JOS

Journal of Official Statistics

SCB September 2019

Journal of Official Statistics vol. 36, 1 (Mar 2020)

sciendo

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 1–24, http://dx.doi.org/10.2478/JOS-2020-0001

A Framework for Official Temporary Population Statistics

Elin Charles-Edwards[1], *Martin Bell*[1], *Radoslaw Panczak*[1], *and Jonathan Corcoran*[1]

There is considerable demand for official statistics on temporary populations to supplement statistics on resident and working populations. Progress has been slow, with temporary population statistics not part of the standard suite of measures produced by national statistical offices. This article adopts the framework for official statistics proposed by Raymer and colleagues as a guide to aspects relating to society, concepts, data, processing, outputs and validation. The article proposes a conceptual framework linking temporary population mobility, defined as a move more than one night in duration that does not entail a change in usual residence, and temporary populations. Using Australia as an example, we discuss various dimensions of temporary mobility that complicate its measurement. We then report the outcomes of a survey of user needs for temporary population statistics along with a desktop review of OECD countries to identify the best formulation of temporary population statistics, and current international practice respectively. The article concludes by proposing two related concepts for temporary populations: population present and person-time, which overcome a number of issues currently impeding progress in this area and discuss their potential implementation.

Key words: Seasonal; ambient; mobility; estimates.

1. Introduction

Accurate and timely population statistics are fundamental to understanding society. Official statistics take a range of forms, reflecting different population concepts, varying data availability, and changing societal needs (Raymer et al. 2015). The two most commonly produced forms of population statistics are de jure counts, which measure the population usually resident in an area at a particular point in time, and de facto counts, which capture the population present in an area. Both de jure and de facto estimates tend to capture a snapshot of populations at a single point in time, and miss the daily, weekly and seasonal flux in populations that are driven by diurnal and temporary population movements. There is an established demand for estimates that account for short-term shifts in populations to serve as the denominator for health, crime and other statistics, to

[1] The University of Queensland, St Lucia, Brisbane, Australia, 4072. Emails: e.charles-edwards@uq.edu.au, martin.bell@uq.edu.au, r.panczak@uq.edu.au and jj.corcoran@uq.edu.au

Acknowledgments: The authors wish to thank the Australian Bureau of Statistics (ABS) for facilitating the collection of data on which the paper is based. However, the interpretations of the analysis are solely those of the authors and do not necessarily reflect the views and opinions of the ABS or any of their employees. This research was funded by the Australian Government through the Australian Research Council Linkage project scheme, LP160100305.

assist in emergency preparedness and response, and for the planning and provision of local goods and services (Smith 1989; Deville et al. 2014; Kounadi et al. 2018). The United Nations (2017) has called for estimates of temporary or service populations "if a population present count or usual resident population count does not accurately represent the demand for, or provision of services in a country or part of a country" (United Nations 2017, 180). Improved data availability, particularly from mobile phones and related technologies, has led to an increased interest in enumerating short-term population change. However, such estimates have been produced outside of official statistical frameworks, often for a single area, and intermittently. Despite their importance as a complement to common population statistics, few statistical agencies currently produce official temporary population estimates, and it remains unclear how such estimates are best conceptualised and generated.

This article explores issues associated with the production of official temporary population statistics. The production of estimates is complex due to variability in the nature and dynamics of the underlying mobility driving short-term shifts in population numbers, a lack of clarity with respect to user needs, limited data availability and nascent conceptual development. We seek to advance the field of temporary population estimates by systematically exploring these issues, and through this process, provide greater conceptual clarity on the formulation of estimates. We begin in Section 2 with a discussion of the spatiotemporal dynamics of various forms of temporary population mobility and their impacts on temporary populations, using Australia as a case study. Understanding the dynamics of mobility is an essential prerequisite to the production of a robust and useful suite of measures of temporary populations. While statistics must be responsive to national contexts, Australia offers a useful testbed, since it features both the diverse forms of mobility and the differing types of data that are commonly found in developed countries.

In Section 3, we report the results of a survey of Australian government officials and planners as to their requirements for temporary population statistics. We then examine currently published outputs of temporary population statistics in a number of OECD countries in Section 4, before concluding by outlining a consistent conceptual basis for temporary population statistics.

2. Temporary Mobility in Australia

Temporary population mobility can be defined *as displacements from a usual residence which are one night or more in duration but do not entail a change in usual residence*. Academic interest in temporary population mobility is longstanding (Chapman and Prothero 1983). Early research was directed at temporary mobility in the developing world, where temporary mobility was viewed as both a customary process and as a means of adapting to transformations brought about by colonisation and industrialisation (Chapman 1978; Chapman and Prothero 1983; Prothero and Chapman 1985; Mitchell 1969; Taylor 1986). Not until later did researchers in the developed world turn to the study of temporary population mobility, spurred by Zelinsky's (1971, 230) 'Hypothesis of the Mobility Transition', which postulated an advanced society characterised by '[V]igorous accelerating circulation, particularly the economic and pleasure-oriented, but other varieties as well'. There is now a considerable literature examining various forms of temporary mobility in developed world settings.

In Australia, as elsewhere, temporary movements are undertaken in response to a range of demographic, social, economic and environmental stimuli, and are facilitated by transport technology (Figure 1). Movements are conveniently classified according to whether they are undertaken for purposes of production or consumption at the destination (Williams and Hall 2000). Production-related moves are exemplified by fly-in fly-out/drive-in drive-out mobility in the resources sector, long-distance commuting, seasonal agricultural mobility along with short-term business trips. Consumption-related mobility includes short-term tourist stays, visits to second homes, the extended cross-continental sojourns of elderly grey nomads, mobility to access health care, the mobility of indigenous peoples to participate in customary activities and visits to friends and relatives (Bell and Ward 2000). Critically, these different forms of mobility have distinct spatial and temporal signatures, the interaction of which underpin the short-term fluctuation in temporary populations stocks.

Bell (2004) identified nine dimensions of temporary mobility that vary for different movement types. These are movement intensity, spatial impact, connectivity, movement distance, spatial circuits, duration, seasonality, frequency, and periodicity. Table 1 illustrates variation in these nine dimensions for eleven types of mobility that are nationally or locally significant in Australia (Bell 2001; Charles-Edwards et al. 2008). We now describe these in turn.

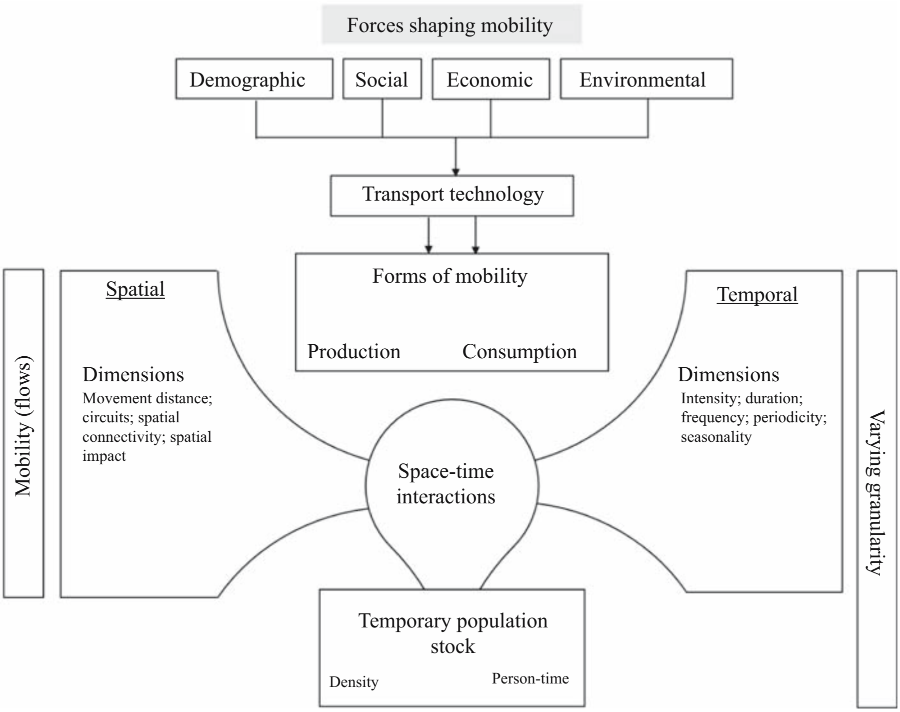


Fig. 1. Conceptualising the mobility system.

Table 1. Movement types and dimensions of mobility, Australia.

	DIMENSIONS OF MOBILITY								
	SPATIAL					TEMPORAL			
Movement type	Intensity	Spatial Impact	Connectivity	Circuits	Distance decay	Duration	Seasonality	Frequency	Periodicity
International tourism	High	Moderate-high	Concentrated	Complex	Low	Days-months	Moderate	Low	Low
Domestic tourism	High	Moderate-high	Concentrated	Simple	Medium	Days-weeks	High	Moderate	Moderate
Business travel	High	Low	Concentrated	Simple	Low	Days-weeks	Low	High	Moderate
Visits to friend and relative	High	Low	Dispersed	Simple	Low	Days-weeks	Moderate	High	High
Children in shared care	Moderate/ High	Low	Dispersed	Simple	High	Days	Low	High	High
Second homes	Moderate	Moderate-High	Concentrated	Simple	High	Days-weeks	Moderate	Moderate	HIgh
Fly-in fly-out/ Drive-in drive-out mining	Low	High	Concentrated	Simple	Low	Weeks	Low	High	HIgh
Grey nomads	Low	Low-moderate	Dispersed	Complex	Low	Weeks-months	High	Low	Low
Long distance business commuting	Low	Low	Concentrated	Simple	Low	Days-weeks	Low	High	High
Seasonal agricultural labour	Low	Moderate	Concentrated	Complex	Low	Weeks-months	High	Low	Low
Indigenous mobility	Low	Moderate	Dispersed	Complex	Low	Weeks-months	Moderate	High	High

Intensity refers to the overall level of movement within a population. This dimension allows numerically significant forms of mobility to be identified and also provides insights into the composition of temporary mobility in Australia. While data are incomplete, available information suggests that tourism, both international and domestic, business travel, and visits to friends and relatives have the highest intensity of temporary mobility in Australia (Charles-Edwards and Bell 2015). In 2018, there were more than 8.1 million short-term international visitors arrivals visiting for holidays, to see friends and relatives, on business, and for conferences or conventions (ABS 2019) and an estimated 42 million domestic tourism trips, 23 million business trips, and 33 million trips to visit friends and relatives taken by Australians aged 15 and older in the year ending September 2018 (Tourism Research Australia 2018). A distinct form of mobility tied to family is that of children in shared care arrangements following parental separation. One in five Australian children will experience parental separation before the age of 17 (Halford 2018), with 49% of children staying overnight with their non-resident parent in 2012–2013 (ABS 2015). Other forms of mobility are locally important. Mobility tied to second homes is significant in several coastal areas, with second homes accounting for up to 50% of the housing stock (Paris et al. 2014). Fly-in fly-out and drive-in drive-out mining, whereby workers in the resource sector live remotely and travel to the mine site for extended shifts (two weeks onsite, one week off, for example), are a significant component of the population in resource regions, such as the Bowen Basin coal mining region of Queensland (QGSO 2018) or Western Australia's remote north-west (Houghton 1993). Other forms of long-distance commuting, for instance travel between regions and cities on a weekly basis, also occur, but there are no reliable statistics on the intensity of these movements. Grey nomads, older people who travel the country in recreational vehicles, are estimated at 2% of the Australian population (Davis 2011) and concentrate in northern Australia during the winter months. Other forms of mobility, such as seasonal agricultural mobility undertaken by international working holidaymakers and itinerant groups, are an important component of the workforce in many rural areas (Hanson and Bell 2007). Indigenous mobility is another significant form of movement, with extensive circuits of movement tied to customary activity and access to services across remote and rural Australia (Taylor and Bell 2012).

The demographic impact of temporary population mobility is determined by its intensity, but also by the degree to which temporary population mobility redistributes population across the settlement system or its *spatial impact*. Borrowing from conventional measures of permanent migration, this can be captured by measures of movement effectiveness, that is, the degree to which flows from one area to another are balanced by counter-flows in the opposite direction. A system with unidirectional flows will be highly effective at redistributing populations, while systems with balanced flows will result in minimal redistribution of the population, even though there may be a high intensity of movement. Although there is substantial evidence demonstrating the spatial concentrations generated by particular types of temporary mobility (see, for example Bell and Ward 1998), little work has been undertaken to directly measure movement effectiveness among temporary populations. That said, some assessment of particular types of movement can be made *a priori*. For example, seasonal agricultural labour, fly-in fly-out or drive-in drive-out mining are likely to be highly effective at redistributing populations due to the spatial concentration of the activities that trigger these moves. By contrast, business travel and visits to friends and relatives tend to more closely reflect national settlement patterns and are more likely to be balanced by counter-flows.

Another spatial aspect is *connectivity*, also termed spatial focusing (Plane and Mulligan 1997). In any system of interregional mobility, the magnitude of flows varies between origin and destination pairs. This reflects both the size of the populations at origins and destinations along with the distance between them, and also indicates the strength of functional linkages (Bell et al. 2002). While empirical evidence is scant, a high degree of spatial focusing at destinations is likely when mobility is triggered to access spatially concentrated goods or services. Second home mobility, for example, is concentrated in high amenity areas, often within a few hours of major population centres (Back and Marjavaara 2017).

With respect to movement *distance*, temporary movements in Australia involve longer distances on average than permanent migration (Bell and Brown 2006); however, the rate of distance decay varies according to the purpose of the move. Tourism movements (McKercher 2018) and second home mobility (Müller et al. 2004), exhibit high distance decay, whereas visits to friends and relatives and fly-in fly-out mining are less affected. Distance can also be a driver of temporary mobility. For example, mobility associated with fly-in fly-out/ drive-in drive-out mining substitutes for permanent moves in remote regions, where the costs of establishing permanent settlements are prohibitive (Houghton 1993).

The final spatial dimension of mobility concerns movement *circuits*. While some forms of mobility involve a simple oscillation between a single origin and destination (e.g., travel to and from second homes), others involve complex itineraries linking multiple destinations (Bell 2001). The mobility of seasonal agricultural workers is one such example, with working holiday makers, retirees and permanent itinerants following a series of harvest trails to meet seasonal demand for horticultural labour across regional Australia (Hanson and Bell 2007).

Shifting to the temporal dimensions of mobility, temporary movements are of variable *duration,* ranging from a single night (i.e., business travel) to sojourns extending over many months (i.e., grey nomads). Duration can be measured with respect to the length of absence from an origin, or the length of a visit at the destination. The majority of movements undertaken in Australia are of short duration, with around half of all movements undertaken within Australia being fewer than two nights in duration. Longer trips, however, make-up three-quarters of all nights away from home (Tourism Research Australia 2018), and thus have a disproportionate impact on temporary population stocks at origins and destinations.

Seasonality is a key aspect of mobility that differentiates it from permanent migration. Institutional seasonality, which reflects the timing of school and public holidays and religious festivals, such as Easter and Christmas, impact the timing of both discretionary tourism and business flows. Natural seasonality, driven by climatic factors, is evident in the mobility of tourists, grey nomads and seasonal agricultural labour. The areal expanse of Australia, spanning multiple climatic zones, produces a north-south gradient in seasonality. Visits peak in the north of the country during the southern hemisphere winter, or “dry season”, at which time climate is comfortable and roads are accessible. By contrast, visits in southern Australia peak over the summer months (Charles-Edwards and Bell 2015). Seasonality means that the timing of peaks in temporary populations varies across the country.

The *frequency* of mobility refers to the number of moves undertaken by a person in a fixed interval. Some forms of mobility are undertaken at frequent intervals, such as the movements associated with long-distance business commuting or second home ownership. Other movements are more sporadic, for example, occasional tourist trips. Like a number of other dimensions of mobility, frequency can be measured with respect to absences from an origin, and may involve multiple destinations for different purposes, or may reflect frequent visits to a particular destination. Movement frequency can have implications for estimates of temporary populations, as it leads to a divergence between the number of *moves* and the number of *movers* as the measurement interval increases. Data from the Australian National Visitor Survey reveal that one quarter of Australians aged 15 and over make an overnight trip in any given four week period. Of these, only 77% make a single trip, but 45% of trips are made by repeat movers (Tourism Research Australia 2018). This suggests that repeat movers account for a disproportionate share of temporary populations.

Periodicity combines information on frequency and duration to capture the sequences of movements (Taylor and Bell 2012). Fly-in fly-out mobility, mobility tied to second homes and many forms of customary mobility undertaken by indigenous Australians can all demonstrate a high degree of periodicity that differentiates them from other forms of mobility of similar duration. Periodicity may be of interest to planners and policy makers as it can impact the level of place attachment and different service requirements of visitors at destinations, with regular visitors having different demands to those visiting a region on a one-off basis.

The nine dimensions of mobility proposed by Bell (2004) highlight the complex spatiotemporal behaviours that characterise temporary forms of mobility in Australia. Useful progress has been made in developing robust measures that capture these multiple dimensions of mobility (see, for example Charles-Edwards and Bell 2015; Taylor and Bell 2012), but implementation is commonly hampered by a lack of consistent, reliable data. Moreover, even the more straightforward metrics, such as intensity and duration, depend on whether the movement is measured at the origin or destination. Equally challenging is whether to measure moves or movers, more or less identical when the observation interval is short, but divergent as the interval lengthens due to repeat mobility. The dimensions described above provide important insights into the dynamics of mobility, and the processes that generate shifts in the population surface from day to day, week to week and month to month, but they are not necessarily the measures that are best suited to the needs of users, nor are they readily estimated by statistical agencies.

3. Survey of User Needs

An understanding of user needs is fundamental to the production of official statistics (Raymer et al. 2015). While there have been long-standing calls for the estimation of temporary populations (Smith 1989; Cook 1996, 1998; Hugo and Harris 2013), little is known about user requirements with respect to population coverage, geography, the frequency of estimates, population characteristics and the types of metrics that may be of most use. To gain insight into user needs, an online survey was distributed with support from the Australian Bureau of Statistics to a range of stakeholders including government agencies, state statistical offices, local government associations and the private sector. The survey was initially distributed to a list of over 100 individuals and organisations, with users encouraged to share the survey link with others in their network. A total of 57 responses were received. Most respondents worked in the government sector, with 25 employed in local government, and a further 26 in state or federal agencies. Four respondents were private sector employees, while one was employed in academia. One respondent did not state their employment sector. The survey asked fifteen questions relating to the potential uses of temporary population estimates, desired population coverage, the temporal resolution of estimates, the output geography and population characteristics of interest. We note that the sample is largely comprised of local government planners and officials and recognise that other users may have different needs.

