estimates when the missing-data mechanism in the subsample of initial nonrespondents is MAR, regardless of the missing-data mechanism in Phase I. The simulation studies also show that it is beneficial to use the NTS scheme, even when collecting data from only a small proportion of the nonrespondents.

Previous studies suggest that the missing-data mechanism in QOL outcomes was probably not MCAR (Bonetti et al. 1999; Fielding et al. 2009). Therefore, NSMI is considered in this applied example, which utilizes two-phase sampling to obtain data from a subsample of the initial nonrespondents in the Children in the Community study. Using the proposed NSMI method, white race was not found to be significantly associated with

Table 5. Regression analysis of a QOL dataset: resources.

	CC1					CC2					IL1				
	Est.	S.E.	LCL	UCL	*p*-value	Est.	S.E.	LCL	UCL	*p*-value	Est.	S.E.	LCL	UCL	*p*-value
Outcome															
Mean	77.26	0.69	75.9	78.62	<.001	77.20	0.67	75.89	78.51	<.001	77.03	0.72	75.6	78.46	<.001
Regression															
Intercept	62.09	6.07	50.19	73.99	<.001	63.85	5.9	52.28	75.43	<.001	62.32	6.15	50.19	75.46	<.001
Sex (male vs. female)	−2.48	1.38	−5.19	0.23	0.056	−2.70	1.33	−5.30	−0.10	0.043	−2.44	1.31	−5.02	0.14	0.054
Race (white vs. nonwhite)	5.64	2.47	0.80	10.49	0.038	5.35	2.36	0.72	9.98	**0.041**	**5.00**	**2.44**	**0.19**	**9.81**	**0.044**
Education (≥HS* vs.<HS)	1.83	1.45	−1.02	4.68	0.104	1.61	1.40	−1.14	4.36	0.125	1.71	1.45	−1.14	4.56	0.119
Age	0.46	0.26	−0.06	0.97	0.053	0.39	0.25	−0.10	0.89	0.060	0.47	0.28	−0.08	1.02	0.056

	IL2					NSMI				
	Est.	S.E.	LCL	UCL	*p*-value	Est.	S.E.	LCL	UCL	*p*-value
Outcome										
Mean	77.09	0.67	75.77	78.42	<.001	77.07	0.71	75.67	78.47	<.001
Regression										
Intercept	64.64	5.73	53.4	75.88	<.001	66.84	5.64	55.77	77.91	<.001
Sex (male vs. female)	−2.27	1.41	−5.07	0.52	0.054	−2.35	1.28	−4.87	0.16	0.053
Race (white vs. nonwhite)	**4.84**	**2.25**	**0.43**	**9.26**	0.042	3.46	2.31	−1.07	7.99	0.067
Education (≥HS* vs.<HS)	2.04	1.33	−0.57	4.66	0.063	1.74	1.32	−0.85	4.34	0.094
Age	0.36	0.25	−0.13	0.84	0.075	0.33	0.24	−0.15	0.80	0.084

*HS: high school. LCL: lower confidence limit. UCL: upper confidence limit.

increased resources, while other alternatives suggested a significant association between race and resources.

By exploring the relationship between missing values and observed data, the NSMI methods use the information of the fully observed variables and improve the efficiency of the estimation. The method of multiple imputation by chained equations provides a valid way of utilizing the information in other variables when imputing the missing values. The NSMI method not only provides a valid estimation of the marginal distribution of the outcome (e.g., mean), but also of the conditional distribution of the outcome on covariates (e.g., regression).

Missing data are ubiquitous and all methods for handling missing data rely on untestable assumptions. NTS provides a valid way to relax these untestable assumptions in part. Ideally, Phase II sampling takes a random sample of Phase I nonrespondents. However, these random subsamples may still be subject to nonresponse. In cases when the sampling yields a missing-data mechanism of MAR for Phase I nonrespondents, the proposed NSMI method is valid, regardless of the first-stage mechanism. In the event that both first- and second-stage missing-data mechanisms are MNAR, neither NSMI nor any multiple-imputation methods that ignore missing-data mechanisms are free of bias. Of course, in practice, assessing MNAR directly is not typically possible—the motivation for the NSMI approach is that, if the Phase I missingness mechanism is strongly MNAR, the Phase II missingness may be less so, because the Phase I nonrespondents may share some common characteristics that make the NS-MAR assumption plausible.

The NTS scheme considered in this article involves collecting data from nonrespondents. This is challenging in practice, but may be achieved by giving an abridged version of the questionnaire, by giving incentives for response, or by using other advanced survey techniques, such as tailoring the questionnaire to the interviewees (Groves and Couper 1998). In the second-stage subsampling within a fixed budget, there is a balance between reducing the nonresponse rate and subsampling more subjects, because by focusing on a moderate number of nonrespondents, it is possible to obtain a high response rate and therefore reduce the nonresponse bias (Elliott et al. 2000). This aspect of the problem is currently under investigation.

Finally, it should be noted that use of the NSMI approach is not fail-safe. As the simulation studies show, if Phase I missingness is MCAR, there is no gain in using the NSMI approach; if Phase I is MCAR or MAR, and Phase II is MNAR, substantial bias can be introduced relative to MAR methods that ignore the Phase II data. While Phase I MAR and MNAR mechanisms cannot be distinguished from observed data, some evidence for Phase I MCAR can be deduced from the observed data. Hence, methods that consider the evidence for MCAR and 'trade off' Phase I versus Phase II imputation may be desirable to enhance robustness under all different mechanisms. In addition, follow-up nonresponse designs that devote more intensive effort to minimizing Phase II MNAR though use of techniques that may not be practical or cost-effective to implement during Phase I data collection (use of targeted incentives, expensive but high response rate data-collection modes such as face-to-face interviews) might be implemented to make NSMI assumptions more plausible. Future research is needed into analytic methods both to improve robustness to NSMI assumption failures and to consider data-collection methods that better meet NSMI assumptions.