3.1. Why are Estimates of Temporary Populations Needed?

Fundamental to the creation of official statistics is an understanding of the need for, and utility of, any output statistics. Respondents were asked an open-ended question on the need for estimates of temporary populations. All respondents (57) answered this question. Responses were manually coded using an inductive approach. Codes were first created based on a 50% sample of responses. These codes were then reapplied to this sample to validate before being applied to the remaining responses. Results are shown in Figure 2. The major application of temporary population estimates was seen to lie in better planning and provision of local goods and services (36 of 57 responses) to cater for peak and seasonal variations in demand. A second commonly cited purpose was to provide a more robust basis for the equitable distribution of Commonwealth Government financial resources to local government authorities (10/57). These are currently allocated using a formula based principally on de jure population estimates prepared by the Australian

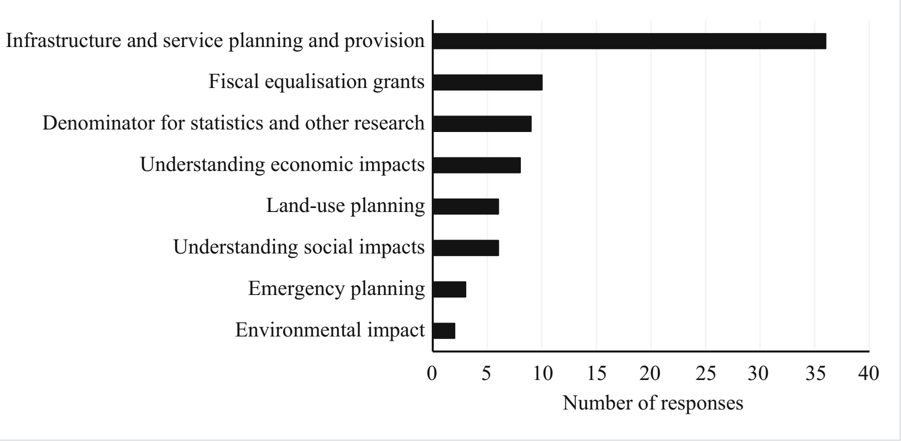


Fig. 2. Need for estimates.

Bureau of Statistics, which are seen as disadvantaging local authorities that host large temporary populations.

Other responses included the need for appropriate denominators for crime and health statistics, a theme that has emerged in the academic literature, as well as information to better understand the nature of the temporary populations themselves (e.g., fly-in-fly-out populations). A number of respondents suggested that estimates were needed to better understand the economic (8/57), social (6/57) and environmental impacts (2/57) of temporary populations. This included the need for statistics to better model labour market impacts and local economic effects, including the impact of temporary population on housing affordability arising from the short-term letting market. Land-use planning, particularly as it relates to land supply, and emergency planning and preparedness, were also nominated. The results overwhelmingly focused on local impacts, perhaps unsurprising given the high proportion of local government officials among respondents, but there was also a desire for statistics with wider geographic coverage for modelling and research purposes.

3.2. Who Should be Captured?

Respondents were asked to identify the groups of visitors that are significant in their region of interest from a closed list, with multiple responses accepted (Figure 3). Domestic tourists (48/57) and international tourists (44/57) topped the list, reflecting the high overall intensity of these movements. Second home owners (36/57) were also of strong interest, followed by grey nomads (33/57), fly-in fly-out (30/57) and drive-in drive-out workers (29/57). Indigenous peoples, seasonal and itinerant workers, homeless populations, people visiting friends and relatives, and international workers received fewer mentions. Two respondents noted that a single index capturing all forms of temporary population, irrespective of motives, would be most valuable.

In addition to information on overnight visitors, respondents expressed a need for estimates of daytime populations, including commuters (28/57) and those travelling for

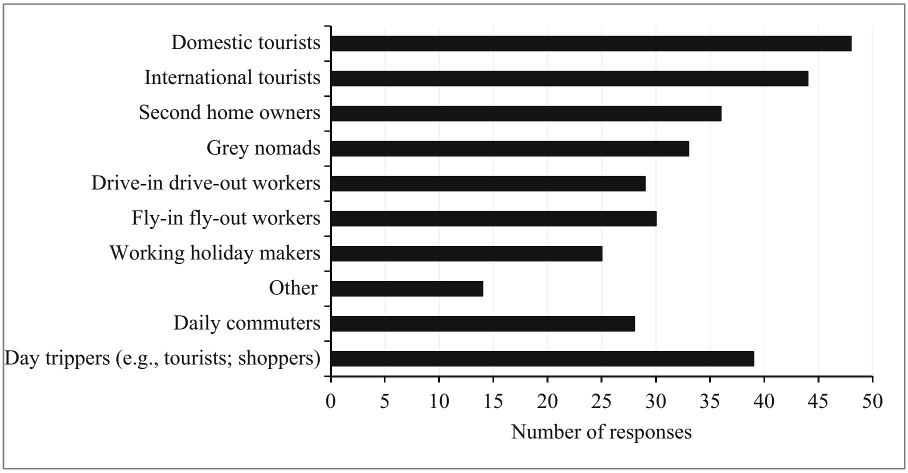


Fig. 3. Types of visitors.

consumption-related purposes (39/57). Most of these preferred separate estimates of daytime as against overnight populations (38/55). Respondents were also asked about the need to identify the purpose of the move. A total of 47 of 57 respondents wanted to distinguish tourists, while 46 wanted to separately identify business travellers. Thirty-eight respondents wanted to know if visitors accessed goods and services at the destination and 45 thought it was important to distinguish between occasional and repeat visitors (e.g., second home owners). What emerges from these results is the diverse composition of temporary populations across Australian regions. For official statistical purposes, targeting population subgroups for estimation may produce locally useful results, but will not serve as a national standard.

3.3. Temporal Framework

A key feature of temporary populations is their variation over time, therefore a single point estimate is unlikely to adequately represent the temporary population of an area. Respondents were asked about the temporal variations in populations they were most interested in capturing from a closed set of responses (Figure 4). Seasonal variation in population numbers was the most common response (49/56), followed by variations between weekday and weekend populations (35/56). Estimates capturing holiday populations were nominated by 34 respondents. Daily estimates were nominated by 16. Significantly, there was little desire for estimates capturing variation over the course of a single day (12/56). Other time periods were nominated by 13 respondents; these included estimates capturing monthly variations (3), periods coincident with

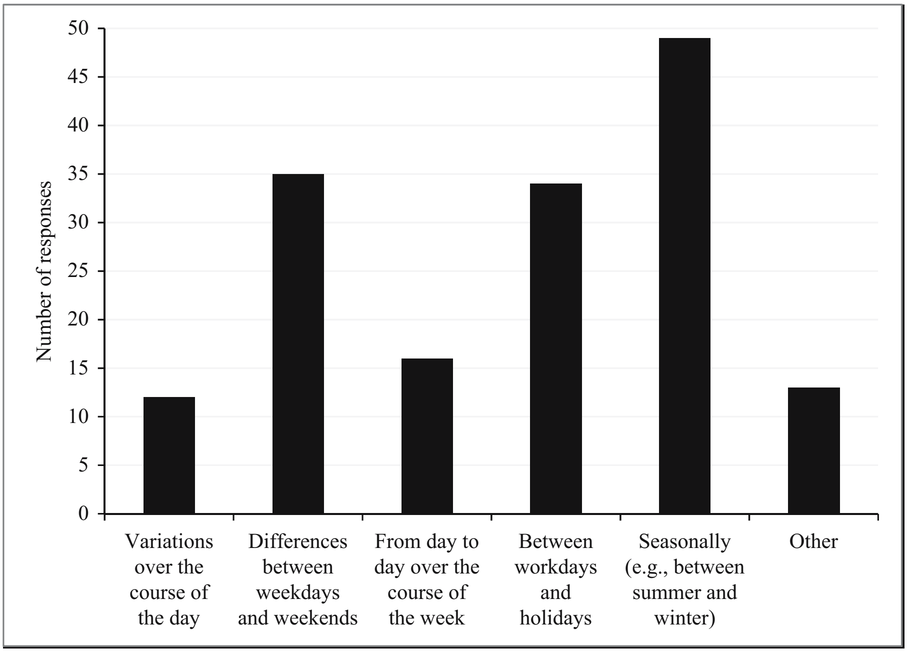


Fig. 4. Temporal variation to be captured.

agricultural harvest seasons (2), estimates timed to capture specific events (2), and single point estimates to facilitate comparison with other Australian population statistics. The results suggest that daily estimates would best meet the needs of most users; however, the data and processing needed for continuous estimates are significant. Monthly estimates offer a potential compromise providing a balance between temporal specificity and data demands.

3.4. Geography

Respondents were asked about the geographic scale at which estimates were needed. Results are summarised in Figure 5, differentiating three levels in the hierarchy of spatial units that make up Australia's regional statistical framework: States and territories (of which there are nine), Local Government Areas (LGAs – 563), and Statistical Areas Level 2 (SA2s – 2310). Local Government Areas (44/57) emerged as the spatial unit for which such estimates were most widely sought, unsurprisingly given that local government officials comprised almost half the respondents. However, a large proportion of respondents underlined the need for estimates at the small area level. SA2s, a geographic unit with an average population of around 10,000 people (ABS 2016), were nominated by 42 respondents. Other responses requested estimates for individual towns, suburbs and discrete communities, as well as for Level 1 Statistical Areas, which have an average population of just 400 people. At the other end of the spatial scale, States and territories were nominated by nine respondents, while others pointed to a need for custom geographies, included gridded population data. The results confirm that estimates are needed at relatively high spatial resolution, and also reveal a desire to aggregate estimates over space. Estimates must therefore be in a form that allows summation over multiple spatial units without risking double counting of populations.

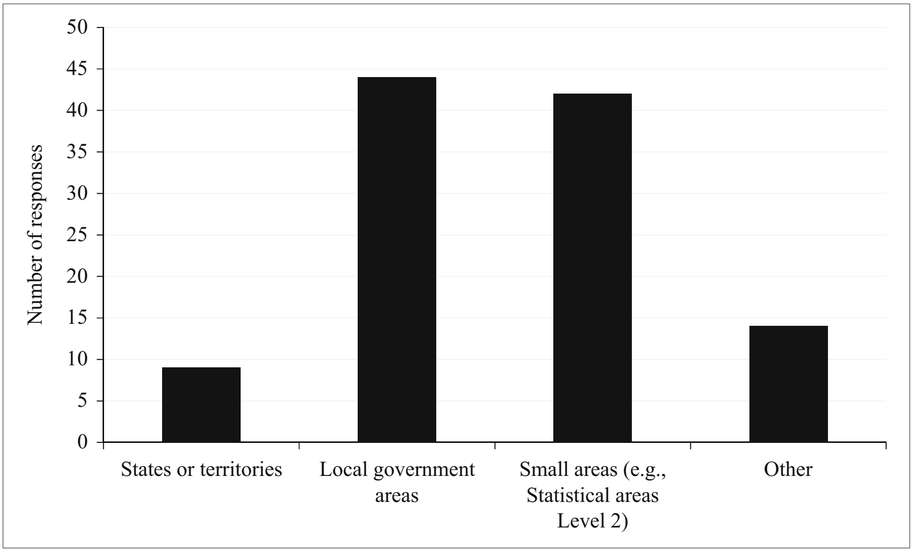


Fig. 5. Australian geographic units for which respondents sought estimates of temporary populations.

3.5. Measures

To gain a better understanding of the most useful output statistic, four measures of temporary populations that commonly appear in the literature as options were put to respondents (see Figure 6). These were:

1. The peak visitor population;
2. The total number of visitors in some period (e.g., a week, month or year);
3. The population present at a defined point in time; and
4. Visitor nights (or Person-time).

Estimates of the peak visitor population were the most popular option in the survey, nominated by 42 of 57 respondents. There are two ways in which measures of peak visitor populations are commonly implemented. The first is a measure of the capacity of an area, that is, the maximum number of people that can be accommodated in private and commercial accommodation (Planning Information and Forecasting Unit 2006). Estimates of capacity can be derived from tourist accommodation surveys, counts of unoccupied or second dwellings (McKenzie and Canterford 2018), or employer-provided housing, such as mining camps, as well as data from sources such as (AirDNA 2019). The second approach is to estimate the peak in actual visitor numbers, either directly from survey data or indirectly using symptomatic data (Smith 1989) to model the change in population numbers over time, benchmarked against the usually resident population. Examples include the use of wastewater data (SGS Economics and Planning 2007), retail spending statistics (Smith 1994), and mobile phone activity (Edmondson et al. n.d.). There are limitations to both the capacity approach and measures of the actual peak in population numbers. Estimates of capacity do not capture the timing of visits, and when aggregated over multiple spatial units, will produce a figure many times larger than the population, actually present. By contrast, it may be possible to capture the timing of population peaks, but again, estimates cannot be aggregated because the seasonality of temporary movements varies widely across space.

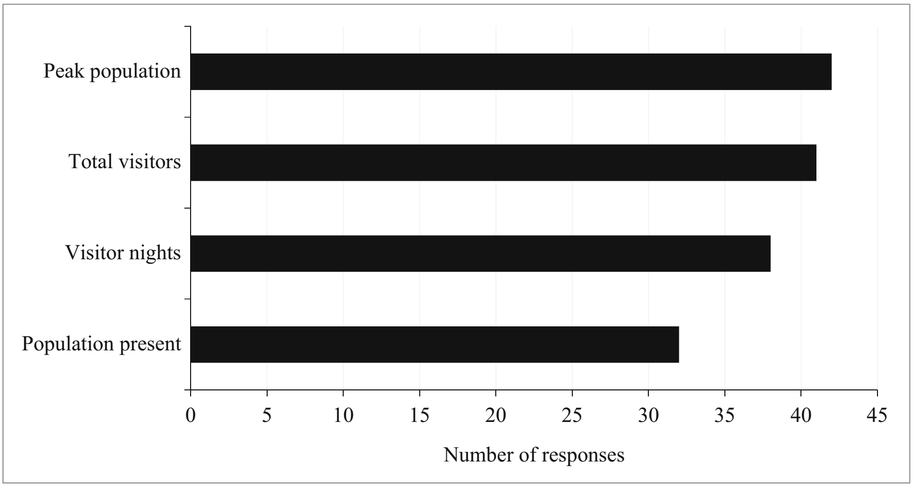


Fig. 6. Potential measures of temporary populations.

Forty-one respondents nominated estimates of the visitor population in a defined interval, such as a month or a year, as a statistic of interest. This measure corresponds closely to the concept of a service population, that is, the population that accesses goods and services in a defined area (ABS 2008; United Nations 2017). In practice, visitor population estimates tend to capture specific groups, such as elderly snowbirds, who migrate seasonally from northern to southern states of the United States over the winter months (Happel and Hogan 2002) or indigenous populations (Markham et al. 2013). Such estimates undoubtedly provide insights into the demand for some goods and services and visitor characteristics, but rarely reflect the actual population in a region at a given point in time. Efforts to estimate total visitor numbers by summing the component visitor groups (e.g., tourists+business travellers+commuters) risk double-counting individuals belonging to multiple populations (StatsNZ 2015). Double-counting is also an issue if estimates are for large areas, as movers can be counted at both their origin (as a resident) and the destination, or at multiple destinations if a trip involves a complex circuit. These issues are compounded if estimates are summed across multiple geographical units.

A third option put to respondents was a measure of visitor nights or person-time spent in a region, which was nominated by 38 of 57 respondents. Visitor nights is a common metric in tourism research (Theobold 2005), while person-time is a concept familiar to demographers and epidemiologists (Vandenbroucke and Pearce 2012), used in the calculation of occurrence-exposure rates. In essence, person-time combines information on the number of visitors (and absentees) in a region with the duration of time they spend in that region. In contrast to instantaneous estimates of the population present, person-time measures capture population over extended discrete time intervals that can be of varying duration. Person-time is not impacted by double-counting and can be aggregated over both space and time. Also, person-time can be used to approximate the average population present in a region in a defined interval by dividing the total person-time by the number of time units in the interval. However, it has not been widely adopted in the literature as a measure of temporary population (for exceptions, see Smith 1989; Batista e Silva et al. 2017).

Population present at a defined point in time was nominated by 32 of 57 respondents as a useful measure of temporary populations. Estimates of the temporary population present capture the number of people in an area at a given point in time and are conceptually equivalent to a de facto population figure. The latter would include both usual residents and visitors to a region. Given the significant seasonal variation of many forms of temporary movement, frequent estimates would be needed to capture peaks and troughs in population numbers. Population present measures have been generated using symptomatic data, including mobile phone data, to track changes in populations over time (Deville et al. 2014). An alternative approach is to estimate the various components of the population at a set point in time (see Swanson and Tayman 2011, for a formulation of this concept). However, in the absence of temporally and spatially synchronised data enumerating different groups, double counting is a potential source of error. Population present estimates have many merits. They are conceptually consistent with de facto population statistics and are not susceptible to double counting, as movers must be physically present in a region to be counted. Population present estimates can be aggregated across multiple geographies, though not over time. If estimates are produced at high frequency, summary metrics, such as the peak and average population of a region, can also be created.

3.6. Other Characteristics

Respondents were asked what was needed in regard to the characteristics of temporary movers. Age was seen as important by 48 of 54 respondents, while sex was nominated by 29. Other characteristics mentioned by respondents included indigenous status (6/54) and employment status (5/54). Given the difficulties in capturing population totals, at this time, estimates of other characteristics are a secondary concern, although information on age and sex would have clear value.

Synthesising the results of the survey, some principles for temporary population statistics can be identified. Population outputs need to include all visitors to a region, including domestic and international visitors, and occasional and regular visitors. The highest demand is for estimates that capture seasonal variation in population numbers and for small geographic areas. In Australia, Local Government Areas were nominated as the preferred spatial unit, reflecting their political function, but Statistical Areas, Level 2, with an average population across the country of around 10,000 persons (ABS 2016), and bespoke geographies were also identified as necessary. Fine spatial units would facilitate user wishes for estimates that can be aggregated into custom regions. Concerning the type of measure, respondents sought a wide range of statistics, including the ability to differentiate seasonal and periodic change. Within that framework, they also called for measures covering various aspects, including the peak population, total visitors and visitor nights. They also wanted information on characteristics and motives, but these emerge as secondary priorities. While this does not provide definitive guidance, it does demonstrate that demand is high and that all forms of data would likely be well received. In the first instance, at least, the way forward therefore, should be guided by considerations of data availability and processing.

4. Official Temporary Population Statistics: A Survey of OECD Countries

Temporary population statistics do not generally form part of standard national statistical outputs. However, a range of information is captured in other collections that may serve as input into temporary population estimates. These include travel and tourism surveys that capture information on domestic and international travel. Questions on temporary populations have been asked in a number of national censuses. For example, the 2011 Census of England and Wales asked “Do you stay at another address for more than 30 days a year?”. A number of countries record information on second homes, but information must be accompanied by data on utilisation to produce population estimates (Back and Marjavaara 2017). In recent years, several statistical agencies have also explored the utility of mobile phone and other “big data” sets as a source of information on population mobility, with pioneering work in this space emerging in Estonia (Ahas et al. 2011). To take stock of progress, we undertook a desktop survey of the national statistical agencies of 35 OECD member countries to determine the type of population estimates currently produced (*de jure*, *de facto* and working) and whether any type of temporary population estimates are available or *undergoing development*. We also sought to identify national travel and tourism surveys that collect data on temporary population mobility that may be used to inform official estimates. For this exercise, temporary population statistics were defined broadly as *any statistic that counts a non-resident population;* this includes both diurnal and overnight visitors. The review was undertaken in the last quarter of 2018 and was limited to national statistical agencies. Other government agencies, such as tourism bureaus or regional government offices might also produce relevant statistics, but these fell outside of our survey frame. As the survey was confined to online resources, results should be viewed as indicative rather than definitive, but do provide a flavour of contemporary approaches in the sample countries.

Results are shown in Table 2. Consistent with the UN Principles and Recommendations (United Nations 2017), all OECD countries produce *de jure* population estimates and most censuses are carried out on a *de jure* basis. However, five member countries also publish *de facto* counts from national population censuses (Australia, Ireland, Israel, Italy, and New Zealand). Comparison of *de facto* and *de jure* counts can provide useful insights into the overall intensity and spatial impacts of temporary mobility (Bell and Ward 2000). In the 2001 Australian Census, nearly 5% of the Australian population were enumerated away from home in a distinctive spatial pattern. Cross-tabulating place of usual residence by place of enumeration also helps identify origin-destination flows among temporary

Table 2. Desktop survey of official population statistics, OECD members.

Country	De jure	De facto	Working population (from census or register)	Temporary population estimates	Travel or tourism survey	Notes
Australia	Yes	Yes	Yes	No	Yes	1
Austria	Yes	No	Yes	No	Yes	
Belgium	Yes	No	Yes	No	Yes	
Canada	Yes	No	Yes	No	Yes	
Chile	Yes	No	Yes	No	Yes	
Czech Republic	Yes	Not known	Yes	No	Yes	
Denmark	Yes	No	Yes	Yes	Yes	2
Estonia	Yes	No	Yes	Yes	Yes	3
Finland	Yes	No	Yes	No	Yes	
France	Yes	No	Yes	Yes	Yes	4
Germany	Yes	Not known	Not known	No	Yes	
Greece	Yes	Not known	Not known	No	Not known	
Hungary	Yes	Not known	Yes	No	Yes	
Iceland	Yes	No	Not known	No	Yes	
Ireland	Yes	Yes	Yes	Yes	Yes	5
Israel	Yes	Yes	Yes	No	Yes	
Italy	Yes	Yes	Yes	Yes	Yes	6
Japan	Yes	No	Yes	No	Yes	
Korea (Republic of)	Yes	Not known	Not known	No	Not known	
Latvia	Yes	No	Yes	No	Yes	
Luxembourg	Yes	No	No	No	Yes	
Mexico	Yes	Not known	Yes	No	Yes	
Netherlands	Yes	No	Yes	Yes	Yes	7
New Zealand	Yes	Yes	Yes	Yes	Yes	8
Norway	Yes	No	Yes	No	Yes	
Poland	Yes	No	Yes	No	Yes	
Portugal	Yes	No	Yes	No	Yes	
Slovak Republic	Yes	Not known	Yes	No	Yes	
Slovenia	Yes	No	Yes	No	Yes	

Table 2. Continued.

Country	De jure	De facto	Working population (from census or register)	Temporary population estimates	Travel or tourism survey	Notes
Spain	Yes	No	Yes	No	Yes	
Sweden	Yes	No	Yes	Yes	Yes	9
Turkey	Yes	Not known	Not known	No	Yes	
United Kingdom	Yes	No	Yes	Yes	Yes	10
United States	Yes	Yes	Yes	No	Yes	

1. The Australian Bureau of Statistics conducted a pilot study to estimate temporary populations based on mobile phone data in 2016. The pilot study has yet to be released. http://www.abs.gov.au/websitedbs/d3310114.nsf/home/ABS+Media+Statements+-+Response+to+reports+about+use+of+aggregate+level+telco+data
2. Holiday dwellings: number and nights spent (https://www.dst.dk/en/Statistik/dokumentation/documentationofstatistics/holiday-dwellings)
3. Feasibility study on the use of mobile positioning data for tourism statistics (https://www.stat.ee/78262?highlight=mobile%2Cphone); Commuting in Estonia. An analysis based on mobile positioning data https://www.stat.ee/65754?highlight=mobile%2Cphone
4. Daytime population (https://www.cso.ie/en/releasesandpublications/ep/p-cp11eoi/cp11eoi/dtpn/)
5. *Population a compete a part: this captures* people counted away from their commune of usual residence, but certain groups are excluded, for example military personnel, people in hospitals, people in convents/monasteries https://www.insee.fr/fr/metadonnees/definition/c1650
6. 2001 Census question on multiple residences (https://unstats.un.org/unsd/demographic/sources/census/quest/ITA2001en.pdf)
7. Mobile phone estimates How many people are here? (https://www.cbs.nl/en-gb/our-services/innovation/project/how-many-people-here-)
8. Using mobile phone data to measure population movements (https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwjsweavzqjgAhXNfH0KHVUDBI0QFjAAegQIChAC&url=http%3A%2F%2Farchive.stats.govt.nz%2F~%2Fmedia%2FStatistics%2Fservices%2Fearthquake-info%2Fusing-cellphone-data-measure-pop-movement.pdf&usg=AOvVaw1vjQVs2_SDGgvNrhlHNwlb)
9. Holiday home areas (https://www.scb.se/en/finding-statistics/statistics-by-subject-area/environment/land-use/concentrations-of-holiday-homes/)
10. Workday population (https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/theworkdaypopulationofenglandandwales/2013-10-31); 2011 Census question on second address (https://www.ons.gov.uk/peoplepopulationandcommunity/housing/bulletins/2011censusnumberofpeoplewithsecondaddressesinlocalauthoritiesinenglandandwales/2012-10-22)

movers, while other census questions also reveal their characteristics. However, these benefits of compositional and spatial detail are offset by the fact that the census provides a snapshot of temporary mobility on a single day of the year, which may not be more broadly representative. In the United States, the American Community Survey (ACS) captures a modified *de facto* population based on a current residence rule, interviewing people living at an address for more than two months. Official guidance suggests that the data produced by the ACS are similar to usual resident counts derived from the decennial Census, except in areas '. . .that include large beach, lake, or mountain vacation areas, or large migrant worker communities. . .' (U.S. Census Bureau 2018, 60). In addition to *de jure* and *de facto* counts, estimates of working populations (i.e., based on place of work or study) are available from the Census or population registers in 28 countries.

Temporary populations statistics, defined in this instance to include daytime populations are produced in 9 of the 35 OECD member countries. Statistics fall into three main types.

The first estimate daytime populations using a combination of working population and usual resident statistics captured by national censuses (see, for example the UK Office of National Statistics (ONS 2014) and the Irish Central Statistics Office (CSO 2017)). These estimates use an accounting framework in which entrants to the population are people working in a region, while exits are usual residents working outside the region. The statistical output is an estimate of the average population present in a region during working hours.

The second group of estimates capture second home populations using surveys (e.g., Denmark), administrative data (e.g., Sweden) or via a census (England and Wales, and Italy). These estimates have varying conceptual bases in different countries. In Sweden, statistics are produced measuring the number of second home dwellings, but not their associated population, in a defined area, which can be used to estimate carrying capacity. In Italy, and England and Wales, statistics capture the number of people who used a second dwelling in the year prior to the Census. In Denmark, estimates are produced of both the number of second homes and the cumulative person-nights spent in those homes.

The third group of estimates capture temporary populations using mobile phone data. Statistics Netherlands has produced experimental estimates of the population present in municipalities at hourly intervals (CBS n.d.). In Australia, mobile phone data have been explored for their use in population estimates (ABS 2018). In addition to instantaneous estimates of the population present, mobile phone data have been used to capture temporary population flows. In Estonia, commuting and tourist flows have been monitored in partnership with academic researchers (Ahas et al. 2011; Ilves and Karus 2014). Statistics New Zealand used mobile phone data to track the temporary movements of people following the Christchurch earthquake in 2011 (StatsNZ 2012).

The desktop review suggests that there is some interest in the production of temporary population statistics in OECD countries, but there is no consensus as to how these are best conceptualised or produced. Approaches vary with respect to population coverage (e.g., daytime populations, second home owners, all visitors), how estimates are conceptualised (e.g., person-time, population present) and the methods and data used to produce the estimates. However, almost all countries do conduct some form of tourism or travel survey that can provide data on common forms of temporary population mobility. While progress is needed on multiple fronts, clear conceptualisation of temporary population statistics is a critical first step.

5. Towards Official Temporary Population Statistics

To date, there has been little progress in the production of temporary population statistics. This is not due to a lack of demand, as calls for temporary population statistics go back decades. What can account for this lack of progress? A paucity of data and inadequate processing procedures are undoubtedly factors, as is the embryonic state of conceptualisation of temporary populations and poor understanding of the underlying dynamics of temporary mobility. “Service population” is the principal concept relating to temporary populations in the official statistical literature, having entered the lexicon of the United Nations Population Division, and national statistical agencies, including Statistics New Zealand and Australia (see, for example StatsNZ 2015; ABS 2008). This concept has proven difficult to implement across multiple geographies and multiple population groups, and suffers from a range of shortcomings associated with estimates of total visitors that are described in Subsection 3.5. Two concepts discussed in this paper may prove useful to progress temporary population estimates: population present and person-time. The strength of both measures is their ability to aggregate numbers across regions (and across time, in the case of person-time) and constrain them to national estimates. Daily population present estimates are consistent with demographic accounts, in that changes in population stocks can be linked to the flows driving dynamic shifts in populations (Rees and Wilson 1973). In practice, the implementation of a full set of multi-regional accounts at daily intervals is likely unworkable due to the complexity of the underlying temporary mobility that we have described in Section 3. However, high frequency estimates of the population present can be generated from symptomatic data for single areas, and for multiple geographies using mobile phone data (CBS n.d.).

While population present estimates are now feasible in some settings, data access remains a key impediment. Mobile phone records are available in some countries, but in others, privacy concerns, fragmentation across multiple providers, and cost make access difficult. Ironically, access to mobile phone data is often easier in developing world countries due to data philanthropy, particularly following natural disasters (Bengtsson et al. 2011; Wilson et al. 2016). Person-time estimates that combine information on the numbers of visitors (and absentees) in a region with the duration of time they spend in that region are an alternative. In contrast to population present estimates, person-time measures population over discrete time intervals that can be of varying duration. To estimate person-time, information is needed on both the intensity of movement and the duration of stay at a destination. Fortunately, such data are collected in travel and tourism surveys that are ubiquitous across OECD countries. These two options are not mutually exclusive. Person-time is consistent with instantaneous measures of population present: if the population present over time is represented by a curve on a population-time chart, person-time is equivalent to the area under the curve. This relationship between person-time and population present offers potential avenues for combining data from multiple sources. A framework that exploits this relationship provides flexibility across national contexts.

6. Whither Temporary Population Estimates?

In this article, we explored issues associated with the production of official temporary population statistics. Guided by the framework developed by Raymer and colleagues (2015) we touched on a number of interrelated elements: the dynamics of the mobility driving short-term shifts in population numbers in a developed world context, user needs for statistics on temporary populations, and existing outputs measuring temporary populations published by OECD countries. The culmination of the article was a discussion of two statistical concepts germane to temporary population statistics: estimates of population present and of person-time. These concepts are, of course, not new to demographers. Integrating under a curve of population counts to calculate person-time is key to the calculation of occurrence-exposure rates, as well as the L_x column in life tables. However, these concepts have been absent from discussions of temporary population estimates. These concepts each have the potential to produce the statistics sought by users and the relationship between the two measures offers a potential avenue for combining data from different sources. Population present might be estimated using mobile phone data, while person-time might be derived from tourism surveys.

Statistical outputs are ultimately dependent upon matching concepts with available data sets. Temporary population mobility has not generally been measured in standard demographic collections, such as censuses or population registers. In instances where this has occurred (e.g., the 2011 Census of England and Wales), data have not been linked explicitly to clearly defined concepts. As data from mobile phone and other ICT develop, it is becoming more feasible to produce such estimates, but a number of challenges remain that relate to privacy, ownership, access and cost. There are also methodological and production challenges relating to bias, groundtruthing, and computing resource demands that are non-trivial (Tam and Clarke 2015), particularly if estimates are sought across multiple regions. In many instances, temporary population statistics will need to be based on a combination of disparate data sources, rather than a single data set. This echoes broader trends in the production of demographic statistics linking administrative data sets. Indeed, many countries are embarking on a transition away from traditional censuses, towards linked administrative data sets capable of producing longitudinal data (Kukutai et al. 2014). For example, Australia proposed replacing the 2016 Census with linked administrative data sets, although plans have since been put on hold (Bell 2015), while the Office of National Statistics (England and Wales) has a programme to develop an administrative census (ONS 2017). Systems are more developed in Nordic countries, with for example Statistics Finland having developed a fully operational statistical system linking population, business and property registers (Ruotsalainen 2018). The shift to longitudinal data offers promising new opportunities to assemble data on both suggested measures, first by providing an effective continuing census from which a snapshot of temporary populations can be extracted, and secondly by cumulating the time spent in particular jurisdictions to generate data on person-time. Also promising are longitudinal statistics on international visitors in countries such as Australia, which can be used to estimate the population physically present in a country by tracking arrivals and departures of international visitors and residents (Burleigh 2018). From this database, it is possible to estimate the persons present at any point in time, but person-time measures can also be generated by summing durations of absence and of stay.
Despite these developments, longitudinal data are not essential for the production of person-time measures and their development should not lead agencies to overlook existing data sets, such as tourism and travel surveys that capture retrospective information on the intensity and duration of moves. Also useful are occupancy statistics collected in tourist accommodation surveys, and from online platforms, such as AirBNB and HomeAway (AirDNA 2019). The need for official temporary population statistics will ultimately vary across national contexts and will be impacted by data availability and processing capability. A clear and consistent conceptual approach is an important first step in the development of widespread temporary population statistics.

7. References

ABS. 2008. *Information Paper: Population Concepts, 2008, Catalogue no. 3107.0.55.006*. Canberra: Australian Bureau of Statistics. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/3107.0.55.006 (accessed March 2019).

ABS. 2015. *Family Characteristics and Transitions, Australia, 2012–13, Catalogue no. 4442.0* Canberra: Australian Bureau of Statistics. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/4442.0 (accessed March 2019).

ABS. 2016. *Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, Catalogue no. 1270.0.55.001*. Canberra: Australian Bureau of Statistics. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.001 (accessed March 2019).

ABS. 2018. Response to reports about use of aggregate level telco data. Canberra: Australian Bureau of Statistics. Available at: https://www.abs.gov.au/websitedbs/d3310114.nsf/home/ABS+Media+Statements+-+Response+to+reports+about+use+of+aggregate+level+telco+data (accessed March 2019).

ABS. 2019. *Overseas Arrivals and Departures, Australia, Dec 2018, Catalogue no. 3401.0*. Canberra: Australian Bureau of Statistics. Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/allprimarymainfeatures/A7FC8F595E0E1D63CA2583BD0076FDF4?opendocument (accessed March 2019).

Ahas, R., S. Silm, A. Aasa, K. Leetmaa, E. Saluveer, and M. Tiru. 2011. "Commuting in Estonia. An analysis based on mobile positioning data." In *Regional Development in Estonia*, edited by E. Narusk, 197. Tallinn: Statistics Estonia. Available at: https://www.stat.ee/publication-download-pdf?publication_id=25596 (accessed May 2019).

AirDNA. 2019. "MarketMinder." Available at: https://www.airdna.co/vacation-rental-data (accessed June 2019).

Back, A. and R. Marjavaara. 2017. "Mapping an invisible population: the uneven geography of second-home tourism." *Tourism Geographies* 19(4): 595–611. DOI: https://doi.org/10.1080/14616688.2017.1331260.

Batista e Silva, F., K. Rosina, M. Schiavina, M. Marin, S. Freire, M. Craglia, and C. Lavalle. 2017. *Spatiotemporal mapping of population in Europe: The'ENACT' project in a nutshell*. Ispra, Italy: European Commission Joint Research Centre. Available at: https://ec.europa.eu/jrc/en/publication/spatiotemporal-mapping-population-europe-enact-project-nutshell (accessed June 2019).

Bell, M. 2001. "Understanding circulation in Australia: Presidential Address." *Journal of Population Research* 18(1): 1–18. DOI: https://doi.org/10.1007/BF03031952.

Bell, M. 2004. "Measuring temporary mobility: dimensions and issues." Cauthe Conference, 10–13 February, 2004, Brisbane, Australia: Council for Australasian Tourism and Hospitality Education.

Bell, M. 2015. "W(h)ither the Census?" *Australian Geographer* 46(3): 299–304. DOI: https://doi.org/10.1080/00049182.2015.1058320.

Bell, M., M. Blake, P. Boyle, O. Duke-Williams, P. Rees, J. Stillwell, and G. Hugo. 2002. "Cross-national comparison of internal migration: issues and measures." *Journal of the Royal Statistical Society A* 165(3): 435–464. DOI: https://doi.org/10.1111/1467-985X.t01-1-00247.

Bell, M. and D. Brown. 2006. "Who are the visitors? Characteristics of temporary movers in Australia." *Population, Place and Space* 12(2): 77–92. DOI: https://doi.org/10.1002/psp.390.

Bell, M. and G. Ward. 1998. "Patterns of temporary mobility in Australia: evidence from the Census." *Australian Geographical Studies* 36(1): 58–81. DOI: https://doi.org/10.1111/1467-8470.00039.

Bell, M. and G. Ward. 2000. "Comparing temporary mobility with permanent migration." *Tourism Geographies* 2(1): 87–107. DOI: https://doi.org/10.1080/146166800363466.

Bengtsson, L., X. Lu, A. Thorson, R.S Garfield, and J. von Schreeb. 2011. "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti." *PLOS Medicine* 8(8): e1001083. DOI: https://doi.org/10.1371/journal.pmed.1001083.

Burleigh, M. 2018. "The physically present population." Australian Population Association Conference, 18–20 July, Darwin, Australia: Australian Population Association.

CBS. n.d. "How many people here?". Statistics Netherlands (CBS). Available at: https://www.cbs.nl/en-gb/our-services/innovation/project/how-many-people-here- (accessed April 2019).

Chapman, M. 1978. "On the Cross-Cultural Study of Circulation." *International Migration Review* 12 (4, Special Issue: Illegal Mexican Immigrants to the United States): 559–569. DOI: https://doi.org/10.1177/019791837801200406.

Chapman, M. and R.M. Prothero. 1983. "Themes on circulation in the third world." *International Migration Review* 17(4): 597–631. DOI: https://doi.org/10.1177/019791838301700402.

Charles-Edwards, E., M. Bell, and D. Brown. 2008. "Where people move and when: temporary population mobility in Australia." *People and Place* 16(1): 21–30. Available at: https://search.informit.com.au/documentSummary;dn=157293264384755;res=IELHSS.

Charles-Edwards, E. and M. Bell. 2015. "Seasonal Flux in Australia's Population Geography: Linking Space and Time." *Population, Space and Place* 21(2): 103–123. DOI: https://doi.org/10.1002/psp.1814.