8. References

Amemiya, T. 1984. "Tobit Models, a Survey." *Journal of Econometrics* 24: 3–61. Doi: http://dx.doi.org/10.1016/0304-4076(84)90074-5.

Bonetti, M., B.F. Cole, and R.D. Gelber. 1999. "A Method-of-Moments Estimation Procedure for Categorical Quality-of-life Data with Nonignorable Missingness." *Journal of the American Statistical Association* 94: 1025–1034. Doi: http://dx.doi.org/10.1080/01621459.1999.10473855.

Chen, H., P. Cohen, S. Kasen, R. Dufur, E.M. Smailes, and K. Gordon. 2004. "Construction and Validation of a Quality of Life Instrument for Young Adults." *Quality of Life Research* 13: 747–759. Doi: http://dx.doi.org/10.1023/B:QURE.0000021700.42478.ab.

Cohen, P., T.N. Crawford, J.G. Johnson, and S. Kasen. 2005. "The Children in the Community Study of Developmental Course of Personality Disorder." *Journal of Personality Disorder* 19: 466–486.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via EM Algorithm." *Journal of the Royal Statistical Society, Series B* 39: 1–38. Available at: http://www.jstor.org/stable/2984875 (accessed May 2016).

Elliott, M.R., R.J.A. Little, and S. Lewitzky. 2000. "Subsampling Callbacks to Improve Survey Efficiency." *Journal of the American Statistical Association* 95: 730–738. Doi: http://dx.doi.org/10.1080/01621459.2000.10474261.

Fielding, S., P.M. Fayers, and C.R. Ramsay. 2009. "Investigating the Missing Data Mechanism in Quality of Life Outcomes: A Comparison of Approaches." *Health and Quality of Life Outcomes* 7: 57. Doi: http://dx.doi.org/10.1186/1477-7525-7-57.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Hansen, M.H. and W.N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association* 41: 517–529. Doi: http://dx.doi.org/10.1080/01621459.1946.10501894.

Harter, R.M., T.L. Mach, J.F. Chapline, and J.D. Wolken. 2007. "Determining Subsampling Rates for Nonrespondents." In Proceedings of ICES-III, June 18–21, 2007. 1293–1300. Montreal, Quebec, Canada. Available at: http://50.205.225.65/meetings/ices/2007/proceedings/ICES2007-000197.PDF (accessed May 2016).

Heitjan, D. and D.B. Rubin. 1991. "Ignorability and Coarse Data." *The Annals of Statistics* 81: 2244–2253. Available at: http://www.jstor.org/stable/2241929 (accessed May 2016).

Heckman, J.J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for such Models." *Annals of Economic and Social Measurement* 5: 475–492. Available at: http://www.nber.org/chapters/c10491.pdf (accessed May 2016).

Little, R.J.A. 1993. "Pattern-Mixture Model for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88: 125–134. Doi: http://dx.doi.org/10.1080/01621459.1993.10594302.

Little, R.J.A. 1994. "A Class of Pattern-Mixture Models for Normal Incomplete Data." *Biometrika* 81: 471–483. Doi: http://dx.doi.org/10.1093/biomet/81.3.471.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley.

Little, R.J.A. and N. Zhang. 2011. "Subsample Ignorable Likelihood for Regression Analysis with Missing Data." *Journal of the Royal Statistical Society, Series C* 60: 591–605. Doi: http://dx.doi.org/10.1111/j.1467-9876.2011.00763.x.

Meng, X.-L. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Sciences* 9: 538–573. Available at: http://www.jstor.org/stable/2246252 (accesssed May 2016).

Nandram, B. and J.W. Choi. 2002. "Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty about Ignorability." *Journal of the American Statistical Association* 97: 381–388. Doi: http://dx.doi.org/10.1198/016214502760046934.

Nandram, B. and J.W. Choi. 2010. "A Bayesian Analysis of Body Mass Index Data from Small Domains under Nonignorable Nonresponse and Selection." *Journal of the American Statistical Association* 105: 120–135. Doi: http://dx.doi.org/10.1198/jasa.2009.ap08443.

The R Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Available at: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (accessed June 2016).

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–592. Doi: http://dx.doi.org/10.1093/biomet/63.3.581.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Srinath, K.P. 1971. "Multiphase Sampling in Nonresponse Problems." *Journal of the American Statistical Association* 66: 583–620. Doi: http://dx.doi.org/10.1080/01621459.1971.10482310.

Statistics Canada. 2011. National Household Survey. Available at: http://www12.statcan.gc.ca/nhs-enm/index-eng.cfm (accessed June 1, 2015).

Van Buuren, S., H.C. Boshuizen, and D.L. Knook. 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine* 18: 681–694. Available at: http://www.stefvanbuuren.nl/publications/Multiple%20imputation%20-%20Stat%20Med%201999.pdf (accessed May 2016).

Van Buuren, S.‘ and K. Groothuis-Oudshoom. 2011. "Mice: Multivariate Imputation by Chained Equation in R." *Journal of Statistical Software* 45: 1–67. Available at: http://doc.utwente.nl/78938/.

Received February 2015
Revised July 2015
Accepted August 2015