Cook, T. 1996. *Demography Working Paper 1996/4 - When ERPs aren't Enough, 1996, Catalogue no* 3112.0. Canberra: Australian Bureau of Statistics. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/3112.0 (accessed March 2019).

Cook, T. 1998. "Overnight visitor counts in Australia and their implications for population estimation." *People Place* 6(1): 60–70.

CSO. 2017. *Census of Population 2016 – Profile 11 Employment, Occupations and Industry*. Cork, Ireland: Central Statistics Office, Ireland. Available at: https://www.cso.ie/en/csolatestnews/presspages/2017/census2016profile11employmentoccupationsandindustry/ (accessed March 2019).

Davis, A. 2011. "On constructing ageing rural populations: 'Capturing' the grey nomad." *Journal of Rural Studies* 27(2): 191–199. DOI: https://doi.org/10.1016/j.jrurstud.2011.01.004.

Deville, P., C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, V.D. Blondel, and A.J. Tatem. 2014. "Dynamic population mapping using mobile phone data." *Proceedings of the National Academy of Sciences* 111(45): 15888–15893. DOI: https://doi.org/10.1073/pnas.1408439111.

Edmondson, B. and Nantucket Data Platform team. n.d. *Making It Count. A Data-Driven Look at Nantucket's Dynamic Population*. Nantucket: Nantucket Data Platform.

Halford, K. 2018. How will my divorce affect my kids? *The Conversation* (September 11, 2018). Available online: http://theconversation.com/how-will-my-divorce-affect-my-kids-101594 (accessed April 2019).

Hanson, J. and M. Bell. 2007. "Harvest trails in Australia: patterns of seasonal migration in the fruit and vegetable industry." *Journal of Rural Studies* 23: 101–117. DOI: https://doi.org/10.1016/j.jrurstud.2006.05.001.

Happel, S.K. and T.D. Hogan. 2002. "Counting snowbirds: the importance of and the problems with estimating seasonal populations." *Population Research and Policy Review* 21: 227–240. DOI: https://doi.org/10.1023.

Houghton, D. 1993. "Long distance commuting; a new approach to mining in Western Australia." *Geographical Journal* 159(3): 281–290. DOI: https://doi.org/10.2307/3451278.

Hugo, G. and K. Harris. 2013. Time and tide: moving towards an understanding of temporal population changes in coastal Australia. Adelaide: The University of Adelaide. Available at: http://www.esc.nsw.gov.au/living-in/about/community-profile-and-population-forecasts/sea_change_taskforce_report_29_april_2013.pdf (accessed January 2019).

Ilves, M. and E. Karus. 2014. "Feasibility study on the use of mobile positioning data for tourism statistics." *Quarterly Bulletin of Statistics Estonia* 2(14):27. Available at: https://www.stat.ee/publication-2014_quarterly-bulletin-of-statistics-estonia-2-14.

Kounadi, O., A. Ristea, M. Leitner, and C. Langford. 2018. "Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies." *Cartography and Geographic Information Science* 45(3): 205–220. DOI: https://doi.org/10.1080/15230406.2017.1304243.

Kukutai, T., V. Thompson, and R. McMillan. 2014. "Whither the census? Continuity and change in census methodologies worldwide, 1985–2014." *Journal of Population Research* 32: 3–22. DOI: https://doi.org/10.1007/s12546-014-9139-z.

Markham, F., J. Bath, J. Taylor, and B. Doran. 2013. *New directions in indigenous service population estimation, CAEPR Working Paper No. 88/2013*. Canberra: Centre for Aboriginal Economic Policy Research. Aavilable at: https://openresearch-repository.anu.edu.au/bitstream/1885/147837/1/WP88_Markham_et_al_Service_delivery_0.pdf (accessed January 2019).

McKenzie, F. and S. Canterford. 2018. *Demographics for bushfire risk analysis: regional Victoria and peri-urban Melbourne*. Melbourne: Department of Environment, Land, Water and Planning. Available at: https://www.planning.vic.gov.au/__data/assets/pdf_file/0035/97685/Demographics-for-Bushfire-Risk-Analysis-web.pdf (accessed October 2018).

McKercher, B. 2018. "The impact of distance on tourism: a tourism geography law." *Tourism Geographies* 20(5): 905–909. DOI: https://doi.org/10.1080/14616688.2018.1434813.

Mitchell, J.C. 1969. "Structural plurality, urbanization and labour circulation in Southern Rhodesia." In *Migration*, edited by J.A. Jacksons: 156–180. London: Cambridge University Press.

Müller, D.K., C.M. Hall, and D. Keen. 2004. "Second home tourism impact, planning and management." In *Tourism, mobility and second homes: between elite landscape and common ground*, edited by C.M. Hall and D.K. Müller: 15–32. Clevedon: Channel View Publications.

ONS. 2014. 2011 Census: Workplace Population Analysis. Newport, UK: Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/workplacepopulationanalysis/2014-05-23 (accessed March 2019).

ONS. 2017. Annual assessment of ONS's progress towards an Administrative Data Census post-2021. Newport, United Kingdom: Office of National Statistics. Available at: https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments/annualassessmentofonssprogresstowardsanadministrativedatacensuspost2021 (accessed March 2019).

Paris, C., C. Thredgold, B. Jorgensen, and J. Martin. 2014. *Second homes and changing populations: Impacts and implications for local government in South Australia*. Adelaide: Centre for Housing, Urban and Reginal Planning (CHURP), University of Adelaide. Available at: http://www.lga.sa.gov.au/webdata/resources/project/2013.35_-_Second_homes_and_changing_populations_-_Impacts_and_implications_for_local_government_in_SA.pdf (accessed March 2019).

Plane, D.A. and G.F. Mulligan. 1997. "Measuring spatial focusing in a migration system." *Demography* 34(2): 251–262. DOI: https://doi.org/10.2307/2061703.

Planning Information and Forecasting Unit. 2006. *Towards estimates of service populations to inform planning in small areas*. Brisbane: Queensland Department of Local Government, Planning, Sport and Recreation.

Prothero, R.M. and M. Chapman. 1985. *Circulation in Third World countries*. Boston: Routledge.

QGSO. 2018. *Bowen Basin population report, 2018*. Brisbane: Queensland Treasury.

Raymer, J., P. Rees, and A. Blake. 2015. "Frameworks for Guiding the Development and Improvement of Population Statistics in the United Kingdom." *Journal of Official Statistics* 31(4): 699–722. DOI: https://doi.org/10.1515/jos-2015-0041.

Rees, P.H. and A.G. Wilson. 1973. "Accounts and Models for Spatial Demographic Analysis I: Aggregate Population." *Environment and Planning A: Economy and Space* 5(1): 61–90. DOI: https://doi.org/10.1068/a050061.

Ruotsalainen, K. 2018. "Techniques for administrative data sources integration in statistical registers." ESTP – Moving towards register based statistical system Valencia, Spain, 12–14 September 2018.

SGS Economics and Planning. 2007. *Impacts of Coastal Population Fluctuations*. Melbourne: Report prepared for Department of Sustainability and Environment, Victorian Government.

Smith, S.K. 1989. "Toward a methodology for estimating temporary residents." *Journal of the American Statistical Association* 84(406): 430–436. DOI: https://doi.org/10.2307/2289926.

Smith, S.K. 1994. "Estimating temporary populations." *Applied Demography* 9(1): 4–7.

StatsNZ. 2012. "Using cellphone data to measure population movements." Wellington: Statistics New Zealand. Available at: https://www.google.com/url?sa=t&rct=

j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwitwO6BwdvkAhUGILcAHU9KBYMQFjAAegQIABAC&url=http%3A%2F%2Farchive.stats.govt.nz%2F~%2Fmedia%2FStatistics%2Fservices%2Fearthquake-info%2Fusing-cellphone-data-measure-pop-movement.pdf&usg=AOvVaw1vjQVs2_SDGgvNrhlHNwlb (accessed April 2019).

StatsNZ. 2015. “Service Population Concept.” Wellington: Statistics New Zealand, Available at: http://datainfoplus.stats.govt.nz/item/nz.govt.stats/66ea8a7a-f1c7-4cbc-80cd-75b5a018df68/0/ (accessed April 2019).

Swanson, D.A. and J. Tayman. 2011. “On Estimating a De Facto Population and its Components.” *Review of Economics & Finance*. Available at: http://www.bapress.ca/Journal-5/On%20Estimating%20a%20De%20Facto%20Population%20and%20Its%20Components%20By%20David%20A.%20Swanson.pdf.

Tam, S.M. and F. Clarke. 2015. *Big Data, Statistical Inference and Official Statistics, Calalogue no. 1351.0.55.054*. Canberra: Australian Bureau of Statistics. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.054 (accessed March 2019).

Taylor, J. and M. Bell. 2012. “Towards comparative measures of circulation: Insights from Indigenous Australia.” *Population, Space and Place* 18(5): 567–578. DOI: https://doi.org/10.1002/psp.695.

Taylor, J. 1986. “Measuring circulation in Botswana.” *Area* 18(3): 203–208. https://www.jstor.org/stable/20002342.

Theobold, W.F. 2005. “The meaning, scope, and measurement of travel and tourism.” In *Global Tourism*, edited by W.F. Theobold. Burlington, MA: Elsevier Science.

Tourism Research Australia. 2018. Travel by Australians. Canberra: Austrade.

United Nations. 2017. *Principles and Recommendations for Population and Housing Censuses*. New York: Department of Economic and Social Affairs Statistics Division.

U.S. Census Bureau. 2018. Understanding and Using American Community Survey Data: What All Data Users Need to Know. Washington D.C.: U.S. Department of Commerce.

Vandenbroucke, J.P. and N. Pearce. 2012. “Incidence rates in dynamic populations.” *International Journal of Epidemiology* 41: 1472–1479. DOI: https://doi.org/10.1093/ije/dys142.

Williams, A.M. and C.M. Hall. 2000. “Tourism and migration: New relationships between production and consumption.” *Tourism Geographies* 2(1): 5–27. DOI: https://doi.org/10.1080/146166800363420.

Wilson, R., E. zu Erbach-Schoenberg, M. Albert, D. Power, S. Tudge, M. Gonzalez, S. Guthrie, H. Chamberlain, C. Brooks, C. Hughes, L. Pitonakova, C. Buckee, X. Lu, E. Wetter, A. Tatem, and L. Bengtsson. 2016. “Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake” *PLOS Currents Disasters* 24(1). DOI: https://doi.org/10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c.

Zelinsky, W. 1971. “The hypothesis of the mobility transition.” *Geographical Review* 61(2): 219–249. DOI: https://doi.org/10.2307/213996.

Received April 2019
Revised July 2019
Accepted September 2019

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 25–48, http://dx.doi.org/10.2478/JOS-2020-0002

Identifying the Direction of Behavioral Dependence in Two-Sample Capture-Recapture Study

Kiranmoy Chatterjee[1] *and Diganta Mukherjee*[2]

With the possibility of dependence between the sources in a capture-recapture type experiment, identification of the direction of such dependence in dual system of data collection is vital. This has a wide range of applications, including in the domains of public health, official statistics and social sciences. Owing to the insufficiency of data for analyzing a behavioral dependence model in dual system, our contribution lies in the construction of several strategies that can identify the direction of underlying dependence between the two lists in the dual system, that is, whether the two lists are positively or negatively dependent. Our proposed classification strategies would be quite appealing for improving the inference as evident from recent literature. Simulation studies are carried out to explore the comparative performance of the proposed strategies. Finally, applications on three real data sets from various fields are illustrated.

Key words: Classification; direction of behavioral dependence; human population; randomized rule; recapture probability.

1. Introduction and Motivation

Estimation of the size of a given population is an important statistical concern that has vast application in the field of public health, population studies and animal abundance. In practice, it is mostly impossible to count all the individuals in a population accurately by any attempt, especially when the population is large enough or very hard to reach. As a remedy, more than one attempt is carried out independently and the population size (N) is estimated by matching the available (two or more) lists of information. This kind of data structure is known as a multiple-record system, which is equivalent to the capture-recapture system popularly relevant to abundance of animal population. However, in the context of a closed human population, use of more than two sources of information is uncommon in the official registration systems of most countries. When two attempts have been made to estimate the N in capture-recapture format, then the resulting data structure is known as a dual-record system (DRS), which is presented in Table 1. Estimation of census coverage error (Gerritse et al. 2017; Chatterjee and Mukherjee 2016a), epidemiological events (Iñigo et al. 2003; Granerod et al. 2013), size of hard-to-count

[1] Department of Statistics, Bidhannagar College, Salt Lake City, Kolkata, 700064, India. Email: kiranmoy07@gmail.com
[2] Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata, 700108, India. Email: diganta@isical.ac.in

Acknowledgments: Authors gratefully acknowledge the partial financial support by the research fellowship award (Office Order No. DS/JSTK-CC-0020F Dated 11 July, 2013) received by the first author from Indian Statistical Institute, India during his attachment with the institute.

Table 1. Data structure in Dual-Record System (DRS).

	List 2		
List 1	In	Out	Total
In	x_{11}	x_{10}	$x_{1\cdot}$
Out	x_{01}	x_{00}	$x_{0\cdot}$
Total	$x_{\cdot 1}$	$x_{\cdot 0}$	$x_{\cdot\cdot} = N$

population (Ruiz et al. 2016) are the primary applications of DRS for human population. ChandraSekar and Deming (1949) proposed homogeneous post-stratification of the data obtained in DRS structure in order to reduce heterogeneity with respect to the capture probabilities among individuals. This proposal has been commonly applied in most of the official statistics, as well as in epidemiological data sets (Eckberg 2000; Iñigo et al. 2003). Bohning et al. (2017) discussed some recent developments of the applications of various capture-recapture models in the arena of epidemiology, medical and social science.

After the construction of such mutually exclusive post-strata, which are within homogeneous but between heterogeneous, relevant statistical models for DRS can be analyzed for each of those post-strata. Model M_t (Otis et al. 1978), equivalent to the Petersen Model (Wolter 1986), has received much attention among practitioners because of its simplicity and identifiability in the vicinity of DRS for a human population. In Table 1, $x_{1\cdot}$, $x_{\cdot 1}$ and x_{11} refer to the number of individuals present in the first list (List 1), second list (List 2) and their common list, respectively. Following the underlying assumption of list-independence between the two lists, the estimator of N based on the model M_t is found to be $\frac{x_{1\cdot} x_{\cdot 1}}{x_{11}}$ (Wolter 1986; Chatterjee and Mukherjee 2016b). This estimator is popularly known as the Petersen estimator in epidemiology, or the ChadraSekar-Deming estimator in the domain of demography or population studies. More details on this model, including various likelihood based estimating approaches, can be found in Chatterjee and Mukherjee (2016b).

However, the assumption of list-independence may seriously mislead in many situations for human populations. Several methodologists and practitioners (ChandraSekar and Deming 1949; Greenfield 1975; El-Khorazaty 2000; Jarvis et al. 2000; Chao et al. 2001) argued that the list-independence assumption may not be justifiable in reality. An efficient brief review was done by Brittain and Bohning (2009) on the various methods available by relaxing the independence assumption and associated comparative study was undertaken in DRS context. The assumption of list-independence is often violated due to the presence of *behavioral response variation* at the time of a second capture attempt in DRS. An individual who is enlisted in List 1 may be more likely to also be included in List 2 than the individual who has not been enlisted in List 1. Hence, the corresponding population is treated as *recapture prone*. This kind of behavioral connection at the second time attempt is commonly encountered in epidemiological (Chao et al. 2001; Granerod et al. 2013) and demographic (Bell 1993; Griffin 2014) studies. Otherwise, in reverse cases, populations become *recapture averse*, for example hard-to-count population, drug addicted population. An interesting attempt to judge the effect of violation of the list-independence in this context was done by Gerritse et al. (2017). These kinds of changes in the behavior of an individual, while he/she attempts to be included in List 2, is grossly known *as behavioral response variation* (Wolter 1986; Chatterjee and Mukherjee 2018). O'Connell and Pollock (1992) introduced a strategy using demographic co-variates for partitioning a population in terms of direction of behavioral dependence. When both the *time variation* and *behavioral response variation* act together, model M_{tb}, to be defined later, can be treated as the most general and suitable statistical model for analyzing capture-recapture data under homogeneity. Gosky and Ghosh (2011) showed the appropriateness of this model among all the capture-recapture models proposed in Otis et al. (1978).

Now we state some basic notation before proceeding further. Let p_{ij} be the probability attached to each individual to be included in the count x_{ij} in (i, j)th cell of Table 1, where $i, j \in \{0, 1, \cdot\}$. In addition we denote

Pr (an individual is captured in List 2 | s/he is listed in List 1) $= \frac{p_{11}}{p_{1\cdot}} = c$,

Pr (an individual is captured in List 2 | s/he is not listed in List 1) $= \frac{p_{01}}{1-p_{1\cdot}} = p$, following Wolter (1986). In this article, we consider $c \neq p$ which refers a violation of independence between the two lists. Thus, there always exists some constant $\phi(>0)$ such that $c = \phi p$. This ϕ is termed *as behavioral response effect*. Note that $\phi > 1$ (equivalently, $c > p$) refers to positive association between the two sources and the associated population is said to be *recapture prone* and $\phi < 1$ (equivalently, $c < p$) refers to negative association and the associated population is said to be *recapture averse*. A capture-recapture model with *time variation* and *behavioral response variation*, denoted as M_{tb}, incorporates *behavioral dependence* between lists. However, model M_{tb} suffers from a problem in DRS, as ϕ or p is not estimable separately, but their product $\phi p = c$ is. This unidentifiability of the realistic model M_{tb} in DRS is discussed in Chao et al. (2000) and Chatterjee and Mukherjee (2016c).

One important aspect of the model M_{tb} is that if some proper knowledge on the direction of ϕ is available, then it might help to draw a reasonably good inference on the N, as evident from literature (Nour 1982; Chatterjee and Mukherjee 2016c, 2018). Particularly, if underlying ϕ for a population is correctly known to be greater than 1 (or less than 1), then uncertainty on ϕ will be reduced, since the domain of ϕ shrinks to $(1, \infty)$ (or $(c, 1)$), where c is the recapture probability of an individual, defined earlier. Hence, one can expect that inference would likely be better if that available knowledge is used. This issue has been proved empirically in Chatterjee and Mukherjee (2016c, 2018). For example, in a demographic study or homicide death study, two sources (or, lists) are commonly positively depended on, which leads to a high estimate of c. On the one hand, in economic surveys, people often want to avoid to be enlisted repeatedly in both the lists, and that leads to low estimate of the recapture probability c. Therefore, one may intuitively suggest that a given population is *recapture prone* if the maximum likelihood estimate of c, denoted by $\hat{c}$, is very close to its upper bound 1. On the other hand, if $\hat{c}$ is very close to its lower bound 0, then the associated population would be *recapture averse* with high probability. Giving an idea about the possible direction of ϕ is always a challenging job if $\hat{c}$ is neither close to 1, nor close to 0. As per our knowledge, no strategy has been developed yet to infer the direction of ϕ from capture-recapture data only. Therefore, in this article, our aim is to develop some classification strategies with which a given population can be identified as *recapture prone*, *recapture averse*, or *list-independent* with high probability in connection with DRS.

In this article, based on various considerations, we propose a large number of classification strategies in Section 2 to identify the direction of list-dependency in DRS. Performance of the proposed strategies are thoroughly investigated with simulated data in Section 3. Based on their performances, a subset of the proposals are applied to three real data sets from various fields of application including demography, crime statistics and economics in Section 4. We illustrate our classification strategies through inference about the direction of inherent dependence in DRS corresponding to each of the three data sets. Finally, in concluding Section 5, we discuss the implications, advantages and possible extensions of the proposed classification strategies.

2. Proposed Classification Strategies

In this section, we propose strategies to identify whether individuals belonging to a given population are *recapture prone* or *recapture averse* or their inclusion statuses in DRS are *list-independent*. To formulate such strategies, we need to impose an assumption on a parameter of the model M_{tb}. For the present context, we consider that the conditional probability $p \geq \pi$, where a value of π can be chosen. Note that this assumption allows a larger interval of possible values of p,$[\pi, 1)$, as one chooses a smaller π. We will discuss this choice later. For a justification of this, we refer to a parallel assumption adopted in Nour (1982) for estimating total number of vital events (e.g., birth, death) where the capture probability in each of the two lists (i.e., $p_{1\cdot}$ and $p_{\cdot 1}$) is assumed to be greater than 0.5. Now, from the definition of p, stated in Section 1, one can write that

$$p = \frac{p_{01}}{1 - p_{1\cdot}} = \frac{x_{01}}{N - x_{1\cdot}},$$

since DRS satisfy $Np_{ij} = x_{ij}$ for all $i, j \in \{0, 1, \cdot\}$. The assumption of $p \geq \pi$ can be rewritten as $(c - \phi\pi) \geq 0$, since $c = \phi p$. By virtue of the *mle*, $\hat{c} = (x_{11}/x_{1\cdot})$ being consistent, as well as an efficient estimator of c, we replace c by $\hat{c}$ and use the approximate relation $\phi = \hat{c}p^{-1}$ for sufficiently large N. Therefore, as $N \geq x_0$, where x_0 denotes the total number of distinct individuals counted in DRS,

$$\begin{aligned} \phi = \hat{c}p^{-1} = \frac{x_{11}}{x_{1\cdot}} \frac{N - x_{1\cdot}}{x_{01}} &\geq \frac{x_{11}}{x_{1\cdot}} \\ or,\ (x_{1\cdot}\phi - x_{11}) &\geq 0. \end{aligned} \tag{1}$$

Thus, using the inequalities in Equation (1) and $(\hat{c} - \phi\pi) \geq 0$, obtained earlier with the consideration of approximation of c by $\hat{c}$, we consider the (composite) inequality

$$\frac{(\hat{c} - \phi\pi)(x_{1\cdot}\phi - x_{11})}{x_{\cdot 1}\phi} \geq k$$

for $x_{\cdot 1} > 0$, where k is some non-negative real number. The choice of k will be explained later. Hence, the above inequality may be expressed as

$$\phi^2 + \phi \, \frac{kx_{\cdot 1}(\pi + 1)x_{11}}{\pi x_{1\cdot}} + \hat{c}^2 \pi^{-1} \le 0 \tag{2}$$

$$\text{or, } (\phi - \phi_0)(\phi - \phi_1) \le 0, \tag{3}$$

where ϕ_0 and ϕ_1 are two real roots of the quadratic equation corresponding to (2), which are functions of the unknown k and satisfy $\phi_0 + \phi_1 = \frac{(\pi+1)x_{11} - kx_{\cdot 1}}{\pi x_{1\cdot}}$ and $\phi_0\phi_1 = \hat{c}^2\pi^{-1}$ with $\phi_0 \le \phi \le \phi_1$ from Equation (3). In fact, the two roots are positive real valued under some condition on k. All the mathematical proofs and justifications on the nature of the two said roots of ϕ are rigorously discussed in the Appendix (Section 6). Since the arithmetic mean of any two positive real numbers is always greater than or equal to their geometric mean, therefore one will have

$$k \leqslant \frac{x_{11}}{x_{\cdot 1}} (1 - \sqrt{\pi})^2,$$

equality holds only when $\phi_0 = \phi_1 = \phi = \hat{c}\pi^{-1/2}$. From the Appendix we know that the above upper bound of k is the necessary and sufficient condition for both the roots ϕ_0 and ϕ_1 to be positive real-valued. In addition, as $k \ge 0$ holds under the assumption that $p \ge \pi$, the values of ϕ_0 and ϕ_1 corresponding to the lower bound of k are $\hat{c}$ and $\hat{c}\pi^{-1}$ respectively. Furthermore, the root ϕ_0 is a monotonically increasing function of k, while ϕ_1 is decreasing in k. This implies

$$\hat{c} \le \phi_0 \le \hat{c}\pi^{-1/2}$$

and

$$\hat{c}\pi^{-1/2} \le \phi_1 \le \hat{c}\pi^{-1}.$$

Proposals. By combining the two above inequalities for the two roots ϕ_0 and ϕ_1, we have

$$\hat{c} \le \phi \le \hat{c}\pi^{-1}.$$

Now, arithmetic (A.M.), geometric (G.M.) and harmonic (H.M.) means of the two limits of the above interval are, $\frac{(1+\pi)\hat{c}}{2\pi}$, $\hat{c}\pi^{-1/2}$ and $\frac{2\hat{c}}{1+\pi}$, respectively. We want to consider a reasonably moderate value of k to allow the possibility of a small value of π. As discussed earlier, ϕ_0 and ϕ_1 are respectively increasing and decreasing functions of k, we consider $\phi_0 \ge$ H.M. to accommodate such reasonable value of k and it automatically implies $\phi_1 \le$ A.M. as $\phi_0\phi_1 = \hat{c}^2\pi^{-1}$. Since $\phi_0 \le \phi \le \phi_1$, one would therefore readily have H.M. $\le \phi \le$ A.M., that is,

$$\frac{2\hat{c}}{1+\pi} \le \phi \le \frac{(1+\pi)\hat{c}}{2\pi}. \tag{4}$$

We call this proposed interval for ϕ as **Proposal 1**. One may further propose tighter bounds for the classification strategies as per the following reasoning.

Proposal 2: As $\hat{c}\pi^{-1/2} \le \phi_1$, we may propose the bound for ϕ as H.M. $\le \phi \le$ G.M., since G.M. is the smallest upper bound for ϕ. So,

$$\frac{2\hat{c}}{1+\pi} \le \phi \le \hat{c}\pi^{-1/2}. \tag{5}$$

One can alternatively propose,

Proposal 3: As $\phi_0 \leq \hat{c}\pi^{-1/2}$, we may propose the bound for ϕ as G.M. $\leq \phi \leq$ A.M., since G.M. is the greatest lower bound for ϕ. So,

$$\hat{c}\pi^{-1/2} \leq \phi \leq \frac{(1+\pi)\hat{c}}{2\pi}. \tag{6}$$

Now, we present classification rules to identify the direction of behavioral dependence among individuals in a population. Three alternative rules for each of the above three proposals are designed to indicate whether underlying ϕ is greater than 1 (i.e., proneness), less than 1 (i.e., aversion) or very close to 1 (i.e. list-independence). Let us consider that ϕ_l and ϕ_u respectively are the lower and upper tolerance limits of ϕ, equidistant from 1, beyond which one may say that underlying DRS is list-dependent. For example, if we consider 5% tolerance, then $(\phi_l, \phi_u) = (0.95, 1.05)$. We outline the classification strategies in detail for Proposal 1 based on the bounds in Equation (4), corresponding to Proposal 1. The variations for other proposals are analogously defined, based on the bounds in Equations (5) and (6).

Rule 1L. We set the lower bound of the interval, $\frac{2\hat{c}}{1+\pi} \leq \phi \leq \frac{(1+\pi)\hat{c}}{2\pi}$, that is, $\frac{2\hat{c}}{1+\pi}$ as a threshold to infer about the direction of dependence. So, if $\frac{2\hat{c}}{1+\pi} > \phi_u$ we say the population is *recapture prone*. Again if $\frac{2\hat{c}}{1+\pi} < \phi_l$, we say the population is *recapture averse*. Further, the population will be termed as *list-independent* if $\phi_l \leq \frac{2\hat{c}}{1+\pi} \leq \phi_u$. This lower bound technique may be conservative as it has a tendency towards indicating recapture aversion (see Section 3).

Analogously, if we replace the threshold in Rule *1L* by the upper bound of the interval, we can define Rule *1U*. Note that Rule *1U* may be conservative towards recapture proneness (see Section 3).

Admitting results of the previous classification rules, now we propose a randomized rule, similar to the statistical hypothesis test rule (as in the Neyman-Pearson tradition), that will safeguard the decision rule from possible threats of bias towards recapture proneness or aversion. Here, the randomized decision is taken when the tolerance limits ϕ_l or ϕ_u lie in the interval (4).

Rule 1R. We consider the following steps involving randomized decisions to infer about the behavioral classification of a given population.

Step 1: Carry out a randomized trial based on a Bernoulli r.v., say *Xp*, with the following probability function in favour of recapture proneness of the given population.

$$\psi_p(\hat{c}) = \begin{cases} 1 & \textit{if} \quad \dfrac{2\hat{c}}{1+\pi} > \phi_u \\ \delta_p & \textit{if} \quad \dfrac{2\hat{c}}{1+\pi} \leq \phi_u < \dfrac{(1+\pi)\hat{c}}{2\pi} \\ 0 & \textit{if} \quad \dfrac{(1+\pi)\hat{c}}{2\pi} \leq \phi_u \end{cases},$$

where

$$\delta_p = \max\left\{0, 1 - \left(\phi_u - \frac{2\hat{c}}{1+\pi}\right) \Big/ \left(\frac{(1+\pi)\hat{c}}{2\pi} - \frac{2\hat{c}}{1+\pi}\right)\right\},$$

Step 2: If the given population is not found to be recapture prone in Step 1 that is, if X_p is not observed to be 1, carry out another randomized trial based on Bernoulli r.v., say, X_a with the following probability function in favor of recapture aversion.

$$\psi_a(\hat{c}) = \begin{cases} 1 & \mathit{if} \quad \dfrac{(1+\pi)\hat{c}}{2\pi} < \phi_l \\ \delta_a & \mathit{if} \quad \dfrac{2\hat{c}}{1+\pi} \le \phi_l \le \dfrac{(1+\pi)\hat{c}}{2\pi} \\ 0 & \mathit{if} \quad \dfrac{2\hat{c}}{1+\pi} > \phi_l, \end{cases},$$

where

$$\delta_a = \left(\phi_l - \frac{2\hat{c}}{1+\pi}\right) \Big/ \left(\frac{(1+\pi)\hat{c}}{2\pi} - \frac{2\hat{c}}{1+\pi}\right).$$

Step 3: If the given population is not found to be recapture averse in Step 2 that is, if X_a is observed to be 0, the given population is classified as list-independent.

It is to be noted that when the above probability $\psi_p(\hat{c})$ ($\psi_a(\hat{c})$) in Step 1 (Step 2) is neither 1 nor 0, one has to perform a Bernoulli experiment with probability of recapture proneness (recapture aversion) equal to δ_p (δ_a), in order to classify whether a given population is *recapture prone* (*averse*).

As we have defined Rules *1L*, *1U* and *1R* for Proposal 1, one can analogously define Rules *2L*, *2U* and *2R* for Proposal 2 based on the bounds in Equation (5). Similarly, Rules *3L*, *3U* and *3R* can be defined for Proposal 3 based on the bounds in Equation (6). We omit the details and proceed to an empirical evaluation of these rules.

The probabilities for considering an individual as *recapture prone (RP), recapture averse (RA), or list-independent (LI)* can be obtained effectively by computing the probabilities $\Pr(X_p = 1)$, $\Pr(X_a = 1, X_p = 0)$ and $\Pr(X_a = 0, X_p = 0)$ respectively. These probabilities are computed based on the asymptotic normality of the *mle* $\hat{c}$. These probabilities, corresponding to the randomized rule under proposal 2 (i.e., Rule *2R*), are presented in the following theorem and the associated proof is sketched in the Appendix.

Theorem 1. For a large population, probabilities for considering an individual to be recapture prone, recapture averse or list-independent under Rule 2R respectively, are as follows:

$$Pr\,(RP) = 1 - \left[\delta_p F(\phi_u\sqrt{\pi}) - (1-\delta_p)F\left(\frac{(1+\pi)\phi_u}{2}\right)\right],$$

$$Pr\,(RA) = \delta_a F\left(\frac{(1+\pi)\phi_l}{2}\right) + (1-\delta_a)F(\phi_l\sqrt{\pi}) \quad \mathit{if} \quad \frac{(1+\pi)\phi_l}{2} < \phi_u\sqrt{\pi},$$

$$(1-\delta_a)F(\phi_l\sqrt{\pi}) + \delta_a(1-\delta_p)F\left(\frac{(1+\pi)\phi_l}{2}\right) + \delta_a\delta_p F(\phi_u\sqrt{\pi})$$

$$\mathit{if} \quad \frac{(1+\pi)\phi_l}{2} > \phi_u\sqrt{\pi}$$

$$
\begin{aligned}
Pr\,(LI) = &\left[\delta_p F(\phi_u\sqrt{\pi}) + (1-\delta_p)F\left(\frac{(1+\pi)\phi_u}{2}\right)\right] \\
&- \left[\delta_a F\left(\frac{(1+\pi)\phi_l}{2}\right) - (1-\delta_a)F(\phi_l\sqrt{\pi})\right] \;\; if \;\; \frac{(1+\pi)\phi_l}{2} < \phi_u\sqrt{\pi}, \\
&\left[\delta_p(1-\delta_a)F(\phi_u\sqrt{\pi}) + (1-\delta_p)F\left(\frac{(1+\pi)\phi_u}{2}\right)\right] \\
&- \left[\delta_a(1-\delta_p)F\left(\frac{(1+\pi)\phi_l}{2}\right) + (1-\delta_a)F(\phi_l\sqrt{\pi})\right] \;\; if \;\; \frac{(1+\pi)\phi_l}{2} > \phi_u\sqrt{\pi},
\end{aligned}
$$

where $F(\cdot)$ refers the cumulative distribution function of normal variate $\hat{c}$ with asymptotic mean and variance are c and $\sigma_{\hat{c}^2}$, respectively.

Theorem 1 is useful for applied work in that it provides a quite important and reasonably simple empirical strategy for detecting behavioral dependence without any need for additional information. All calculations are made here under the normality assumption that is satisfied for reasonably large populations. Note that for some configuration of $\mathrm{p}'_{ij}s$ (as defined in the introduction), the boundary constraints in the probability calculations may become binding. In such cases, the conclusion from Theorem 1 would be approximate in nature. Apart from such cases, the strategy will work well. A graphical comparative study between the three probabilities computed in Theorem 1 is carried out in Subsection 3.2.

3. Simulation Study

3.1. *Evaluation of Classification Rules*

In this section, we perform an extensive simulation study for comparing the performances of the proposed classification strategies. Let us consider ten simulated populations, comprising three populations for each of the three absolute difference values, (0.1, 0.15, 0.20), between $p_{1\cdot}$ and $p_{\cdot 1}$ and one additional population with very high capture probabilities ($p_{1\cdot} = 0.95, p_{\cdot 1} = 0.85$). Further, we also consider three values of ϕ, viz. 1.50, 0.60, 1.00, in order to represent three situations of behavioral dependence, (*i*) recapture proneness, (*ii*) recapture aversion, and (*iii*) list-independence, respectively. These simulated populations for each of three said ϕ values together encompass all possible combinations with true population size, $N = 1{,}000$. The compositions of the simulated populations are shown in Table 2. The true value of the parameter c, calculated using the given $p_{1\cdot}, p_{\cdot 1}$ and ϕ, is also presented corresponding to each of the ten simulated populations. Tables 3, 4 and 5 present the performance evaluation of the classification strategies based on Proposals 1, 2 and 3, respectively, developed in Section 2, in terms of *correct classification rate* (CCR) of the underlying directional nature of ϕ. CCR is presented in percentage (%) after computing the number of correct classifications out of 5,000 replications.

From the simulation analysis we observed that $\pi = 1/3$, 1/4 have overall better results across the three proposals (1, 2 and 3) and three dependence situations (proneness, aversion and list-independence). Hence, we make a suggestion for choosing $\pi = 1/3$ or

Table 2. Compositions of simulated populations with N = 1,000.

		Recapture prone		Recapture averse		Causally independent	
		$\phi = 1.50$		$\phi = 0.60$		$\phi = 1.00$	
$p_{1\cdot}$	$p_{\cdot 1}$	Population	c	Population	c	Population	c*
0.50	0.60	P1	0.720	A1	0.450	I1	0.60
0.70	0.60	P2	0.667	A2	0.500	I2	0.60
0.70	0.80	P3	0.889	A3	0.667	I3	0.80
0.45	0.60	P4	0.735	A4	0.439	I4	0.60
0.70	0.55	P5	0.611	A5	0.458	I5	0.55
0.80	0.65	P6	0.696	A6	0.574	I6	0.65
0.45	0.65	P7	0.796	A7	0.476	I7	0.65
0.75	0.55	P8	0.600	A8	0.471	I8	0.55
0.85	0.65	P9	0.684	A9	0.591	I9	0.65
0.95	0.85	P10	0.864	A10	0.823	I9	0.85

*Under list-independence, that is, $\phi = 1$, $c = p_{\cdot 1}$

1/4 based on empirical performance. It is noted that the upper bound for Proposal 2 is same as the lower bound for Proposal 3 and their union is the bounds for Proposal 1. Hence, we first compare Proposals 2 and 3, which are based on disjoint intervals. By construction, Proposal 2 (Proposal 3) goes against the *recapture prone* (*averse*) conclusions for the populations. Thus, comparison on the basis of performances in case of truly *list-independent* populations will be relatively unbiased. On the basis of empirical performance, Proposal 2 is found to be better than Proposal 3. In addition, it is observed across the simulated populations that Proposal 2 also performs better than Proposal 1. Overall, we can conclude that Proposal 2 outperforms Proposal 1 which is indeed better than Proposal 3. Choice of the middle values for π (e.g., 1/3, 1/4) are better than the extreme values (e.g., 2/5, 1/5). Henceforth, for real data analysis, we prescribe the use of Proposal 2 or 1 with π values in the range of $\{1/3, 1/4\}$.

While comparing the rules L, U and R over the chosen proposals and the values for π, we see that rules R have a more balanced performance in terms of CCR than that of L and U rules. Performance of rule R for the best choice of proposal (i.e., Proposal 2) would be better understood through further analysis sketched in the next subsection.

3.2. *Performance Study of Rule 2R*

Graphical analysis of the probabilities computed based on the result in Theorem 1 for the randomized classification Rule *2R* is presented in Figure 1.

Here we consider the same combinations of $(p_{1\cdot}, p_{\cdot 1})$, which are considered in Table 2 in Subsection 3.1 except the 10th combination (0.95, 0.85), with varying ϕ values over the domain (0.40, 2.00). In general, the performance of Rule *2R* is seen to be quite good, except for a few combinations of $(p_{1\cdot}, p_{\cdot 1})$ values. Overall, we can see that our theoretical intuition put forth in the discussion after Theorem 1 and the findings in Subsection 3.1 are carried over here.

Table 3. Evaluation of the classification rules based on Proposal 1.

	Rule 1L	Rule 1U	Rule 1R		Rule 1L	Rule 1U	Rule 1R		Rule 1L	Rule 1U	Rule 1R
Population	CCR	CCR	CCR	Population	CCR	CCR	CCR	Population	CCR	CCR	CCR
					$\pi = 2/5$						
P1	0.74	100.00	67.86	A1	100.00	99.78	99.98	I1	8.90	90.08	75.14
P2	0.00	98.02	31.18	A2	100.00	76.90	97.32	I2	5.64	93.74	75.76
P3	100.00	100.00	100.00	A3	100.00	76.90	97.32	I3	2.52	0.00	0.12
P4	4.28	100.00	77.16	A4	100.00	99.92	100.00	I4	9.60	89.08	75.20
P5	0.00	16.66	1.68	A5	100.00	99.92	100.00	I5	0.00	97.00	36.06
P6	0.00	100.00	54.72	A6	99.96	0.02	43.94	I6	88.26	10.32	80.62
P7	91.36	100.00	99.38	A7	100.00	94.66	99.38	I7	81.20	17.34	77.74
P8	0.00	5.38	0.40	A8	100.00	99.00	99.94	I8	0.00	97.26	35.62
P9	0.00	99.98	45.22	A9	99.10	0.00	29.96	I9	88.80	9.66	80.78
P10	100.00	100.00	100.00	A10	99.10	0.00	29.96	I10	0.00	0.00	0.00
					$\pi = 1/3$						
P1	26.24	100.00	92.60	A1	100.00	50.34	92.16	I1	49.86	1.18	62.28
P2	0.00	100.00	68.08	A2	100.00	0.30	60.24	I2	49.12	0.52	63.00
P3	100.00	100.00	100.00	A3	100.00	0.30	60.24	I3	0.00	0.00	0.00
P4	53.90	100.00	96.14	A4	100.00	68.18	95.68	I4	50.66	1.72	61.78
P5	0.00	99.92	39.98	A5	100.00	32.40	88.80	I5	0.44	50.22	66.50
P6	1.26	100.00	83.54	A6	93.28	0.00	13.60	I6	99.88	0.00	39.20
P7	99.94	100.00	100.00	A7	100.00	13.74	77.40	I7	98.54	0.00	40.00
P8	0.00	99.58	33.06	A8	100.00	11.66	80.40	I8	0.28	50.68	66.02
P9	0.06	100.00	77.70	A9	70.60	0.00	5.96	I9	99.80	0.00	38.78
P10	100.00	100.00	100.00	A10	70.60	0.00	5.96	I10	0.00	0.00	0.00

Table 3. Continued.

	Rule 1L	Rule 1U	Rule 1R		Rule 1L	Rule 1U	Rule 1R		Rule 1L	Rule 1U	Rule 1R
Population	CCR	CCR	CCR	Population	CCR	CCR	CCR	Population	CCR	CCR	CCR
					$\pi = 1/4$						
P1	94.74	100.00	99.84	A1	100.00	0.00	41.44	I1	96.06	0.00	27.22
P2	12.16	100.00	93.02	A2	99.94	0.00	15.48	I2	97.48	0.00	27.30
P3	100.00	100.00	100.00	A3	99.94	0.00	15.48	I3	0.00	0.00	0.00
P4	98.66	100.00	99.96	A4	100.00	0.00	46.92	I4	94.94	0.00	27.56
P5	0.00	100.00	76.64	A5	100.00	0.00	35.12	I5	25.74	0.00	43.02
P6	71.02	100.00	99.20	A6	25.82	0.00	0.58	I6	98.86	0.00	11.56
P7	100.00	100.00	100.00	A7	100.00	0.00	26.28	I7	95.42	0.00	11.64
P8	0.00	100.00	73.46	A8	100.00	0.00	29.12	I8	24.68	0.00	43.42
P9	41.58	100.00	97.72	A9	4.50	0.00	0.08	I9	98.98	0.00	10.78
P10	100.00	100.00	100.00	A10	4.50	0.00	0.08	I10	0.00	0.00	0.00
					$\pi = 1/5$						
P1	99.86	100.00	100.00	A1	99.98	0.00	14.96	I1	99.32	0.00	13.44
P2	64.38	100.00	98.96	A2	98.30	0.00	4.32	I2	99.88	0.00	13.70
P3	100.00	100.00	100.00	A3	98.30	0.00	4.32	I3	0.00	0.00	0.00
P4	99.94	100.00	100.00	A4	100.00	0.00	17.86	I4	0.42	0.00	13.20
P5	0.44	100.00	89.44	A5	100.00	0.00	12.04	I5	30.12	0.00	26.06
P6	98.84	100.00	99.96	A6	3.02	0.00	0.00	I6	0.00	0.00	2.80
P7	100.00	100.00	100.00	A7	99.58	0.00	8.56	I7	0.00	0.00	3.26
P8	0.02	100.00	86.62	A8	100.00	0.00	9.22	I8	29.30	0.00	26.14
P9	93.42	100.00	99.90	A9	0.08	0.00	0.00	I9	0.00	0.00	2.74
P10	100.00	100.00	100.00	A10	0.08	0.00	0.00	I10	0.00	0.00	0.00

Table 4. Evaluation of the classification rules based on Proposal 2.

	Rule 2L	Rule 2U	Rule 2R		Rule 2L	Rule 2U	Rule 2R		Rule 2L	Rule 2U	Rule 2R
Population	CCR	CCR	CCR	Population	CCR	CCR	CCR	Population	CCR	CCR	CCR
					π = 2/5						
P1	0.74	89.30	37.70	A1	100.00	100.00	100.00	I1	8.90	92.08	52.74
P2	0.00	5.36	0.64	A2	100.00	100.00	100.00	I2	5.64	95.04	53.12
P3	100.00	100.00	100.00	A3	100.00	100.00	100.00	I3	2.52	0.00	0.26
P4	4.28	96.66	53.72	A4	100.00	100.00	100.00	I4	9.60	90.96	52.06
P5	0.00	0.00	0.00	A5	100.00	100.00	100.00	I5	0.00	15.50	2.94
P6	0.00	51.60	9.92	A6	99.96	39.78	83.40	I6	88.26	99.78	98.54
P7	91.36	100.00	98.80	A7	100.00	100.00	100.00	I7	81.20	97.94	95.90
P8	0.00	0.00	0.00	A8	100.00	100.00	100.00	I8	0.00	14.56	2.74
P9	0.00	23.20	3.42	A9	99.10	9.92	59.76	I9	88.80	99.78	98.62
P10	100.00	100.00	100.00	A10	99.10	9.92	59.76	I10	0.00	0.00	0.00
					π = 1/3						
P1	26.24	100.00	84.88	A1	100.00	99.94	99.98	I1	49.86	94.06	88.50
P2	0.00	95.92	35.52	A2	100.00	84.36	97.84	I2	49.12	96.98	90.00
P3	100.00	100.00	100.00	A3	100.00	84.36	97.84	I3	0.00	0.00	0.00
P4	53.90	100.00	92.86	A4	100.00	99.96	100.00	I4	50.66	93.72	88.94
P5	0.00	9.22	1.08	A5	100.00	99.96	100.00	I5	0.44	94.64	42.16
P6	1.26	100.00	65.02	A6	93.28	0.14	30.58	I6	99.88	18.64	80.48
P7	99.94	100.00	100.00	A7	100.00	96.80	99.62	I7	98.54	25.56	78.24
P8	0.00	2.66	0.16	A8	100.00	99.54	99.96	I8	0.28	94.98	41.34
P9	0.06	99.88	54.24	A9	70.60	0.00	14.04	I9	99.80	17.80	81.12
P10	100.00	100.00	100.00	A10	70.60	0.00	14.04	I10	0.00	0.00	0.00

Table 4. Continued.

	Rule 2L	Rule 2U	Rule 2R		Rule 2L	Rule 2U	Rule 2R		Rule 2L	Rule 2U	Rule 2R
Population	CCR	CCR	CCR	Population	CCR	CCR	CCR	Population	CCR	CCR	CCR
					$\pi = 1/4$						
P1	94.74	100.00	99.54	A1	100.00	50.34	90.40	I1	96.06	1.18	57.70
P2	12.16	100.00	85.50	A2	99.94	0.30	49.90	I2	97.48	0.52	57.70
P3	100.00	100.00	100.00	A3	99.94	0.30	49.90	I3	0.00	0.00	0.00
P4	98.66	100.00	99.92	A4	100.00	68.18	94.50	I4	94.94	1.72	58.18
P5	0.00	99.92	50.22	A5	100.00	32.40	86.16	I5	25.74	50.22	80.50
P6	71.02	100.00	98.12	A6	25.82	0.00	2.24	I6	98.86	0.00	24.90
P7	100.00	100.00	100.00	A7	100.00	13.74	72.22	I7	95.42	0.00	24.54
P8	0.00	99.58	41.44	A8	100.00	11.66	76.82	I8	24.68	50.68	80.78
P9	41.58	100.00	94.46	A9	4.50	0.00	0.22	I9	98.98	0.00	24.22
P10	100.00	100.00	100.00	A10	4.50	0.00	0.22	I10	0.00	0.00	0.00
					$\pi = 1/5$						
P1	99.86	100.00	100.00	A1	99.98	1.60	58.16	I1	99.32	0.00	30.64
P2	64.38	100.00	97.36	A2	98.30	0.00	23.00	I2	99.88	0.00	31.14
P3	100.00	100.00	100.00	A3	98.30	0.00	23.00	I3	0.00	0.00	0.00
P4	99.94	100.00	100.00	A4	100.00	5.62	67.26	I4	0.42	0.00	30.34
P5	0.44	100.00	75.22	A5	100.00	0.12	52.34	I5	30.12	0.12	55.58
P6	98.84	100.00	99.96	A6	3.02	0.00	0.08	I6	0.00	0.00	6.86
P7	100.00	100.00	100.00	A7	99.58	0.02	38.92	I7	0.00	0.00	7.64
P8	0.02	100.00	69.76	A8	100.00	0.02	42.44	I8	29.30	0.12	56.86
P9	93.42	100.00	99.74	A9	0.08	0.00	0.00	I9	0.00	0.00	6.30
P10	100.00	100.00	100.00	A10	0.08	0.00	0.00	I10	0.00	0.00	0.00

Table 5. Evaluation of the classification rules based on Proposal 3.

	Rule 3L	Rule 3U	Rule 3R		Rule 3L	Rule 3U	Rule 3R		Rule 3L	Rule 3U	Rule 3R
Population	CCR	CCR	CCR	Population	CCR	CCR	CCR	Population	CCR	CCR	CCR
					$\pi = 2/5$						
P1	89.30	100.00	98.12	A1	100.00	99.78	99.98	I1	92.08	90.08	96.84
P2	5.36	98.02	58.86	A2	100.00	76.90	94.96	I2	95.04	93.74	98.42
P3	100.00	100.00	100.00	A3	100.00	76.90	94.96	I3	0.00	0.00	0.00
P4	96.66	100.00	99.42	A4	100.00	99.92	100.00	I4	90.96	89.08	96.42
P5	0.00	16.66	2.92	A5	100.00	99.92	100.00	I5	15.50	97.00	64.56
P6	51.60	100.00	90.86	A6	39.78	0.02	8.12	I6	99.78	10.32	64.60
P7	100.00	100.00	100.00	A7	100.00	94.66	98.96	I7	97.94	17.34	64.30
P8	0.00	5.38	0.76	A8	100.00	99.00	99.90	I8	14.56	97.26	64.18
P9	23.20	99.98	79.76	A9	9.92	0.00	1.40	I9	99.78	9.66	65.04
P10	100.00	100.00	100.00	A10	9.92	0.00	1.40	I10	0.00	0.00	0.00
					$\pi = 1/3$						
P1	100.00	100.00	100.00	A1	99.94	50.34	86.18	I1	94.06	1.18	38.08
P2	95.92	100.00	99.60	A2	84.36	0.30	28.10	I2	96.98	0.52	38.26
P3	100.00	100.00	100.00	A3	84.36	0.30	28.10	I3	0.00	0.00	0.00
P4	100.00	100.00	100.00	A4	99.96	68.18	92.00	I4	93.72	1.72	39.44
P5	9.22	99.92	72.84	A5	99.96	32.40	80.10	I5	94.64	50.22	89.26
P6	100.00	100.00	100.00	A6	0.14	0.00	0.00	I6	18.64	0.00	2.08
P7	100.00	100.00	100.00	A7	96.80	13.74	59.50	I7	25.56	0.00	4.06
P8	2.66	99.58	60.68	A8	99.54	11.66	65.14	I8	94.98	50.68	89.16
P9	99.88	100.00	100.00	A9	0.00	0.00	0.00	I9	17.80	0.00	1.70
P10	100.00	100.00	100.00	A10	0.00	0.00	0.00	I10	0.00	0.00	0.00

Table 5. Continued.

	Rule 3L	Rule 3U	Rule 3R		Rule 3L	Rule 3U	Rule 3R		Rule 3L	Rule 3U	Rule 3R
Population	CCR	CCR	CCR	Population	CCR	CCR	CCR	Population	CCR	CCR	CCR
					$\pi = 1/4$						
P1	100.00	100.00	100.00	A1	50.34	0.00	8.88	I1	1.18	0.00	0.08
P2	100.00	100.00	100.00	A2	0.30	0.00	0.04	I2	0.52	0.00	0.04
P3	100.00	100.00	100.00	A3	0.30	0.00	0.04	I3	0.00	0.00	0.00
P4	100.00	100.00	100.00	A4	68.18	0.00	16.02	I4	1.72	0.00	0.14
P5	99.92	100.00	99.98	A5	32.40	0.00	3.98	I5	50.22	0.00	6.56
P6	100.00	100.00	100.00	A6	0.00	0.00	0.00	I6	0.00	0.00	0.00
P7	100.00	100.00	100.00	A7	13.74	0.00	1.78	I7	0.00	0.00	0.00
P8	99.58	100.00	99.96	A8	11.66	0.00	1.16	I8	50.68	0.00	6.52
P9	100.00	100.00	100.00	A9	0.00	0.00	0.00	I9	0.00	0.00	0.00
P10	100.00	100.00	100.00	A10	0.00	0.00	0.00	I10	0.00	0.00	0.00
					$\pi = 1/5$						
P1	100.00	100.00	100.00	A1	1.60	0.00	0.04	I1	0.00	0.00	0.00
P2	100.00	100.00	100.00	A2	0.00	0.00	0.00	I2	0.00	0.00	0.00
P3	100.00	100.00	100.00	A3	0.00	0.00	0.00	I3	0.00	0.00	0.00
P4	100.00	100.00	100.00	A4	5.62	0.00	0.26	I4	0.00	0.00	0.00
P5	100.00	100.00	100.00	A5	0.12	0.00	0.02	I5	0.12	0.00	0.02
P6	100.00	100.00	100.00	A6	0.00	0.00	0.00	I6	0.00	0.00	0.00
P7	100.00	100.00	100.00	A7	0.02	0.00	0.00	I7	0.00	0.00	0.00
P8	100.00	100.00	100.00	A8	0.02	0.00	0.00	I8	0.12	0.00	0.00
P9	100.00	100.00	100.00	A9	0.00	0.00	0.00	I9	0.00	0.00	0.00
P10	100.00	100.00	100.00	A10	0.00	0.00	0.00	I10	0.00	0.00	0.00

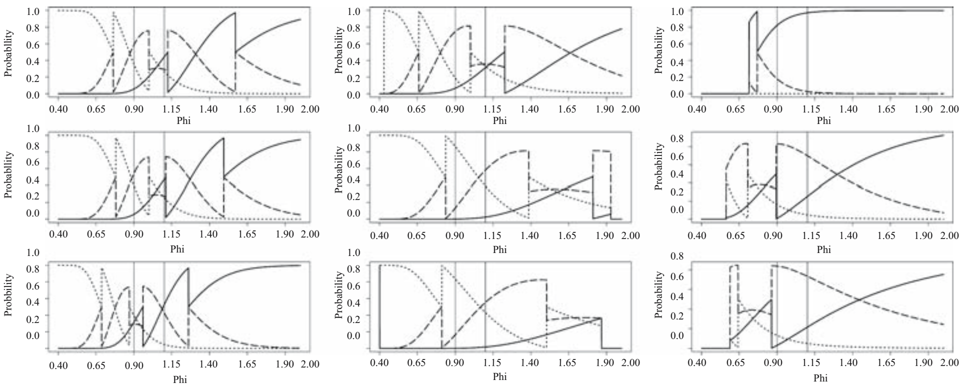


Fig. 1. Graphical comparison of the probabilities of recapture proneness, recapture aversion & list-independence under the Rule 2R, associated with the Proposal 2, with $\pi = 1/3$ over different ϕ values. Continuous, dotted and longdash lines respectively refer the probabilities corresponding to recapture proneness, recapture aversion and list-independence.

4. Real Data Illustration

In this section, we illustrate the classification strategies formulated under the two selected proposals (i.e., Proposals 1 and 2) described in Section 2 through the application on three real data sets from three different fields – demography, crime statistics and social sciences.

Firstly, we consider the Malawi death data obtained from a Population Change Survey to estimate birth, death and migration rates conducted by the National Statistical Office in Malawi between 1970 and 1972. Greenfield (1975) introduced this data set. Later, it has also been used by Nour (1982) and Chatterjee and Mukherjee (2016c). Very large values of $\hat{c}$ for all the strata, except Lilongwe, clearly indicate recapture proneness. Thus, we consider the data for Lilongwe and Other Urban Areas (see top panel of Table 6) for comparative analysis of the proposed classification rules.

Secondly, we use Homicide data, which are analyzed by Eckberg (2000) to establish the utility of dual enumeration methods for estimating the total number of unrecorded murders in South Carolina, 1877–1878. This interesting work was meant for tracing the historic trends in homicide based on the two sources alone and the author claimed that the popular

Table 6. Three real data sets in DRS format which are used for the classification analysis.

		Count			
		List 1	List 2	Matched	Estimate of c
Data	Populations	$x_{1.}$	$x_{.1}$	x_{11}	$\hat{c}$
Malawi death	Lilongwe	324	216	192	0.593
	Other Urban Areas	1960	2450	1645	0.839
Homicide*	Zero-Two	29	102	23	0.793
	Three-Five	56	74	42	0.750
	Six	50	43	32	0.640
Handloom	Ward No. 2	126	107	85	0.675
	Ward No. 16	131	103	50	0.382

*Each of the three populations belong to this data set are named in terms of holding index score.

ChandraSekar-Deming estimator would face a formidable undercount problem. This happens due to a possible positive correlation between two data sources – (i) South Carolina Department of Archives and History and (ii) News and Courier Reports. Dual system data were available for all 33 counties in South Carolina. Following the proposal of ChandraSekar and Deming (1949), all counties (except Charleston) are divided into three homogeneous groups based on a 0–6 point index scale that measures the thoroughness of county archives. For more details on the data source and index scaling, readers are referred to Eckberg (2000, 5–9). Data in DRS form are presented in the middle panel of Table 6.

In addition to the above two data sets, we consider another data set on *handloom workers*. This new data are from a survey aimed to estimate the undercount in the census of handloom workers (master weavers and labours only) attached to the handloom industry residing at Gangarampur in South Dinajpur district of the state West Bengal, India in 2013. Handloom products have a rich tradition in this state and the handloom industry occupies a place second only to agriculture in providing livelihood to the people. In the urban area of Gangarampur, there are sixteen wards; two of these wards (Wards No. 2 and 16) are selected for Post Enumeration Survey (PES) to evaluate the coverage in the original census (SOSU 2014). Surprisingly, the nature of the data on these two wards are found to be different in terms of recapture proportion. Since this counting task is meant for the benefit of the workers attached to the handloom industry, a general thought is that the census and the PES might be positively related. On the other hand, some people may consider that one time enrollment is enough. So, if one is counted at the time of census, he/she may be reluctant at the time of the PES and that implies ϕ to be less than 1. Thus, in both of the possibilities, the ChandraSekar-Deming estimator would not be appropriate. Surveyors reported that workers in Ward No. 16, which is very close to the town center, might be somewhat reluctant to enlist themselves a second time (i.e., at the time of the PES). Moreover, most of them work outside (other districts) and usually come home during particular seasons. That is why Ward No. 16 shows in very low matches compared to Ward No. 2. Moreover, the beliefs of the experts of the Textile Directorate of the Government of West Bengal also drive the idea that the ChandraSekar-Deming estimates (157 and 270 respectively) fail to extract the sizes with precision. However, they expect that the ChandraSekar-Deming method yields a slight undercount for Ward No. 2 and a high overcount for Ward No. 16. The bottom panel of Table 6 presents the DRS data for these two wards.

Table 6 presents the list-wise counts and matched records for each of the three data sets mentioned above. In addition, we also present values of the key statistic $\hat{c} = (x_{11}/\mathrm{x}_{1\cdot})$ for each data set in Table 6. Results on the directional classification for the above three data sets are presented in Table 7. In addition, estimates of the probabilities associated with the directional classifications found are presented for each of the seven real populations in Table 7. These estimated probabilities can be treated as a measure of chance (or uncertainty) behind the directional classifications found.

In the light of the findings from the selected strategies (under proposals 1 and 2 with $\pi = (1/3,\ 1/4)$, the populations Lilongwe and Other Urban Areas are classified as list-independent and recapture prone, respectively. The classification result for Lilongwe is quite interesting, as it indicates a difference from the conventional assumption of positive dependence in DRS on demographic application. All the populations under the Homicide data set exhibit recapture proneness except for the *holding index score 6*, for which rules L

Table 7. Results from the two selected methods for directional classification strategies referring to the status of behavioral dependence in real DRS data sets. Probabilities corresponding to the resulting directional classifications are provided in (). The classification 'list-independent' is written here as 'Indep'.

	Malawi death		Homicide			Handloom	
Rules	Lilongwe	Other urban areas	Zero-two	Three-five	Six	Ward No. 2	Ward No. 16
			Proposal 1 (with $\pi = 1/3$)				
Rule 1L	Averse (0.54)	Prone (0.99)	Prone (0.65)	Prone (0.58)	Indep (0.35)	Indep (0.52)	Averse (0.99)
Rule 1U	Prone (0.80)	Prone (1.00)	Prone (0.89)	Prone (0.92)	Prone (0.76)	Prone (0.92)	Averse (0.85)
Rule 1R	Indep (0.75)	Prone (0.99)	Prone (0.65)	Prone (0.58)	Prone (0.56)	Prone (0.75)	Averse (0.85)
			Proposal 1 (with $\pi = 1/4$)				
Rule 1L	Indep (0.68)	Prone (1.00)	Prone (0.73)	Prone (0.69)	Indep (0.33)	Indep (0.42)	Averse (0.99)
Rule 1U	Prone (0.99)	Prone (1.00)	Prone (0.96)	Prone (0.98)	Prone (0.93)	Prone (0.99)	Indep (0.46)
Rule 1R	Prone (0.74)	Prone (1.00)	Prone (0.73)	Prone (0.69)	Prone (0.86)	Prone (0.98)	Averse (0.90)
			Proposal 2 (with $\pi = 1/3$)				
Rule 2L	Averse (0.54)	Prone (0.99)	Prone (0.65)	Prone (0.58)	Indep (0.35)	Indep (0.52)	Averse (0.99)
Rule 2U	Indep (0.69)	Prone (1.00)	Prone (0.80)	Prone (0.80)	Prone (0.55)	Prone (0.68)	Averse (0.98)
Rule 2R	Indep (0.95)	Prone (0.99)	Prone (0.65)	Prone (0.58)	Indep (0.62)	Indep (0.53)	Averse (0.98)
			Proposal 2 (with $\pi = 1/4$)				
Rule 2L	Indep (0.68)	Prone (1.00)	Prone (0.73)	Prone (0.69)	Indep (0.33)	Indep (0.43)	Averse (0.98)
Rule 2U	Prone (0.80)	Prone (1.00)	Prone (0.89)	Prone (0.92)	Prone (0.76)	Prone (0.92)	Averse (0.85)
Rule 2R	Indep (0.69)	Prone (1.00)	Prone (0.73)	Prone (0.69)	Prone (0.66)	Prone (0.88)	Averse (0.85)
Overall	Indep	Prone	Prone	Prone	Prone	Prone	Averse

find no dependence. For Ward No. 16 under the handloom data set, the strategies agree and classify the ward as recapture averse. However, inference for the population in Ward No. 2 is in favor of recapture proneness. Indeed, in the conflicting cases, we recommend going with the classification made by the randomized Rule *2R*, as Proposal 2 along with randomized rule (R) is found to be better in the simulation based comparative study performed earlier. In the last row of Table 7, we present the overall classification status of the population in each real data set and that is found to be identical with the inference from Rule *2R*.

It is to be noted that inference about the direction of possible behavioral dependence in any data set drawn in literature may not match with the conclusions of the classification strategies prescribed in this present article.

5. Conclusion

From the extensive literature on capture-recapture data analysis on human population, it is quite clear that assumption of *list-independence* does not hold satisfactorily in many instances. As far as homogeneous human population size estimation is concerned, two-sample capture-recapture experiments are very common and the model M_{tb} is best suited. However, this model seriously suffers from the non-identifiability problem and analyses in the literature suggest that the availability of knowledge on the direction of behavioral dependency could improve the inference to a great extent. Eliciting such information is crucial in DRS. To address this issue, we develop several comparable strategies for classification of the given population in terms of the direction of dependence (i.e., whether the given population is *recapture prone* or *averse*) under a mild and realistic assumption. Among the three proposals on bounds for unknown ϕ, Proposal 2 is found to be better suited. On the other hand, the classification strategy based on randomized technique (R) is quite appealing for the development of more efficient inference in the context of M_{tb}-DRS. Hence, the strategy, Rule *2R*, is quite accurate except for particular situations with small recapture probabilities. In real life applications, this strategy provides us with a useful tool for classifying human populations.

With the knowledge inferred in this article on the direction of dependence in DRS, model M_{tb} can be successfully analyzed in the Bayesian paradigm. For example, Chatterjee and Mukherjee (2016c, 2018) have shown that availability of knowledge on the direction of behavioral dependence helps to gain precision in the estimation of population size. One can also analogously improve the Bayesian methodology proposed in Lee and Chen (1998) and Lee et al. (2003) by choosing suitable priors using our strategies. Also, inference on population size assuming any particular dependence type (i.e., recapture prone or averse) can be improved based on bound or inequality. For example, in Greenfield (1975) and Nour (1982), recapture proneness is assumed for demographic data. Their assumption can be verified using our methodology. For both of these approaches, the knowledge on the possible direction of dependence gained from the methods proposed in our present article results in more precise inference. Extension of the proposed behavioral classification strategies may be possible for more than two capture occasions.

Throughout the article, we assume that ordering of the two samples (or lists) in DRS are known, that is, one list is completely prepared before the other. However in some cases, listing by two sources are either parallel or their ordering is not known. In such cases, consider a typical situation where the number of recaptured individuals (x_{11}) is small compared to the number of one-time captured individuals, that is x_{10} or x_{01}. Then, our proposed classification results may differ if ordering of the lists is interchanged as observed in the case of Homicide data, with holding index score Zero-Two. In this case, reversal of ordering for the lists will lead to a conclusion of recapture aversion instead of proneness. As per the definition of grouping, these data exhibit lack of thoroughness of the county archives and hence, x_{11} is small compared to x_{10} or x_{01}.

6. Appendix

6.1. Proof of Theorem 1

Firstly, $Pr(RP) = Pr(X_P = 1)$. Therfore,

$$\begin{aligned} Pr(RP) &= Pr\left(\frac{2\hat{c}}{1+\pi} > \phi_u\right) + \delta_p Pr\left(\frac{2\hat{c}}{1+\pi} \leq \phi_u < \hat{c}\pi^{-1/2}\right) \\ &= 1 - F\left(\frac{(1+\pi)\phi_u}{2}\right) + \delta_p\left[F\left(\frac{(1+\pi)\phi_u}{2}\right) - F\left(\phi_u \pi^{1/2}\right)\right] \\ &= 1 - \left[\delta_p F\left(\phi_u \pi^{1/2}\right) + \left(1-\delta_p\right)F\left(\frac{(1+\pi)\phi_u}{2}\right)\right], \end{aligned}$$

where $F(x)$ denoted as the cumulative distribution function of the normal variate $\hat{c}$ at x such that $\hat{c}$ has mean c and variance equals to $\sigma_{\hat{c}}^2 = V(\hat{c})$. This distribution for $\hat{c}$ is asymptotic. Thus, $F(x) = \Phi\left(\frac{x-c}{\sigma_{\hat{c}}}\right)$, where Φ has its usual meaning, that is, cumulative distribution function of a standard normal variate.

Secondly, $Pr(RA) = Pr(X_p = 0, X_a = 1)$. Therefore,

$$\begin{aligned} Pr(RA) \;=\;& Pr\left(\hat{c}\pi^{-1/2} < \phi_l, X_p = 0\right) + \delta_a Pr\left(\frac{2\hat{c}}{1+\pi} \leq \phi_l \leq \hat{c}\pi^{-1/2}, X_p = 0\right) \\ =\;& Pr\left(\hat{c}\pi^{-1/2} < \phi_l, \hat{c}\pi^{-1/2} \leq \phi_u\right) + \left(1-\delta_p\right)Pr\left(\hat{c}\pi^{-1/2} < \phi_l, \frac{2\hat{c}}{1+\pi} \leq \phi_u < \hat{c}\pi^{-1/2}\right) \\ & +\delta_a\left(1-\delta_p\right)Pr\left(\frac{2\hat{c}}{1+\pi} \leq \phi_l < \hat{c}\pi^{-1/2}, \frac{2\hat{c}}{1+\pi} \leq \phi_u < \hat{c}\pi^{-1/2}\right) \\ & +\delta_a Pr\left(\frac{2\hat{c}}{1+\pi} \leq \phi_l \leq \hat{c}\pi^{-1/2}, \hat{c}\pi^{-1/2} \leq \phi_u\right) \\ =\;& Pr\left(\hat{c}\pi^{-1/2} < \phi_l\right) + \left(1-\delta_p\right).0 + \delta_a Pr\left[\phi_l \pi^{1/2} \leq \hat{c} \leq \min\left(\frac{(1+\pi)\phi_l}{2}, \phi_u \pi^{1/2}\right)\right] \\ & +\delta_a\left(1-\delta_p\right)Pr\left(\phi_u \pi^{1/2} < \hat{c} \leq \frac{(1+\pi)\phi_l}{2}\right). \end{aligned}$$

If ϕ_l and ϕ_u are chosen such that $\frac{(1+\pi)\phi_l}{2} < \phi_u \pi^{1/2}$, therefore,

$$Pr(RA) = \delta_a F\left(\frac{(1+\pi)\phi_l}{2}\right) + (1-\delta_a)F\left(\phi_l \pi^{1/2}\right).$$

Lastly,

$$\begin{aligned}Pr(LI) &= Pr\left(X_p = 0,\ \ X_a = 0\right)\\ &= Pr\left(X_p = 0\right) - Pr\left(X_p = 0, X_a = 1\right)\\ &= \left[\delta_p F\left(\phi_u \pi^{1/2}\right) + \left(1-\delta_p\right)F\left(\frac{(1+\pi)\phi_u}{2}\right)\right]\\ &\quad -\left[\delta_a F\left(\frac{(1+\pi)\phi_l}{2}\right) + (1-\delta_a)F\left(\phi_l \pi^{1/2}\right)\right]\end{aligned}$$

Again, if $\frac{(1+\pi)\phi_l}{2} > \phi_u \pi^{1/2}$ holds,

$$\begin{aligned}Pr(RA) &= F\left(\phi_l \pi^{1/2}\right) + \delta_a\left[F\left(\phi_u \pi^{1/2}\right) - F\left(\phi_l \pi^{1/2}\right)\right]\\ &\quad +\delta_a\left(1-\delta_p\right)\left[F\left(\frac{(1+\pi)\phi_l}{2}\right) - F\left(\phi_u \pi^{1/2}\right)\right]\\ &= (1-\delta_a)F\left(\phi_l \pi^{1/2}\right) + \delta_a\left(1-\delta_p\right)F\left(\frac{(1+\pi)\phi_l}{2}\right) + \delta_a\delta_p F\left(\phi_u \pi^{1/2}\right)\end{aligned}$$

and similarly,

$$\begin{aligned}Pr(LI) &= \left[\delta_p(1-\delta_a)F\left(\phi_u \pi^{1/2}\right) + \left(1-\delta_p\right)F\left(\frac{(1+\pi)\phi_u}{2}\right)\right]\\ &\quad -\left[\delta_a\left(1-\delta_p\right)F\left(\frac{(1+\pi)\phi_l}{2}\right) + (1-\delta_a)F\left(\phi_l \pi^{1/2}\right)\right].\end{aligned}$$

This completes the proof.

6.2. *Mathematical Proofs and Justification on the Two Roots of ϕ*

The quadratic equation, corresponding to inequality (2), can be written as

$$\phi^2 + u\phi + v = 0, \tag{S1}$$

where $u = \frac{kx_{\cdot 1} - (\pi+1)x_{11}}{\pi x_{1\cdot}}$ and $v = \hat{c}^2\pi^{-1}$. This quadratic equation (S1) has two roots, ϕ_0 and ϕ_1, which satisfy

$$\phi_0 + \phi_1 = -u = \frac{(\pi+1)x_{11} - kx_{\cdot 1}}{\pi x_{1\cdot}} \text{ and } \phi_0\phi_1 = v = \hat{c}^2\pi^{-1}.$$

The two roots ϕ_0 and ϕ_1 become real if and only if $u^2 - 4v \geq 0$. In order for both roots to be strictly positive, further restriction $\phi_0 + \phi_1 > 0$, equivalently $u < 0$, is needed, since $\phi_0\phi_1 = v > 0$ as $x_{11} > 0$. Thus, the inequality condition $u^2 - 4v \geq 0$ is equivalent to $(-u) \geq 2\sqrt{v}$ i.e., $u \leq -2\sqrt{v}$. Now, the condition $u \leq -2\sqrt{v}$ means

$$kx_{\cdot 1} - (\pi+1)x_{11} \leq -2x_{11}\sqrt{\pi} \Leftrightarrow k \leq \frac{x_{11}}{x_{\cdot 1}}\left(1-\sqrt{\pi}\right)^2 \tag{S2}$$

since $\hat{c} = (x_{11}/x_{1\cdot})$.

Hence, from the above, it is clear that the two roots ϕ_0 and ϕ_1 are functions of the only unknown k and they satisfy $\phi_0 \le \phi \le \phi_1$ in order to maintain the inequality (3). In particular, equality in the above equation (S2) holds only when $u^2 = 4v$, or equivalently, $(-u)/2 = \sqrt{v}$ or equivalently, $\frac{(\phi_0+\phi_1)}{2} = \sqrt{\phi_0\phi_1}$ holds and this A.M. = G.M. condition holds if and only if $\phi_0 = \phi_1 = \phi = \sqrt{v} = \hat{c}\pi^{-1/2}$. In addition, structurally $k \ge 0$ holds under the assumption of $p \ge \pi$. Hence, note that the upper bound (S2) on the non-negative term k is the *necessary and sufficient* condition for both the roots ϕ_0 and ϕ_1 to be positive real-valued.

The two real positive roots, ϕ_0 and ϕ_1, of the Equation (S1) are

$$\phi_0 = \frac{1}{2}\left(-u - \sqrt{u^2 - 4v}\right) \text{ and } \phi_1 = \frac{1}{2}\left(-u + \sqrt{u^2 - 4v}\right),$$

with the restrictions $u < 0$ and $u^2 - 4v \ge 0$, where u and v are stated in (S1). From the above expressions of ϕ_0 and ϕ_1, we have the values of ϕ_0 and ϕ_1 corresponding to the lower bound of k (i.e., $k = 0$) are $\hat{c}$ and $\hat{c}\pi^{-1}$ respectively. Similarly, when k attains its upper bound (S2), both of ϕ_0 and ϕ_1 are equal to $\hat{c}\pi^{-1/2}$. This implies

$$\hat{c} \le \phi_0 \le \hat{c}\pi^{-1/2} \text{ and } \hat{c}\pi^{-1/2} \le \phi_1 \le \hat{c}\pi^{-1}. \tag{S3}$$

Now, let us examine the nature of the roots in terms of the unknown k.

$$\frac{d\phi_0}{dk} = \frac{d\phi_0}{du}\cdot\frac{du}{dk} = -\frac{x_{.1}}{2\pi x_{1.}}\left(1 + \frac{u}{\sqrt{u^2 - 4v}}\right).$$

Given the stated restrictions $u < 0$ and $u^2 - 4v \ge 0$,

$$u^2 - 4v < u^2 \Rightarrow \sqrt{u^2 - 4v} < -u \Rightarrow \left(1 + \frac{u}{\sqrt{u^2 - 4v}}\right) < 0 \Rightarrow \frac{d\phi_0}{dk} > 0,$$

since, $x_{.1}, x_{1.}, \pi > 0$. Thus, the smaller root, ϕ_0, is monotonically increasing in k. As the product of the two roots is a constant, that is, independent of k, the other root, ϕ_1, is monotonically decreasing. As we already found, the values of ϕ_0 and ϕ_1 corresponding to the lower bound of k are $\hat{c}$ and $\hat{c}\pi^{-1}$ respectively. Therefore, ϕ_0 (ϕ_1) increases (decreases) to $\hat{c}\pi^{-1/2}$ as k increases to its upper bound in (S2) and this finding meets the result (S3). This completes the justifications.

7. References

Bell, W.R. 1993. "Using information from demographic analysis in post-enumeration survey estimation." *Journal of the American Statistical Association* 88: 1106–1118. DOI: http://dx.doi.org/10.2307/2290805.

Bohning, D., P.V.D. Heijden, and J. Bunge. 2017. Capture-Recapture Methods for the Social and Medical Sciences (1st edition). Boca Raton, FL: Chapman Hall CRC Interdisciplinary Statistics.

Brittain, S. and D. Bohning. 2009. "Estimators in capture-recapture studies with two sources." *Advanced Statistical Analysis* 93: 23–47. DOI: https://doi.org/10.1007/s10182-008-0085-y.

ChandraSekar, C. and W.E. Deming. 1949. “On a method of estimating birth and death rates and the extent of registration.” *Journal of the American Statistical Association* 44: 101–115. DOI: https://doi.org/10.1080/01621459.1949.10483294.

Chao, A., W. Chu, and H.H. Chiu. 2000. “Capture-recapture when time and behavioral response affect capture probabilities.” *Biometrics* 56: 427–433. DOI: https://doi.org/10.1111/j.0006-341X.2000.00427.x.

Chao, A., P.K. Tsay, S.H. Lin, W.Y. Shau, and D.Y. Chao. 2001. “Tutorial in biostatistics: The application of capture-recapture models to epidemiological data.” *Statistics in Medicine* 20: 3123–3157. DOI: https://doi.org/10.1002/sim.996.

Chatterjee, K. and D. Mukherjee. 2016a. “An improved estimator of omission rate for census count: with particular reference to India.” *Communication in Statistics: Theory and Methods* 45: 1047–1162. DOI: https://doi.org/10.1080/03610926.2013.854911.

Chatterjee, K. and D. Mukherjee. 2016b. “An improved integrated likelihood population size estimation in dual-record system.” *Statistics & Probability Letters* 110: 146–154. DOI: https://doi.org/10.1016/j.spl.2015.12.019.

Chatterjee, K. and D. Mukherjee. 2016c. “On the estimation of homogeneous population size from a complex dual-record system.” *Journal of Statistical Computation and Simulation* 86: 3562–3581. DOI: https://doi.org/10.1080/00949655.2016.1173695.

Chatterjee, K. and D. Mukherjee. 2018. “A new integrated likelihood for estimating population size in dependent dual-record system.” *Canadian Journal of Statistics* 46: 577–592. DOI: https://doi.org/10.1002/cjs.11477.

Eckberg, D.L. 2000. “A capture-recapture approach to the estimation of hidden historical killings.” Proceedings of the 1999 meeting of the 332 Homicide Research Working Group, Washington, DC: Federal Bureau of Investigation, this volume was edited by P.H. Blackman, V.L. Leggett, B.L. Olson, and J.P. Jarvis.

El-Khorazaty, M.N. 2000. “Dependent dual-record system estimation of number of events: a capture-mark-recapture strategy.” *Environmetrics* 11: 435–448. DOI: https://doi.org/10.1002/1099-095X(200007/08)11:4<435::AID-ENV427>3.0.CO;2-2.

Gerritse, S.C., B.F.M. Bakker, D. Zult, and P.G.M. Van-der-Heijden. 2017. *The impact of linkage errors and erroneous captures on the population size estimator due to implied coverage 2017–16*. Vol. 16. Statistics Netherlands. Available at: https://www.cbs.nl/en-gb/background/2017/39/impact-of-linkage-errors-and-erroneous-captures (accessed February 2020).

Gosky, R. and S.K. Ghosh. 2011. “A comparative study of Bayes estimators of closed population size from capture-recapture data.” *Journal of Statistical Theory and Practice* 5: 241–260. DOI: https://doi.org/10.1080/15598608.2011.10412027.

Granerod, J., S. Cousens, N.W.S. Davies, N.S. Crowcroft, and S.L. Thomas. 2013. “New estimates of incidence of encephalitis in England.” *Emerging Infectious Diseases* 19: 1455–1462. DOI: https://doi.org/10.3201/eid1909.130064.

Greenfield, C.C. 1975. “On the estimation of a missing cell in a 2 × 2 contingency table.” *Journal of the Royal Statistical Society* A 138: 51–61. DOI: https://doi.org/10.2307/2345249.

Griffin, R.A. 2014. “Potential uses of administrative records for triple system modeling for estimation of census coverage error in 2020.” *Journal of Offcial Statistics* 30: 177–189. DOI: https://doi.org/10.2478/jos-2014-0012.

Iñigo, J., A. Arce, J.M. Martn-Moreno, R. Herruzo, E. Palenque, and F. Chaves. 2003. “Recent transmission of tuberculosis in Madrid: application of capture-recapture analysis to conventional and molecular epidemiology.” *International Journal of Epidemiology* 32: 763–769. DOI: https://doi.org/10.1093/ije/dyg098.

Jarvis, S.N., P.J. Lowe, A. Avery, S. Levene, and R.M. Cormack. 2000. “Children are not goldfish-mark/recapture techniques and their application to injury data.” *Injury Prevention* 6: 46–50. DOI: http://dx.doi.org/10.1136/ip.6.1.46.

Lee, S.M. and C.W.S. Chen. 1998. “Bayesian inference of population size for behavioral response models.” *Statistica Sinica* 8: 1233–1247. Available at: http://www3.stat.sinica.edu.tw/statistica/j8n4/j8n414/j8n414.htm (accessed February 2020).

Lee, S.M., W. Hwang, and L. Huang. 2003. “Bayes estimation of population size from capture-recapture models with time variation and behavior response.” *Statistica Sinica* 13: 477–494. Available at: http://www3.stat.sinica.edu.tw/statistica/j13n2/j13n213/j13n213.html (accessed February 2020).

Nour, E.-S. 1982. “On the estimation of the total number of vital events with data from dual collection systems.” *Journal of the Royal Statistical Society*, Series A 145: 106–116. DOI: http://dx.doi.org/10.2307/2981424.

O’Connell, M. and K.H. Pollock. 1992. Wildlife 2001: Populations. Dordrecht: Springer.

Otis, D.L., K.P. Burnham, G.C. White, and D.R. Anderson. 1978. “Statistical inference from capture data on closed animal populations.” *Wildlife Monographs: A Publication of Wildlife Society* 62: 3–135. Available at: https://pubs.er.usgs.gov/publication/70119899 (accessed February 2020).

Ruiz, M.S., A. O’Rourke, and S.T. Allen. 2016. “Using capture-recapture methods to estimate the population of people who inject drugs in Washington DC.” *AIDS and Behavior* 20: 363–368. DOI: http://dx.doi.org/10.1007/s10461-015-1085-z.

SOSU, 2014. *Report on the project survey of looms and work sheds in comprehensive handloom development programme in Dakshin Dinajpur District*. Sampling and Offcial Statistics Unit, Indian Statistical Institute, Commissioned by: Directorate of Textiles, Government of West Bengal.

Wolter, K.M. 1986. “Some coverage error models for census data.” *Journal of the American Statistical Association* 81: 338–346. DOI: http://dx.doi.org/10.2307/2289222.

Received November 2017
Revised August 2019
Accepted November 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 49–62, http://dx.doi.org/10.2478/JOS-2020-0003

The Joinpoint-Jump and Joinpoint-Comparability Ratio Model for Trend Analysis with Applications to Coding Changes in Health Statistics

Huann-Sheng Chen[1], *Sarah Zeichner*[2], *Robert N. Anderson*[3], *David K. Espey*[4], *Hyune-Ju Kim*[5], *and Eric J. Feuer*[1]

Analysis of trends in health data collected over time can be affected by instantaneous changes in coding that cause sudden increases/decreases, or "jumps," in data. Despite these sudden changes, the underlying continuous trends can present valuable information related to the changing risk profile of the population, the introduction of screening, new diagnostic technologies, or other causes. The joinpoint model is a well-established methodology for modeling trends over time using connected linear segments, usually on a logarithmic scale. Joinpoint models that ignore data jumps due to coding changes may produce biased estimates of trends. In this article, we introduce methods to incorporate a sudden discontinuous jump in an otherwise continuous joinpoint model. The size of the jump is either estimated directly (the Joinpoint-Jump model) or estimated using supplementary data (the Joinpoint-Comparability Ratio model). Examples using ICD-9/ICD-10 cause of death coding changes, and coding changes in the staging of cancer illustrate the use of these models.

Key words: Joinpoint model; international classification of diseases (ICD); trend analysis; coding change; cancer staging system; comparability ratio.

1. Introduction

"Is the trend changing?" This question underlies trend analysis in the field of disease prevention, control and surveillance. Data describing disease incidence, mortality and other health series are reported over time. Data items in those series are often recorded or classified based on certain types of coding systems, and sudden changes in code structure or coding rules over time are not uncommon. For example, in 1999, the Tenth Revision of the International Classification of Diseases (ICD-10) replaced the Ninth Revision (ICD-9) for coding causes of death used for mortality statistics (Anderson et al. 2001). In another

[1] Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, MD, U.S.A. Email: Huann-Sheng.Chen@nih.gov and feuerr@mail.nih.gov
[2] Division of Geological and Planetary Sciences, California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, U.S.A. Email: szeichner@caltech.edu
[3] Division of Vital Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, U.S.A. Email: rca7@cdc.gov
[4] Division of Population Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Albuquerque, NM, U.S.A. Email: dke0@cdc.gov
[5] Department of Mathematics, Syracuse University, Syracuse, NY, U.S.A. Email: hjkim@syr.edu

Acknowledgments: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the National Institutes of Health.

example, cancer staging algorithms change across time periods (Amin 2017). Newer staging system are added by cancer registries to keep definitions consistent with the current understanding of diseases. Such coding changes may cause discontinuous increases/decreases, or "jumps," in the data series, even though it may not affect the underlying trend.

In the past, the Joinpoint model, developed by the National Cancer Institute, has been widely used to characterize and report trends in health statistics and other time data series (Kim et al. 2000; Clegg et al. 2009). The model has log linear segments and has distinct advantages over other models, such as the polynomial fitted model, and is easier to interpret (Clegg et al. 2009). However, when there is a discontinuity or jump that occurs in the data, the Joinpoint model could potentially result in a misleading interpretation of trend changes, as the jump could be the result of a one-time external circumstance, which may not be interpreted as a changing point in the data.

In some cases the size of the jump is known or can be estimated. In the case of ICD code changes, a comparison of causes of death coded to both ICD-9 and ICD-10 was conducted. The National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC) double coded 1996 death certificates using both the ICD-9 and ICD-10 algorithms and published "comparability ratios" and their standard errors (National Center for Health Statistics 2009). The comparability ratios estimate the size of the jump in mortality rates associated with the coding change. In this case, we propose a Joinpoint-Comparability Ratio (JP-CR) method to accommodate the jumps, that is, converting the data before the jump by the size of the jump and then applying the Joinpoint model to the converted data. The standard error of the converted data can be adjusted by combining the standard error of the CR and the standard error associated with the original data. This approach may work well if the goal is to capture the trend of the time series data given that the jump size is known or has been estimated based on an external study.

However, the requirement of knowing the jump size makes the analysis more difficult because the size of the jump needs to be estimated, and in many cases, this may not be feasible. Motivated by the need for trend analysis with a sudden jump where the size of the jump is unknown, we propose another method called Joinpoint-Jump (JP-Jump) model. The model minimizes the effect of the jumps on trend analysis. Unlike the JP-CR method, the JP-Jump model simultaneously estimates the size of the jump, as well as the changes in trend.

The remainder of this article is organized as follows. Section 2 briefly reviews the Joinpoint model and presents the JP-CR and JP-Jump models. In Section 3, we consider several applications. In one application, the proposed models are applied to US mortality data where there is a coding change from ICD-9 to ICD-10 starting in 1999. In another application, we apply the proposed models to cancer incidence data to test the difference between two cancer staging systems. Section 4 discusses practical considerations in choosing between the two models.

2. The Joinpoint, Joinpoint-Jump (JP-Jump) and Joinpoint-Comparability Ratio (JP-CR) Models

2.1. *The Joinpoint Model (JP) – A Brief Review*

The Joinpoint model is a segmented linear regression model. Suppose that we observe (x_1,γ_1), . . . ,(x_n, γ_n), γ_i is an age-adjusted cancer incidence/mortality rate at time x_i and

$y_i = \log(\gamma_i)$. The Joinpoint model, propsed by Kim et al. (2000), assumes

$$y_i = log(\gamma_i) = \alpha_1 + \beta_1 x_i + \delta_1(x_i - \tau_1)^+ + \ldots + \delta_\kappa(x_i - \tau_\kappa)^+ + \epsilon_i,$$
$$i = 1, \ldots, n, \quad (1)$$

where ϵ_i are independent errors, the notation $a^+ = a$ if $a > 0$, and $a^+ = 0$ otherwise. In this model, the mean function of y_i is linear segments connected at change-points $\tau_1 < \cdots < \tau_\kappa$. The locations of change-points τ_i as well as the number of change-points κ are assumed to be unknown and need to be estimated from the data. Assuming the number of change-points is given, say $\kappa = K$, the locations of K joinpoints, $\tau_1, \ldots, \tau_K$, are estimated by the grid search method described by Lerman (1980) or a continuous fitting method proposed by Hudson (1966). The overall least squares estimates of the regression coefficients are then obtained based on the estimated joinpoints. Once the least squares fit is obtained for a model with $\kappa = K$, an iterative procedure is used to determine whether addition of joinpoints significantly reduces the residual sum of squares. The procedure iteratively tests the null hypothesis that there are K_0 joinpoints against the alternative hypothesis that there are K_1 joinpoints where $K_1 > K_0$, and usually begins with K_0 and K_1 respectively being the pre-specified minimum and maximum number of joinpoints allowed. Due to the fact that classical asymptotic theory does not work in this situation, a Monte Carlo permutation test is used to determine the p-value of the test. First permute the residuals from fitting the model under the null hypothesis. For each permutation, add the permutated residuals back to the fitted values and refit the permutated data under the alternative hypothesis, and obtain the F-statistics as a goodness-of-fit measure. The p-value of the test is then calculated from the distribution of the goodness-of-fit statistics. If the null hypothesis is rejected, then test $H_0 : \kappa = K_0 + 1$ versus $H_1 : \kappa = K_1$; otherwise, test $H_0 : \kappa = K_0$ versus $H_1 : \kappa = K_1 - 1$. Tests are repeatedly conducted until for some K, the testing of $H_0 : \kappa = K$ versus $H_1 : \kappa = K + 1$ is performed. Because the procedure is based on multiple testing, the significance level of each test is adjusted to maintain the overall level under α, which is the probability of over-fitting the model. Instead of permutation test, model selection methods based on the Bayes Information Criterion (BIC) or a modified BIC, can work as faster alternatives.

2.2. *The JP-Jump Model*

We consider a model where a jump occurs at a known location s, and we allow s to be a possible change-point. Denote the observed rate at the time x_i as γ_i, $i = 1, \ldots, n$. The jump location is incorporated into a joinpoint model by assuming

$$y_i = log(\gamma_i) = \alpha_1 + \beta_1 x_i + \delta_1(x_i - \tau_1)^+ + \cdots + \delta_\kappa(x_i - \tau_\kappa)^+ + \lambda\, \mathrm{I}(x_i \geq s) + \epsilon_i, \quad (2)$$

where $\tau_1 < \cdots < \tau_\kappa$ are unknown change-points, s is a known location of a jump, λ represents the jump size, ϵ_i are independent errors $N(0, \sigma^2)$, and $\mathrm{I}(A) = 1$ if A is true and 0 otherwise. If j denotes the index value such that $\tau_j < s < \tau_{j+1}$, this model can also be expressed as

$$E(y_i|x_i) = \begin{cases} \alpha_1 + \beta_1 x_i & \textit{if } \tau_0 \leq x_i \leq \tau_1 \\ \vdots & \vdots \\ \alpha_j + \beta_j x_i & \textit{if } \tau_{j-1} < x_i \leq \tau_j \\ \alpha_{j+1} + \beta_{j+1} x_i & \textit{if } \tau_j < x_i < s \\ \alpha_{j+1} + \lambda + \beta_{j+1} x_i & \textit{if } s \leq x_i \leq \tau_{j+1} \\ \alpha_{j+2} + \lambda + \beta_{j+2} x_i & \textit{if } \tau_{j+1} < x_i \leq \tau_{j+2} \\ \vdots & \vdots \\ \alpha_{\kappa+1} + \lambda + \beta_{\kappa+1} x_i & \textit{if } \tau_\kappa < x_i \leq \tau_{\kappa+1} \end{cases}, \tag{3}$$

where $\tau_0 = \min\{x_i\}$, $\tau_{\kappa+1} = \max\{x_i\}$, and $\alpha_m + \beta_m \tau_m = \alpha_{m+1} + \beta_{m+1}\tau_m$ for $m = 1, \ldots, \kappa$. Equations (2) and (3) are equivalent when

$$\alpha_l = \alpha_1 - \sum_{u=1}^{l-1} \delta_u \tau_u, l = 2, \ldots, \kappa + 1,$$

$$\beta_l = \beta_1 + \sum_{u=1}^{l-1} \delta_u, \quad l = 2, \ldots, \kappa + 1.$$

To fit the model, at any possible locations of joinpoints $(\tau_1, \ldots, \tau_\kappa) = (t_1, \ldots, t_\kappa) = t'$, we obtain the least squares estimate of $\boldsymbol{\theta} = (\alpha_1, \beta_1, \delta_1, \ldots, \delta_\kappa, \lambda)'$ as

$$\hat{\boldsymbol{\theta}}(t) = (X_t^{'} X_t)^{-1} X_t^{'} y,$$

where $y_1 = (y_1, \ldots, y_n)'$, and

$$X_t = \begin{pmatrix} 1 & x_1 & (x_1 - t_1)^+ & \cdots & (x_1 - t_\kappa)^+ & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_j & (x_j - t_1)^+ & \cdots & (x_j - t_\kappa)^+ & 0 \\ 1 & x_{j+1} & (x_{j+1} - t_1)^+ & \cdots & (x_{j+1} - t_\kappa)^+ & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - t_1)^+ & \cdots & (x_n - t_\kappa)^+ & 1 \end{pmatrix}.$$

The number of change-points κ can be determined by a Monte Carlo permutation test and the locations of change-points τ_i are obtained by a grid search method. A grid search finds the estimate of $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_\kappa)'$, denoted by $\hat{\tau}$, by minimizing the residual sum of squares RSS(t) over all possible choices of t. Given $\hat{\tau}$, the overall least squares estimate is calculated, denoted by $\hat{\boldsymbol{\theta}}(\hat{\tau})$. After fitting the model and calculating the residual sum of squares for the fitted model, the number of change-points can be either selected by using a permutation procedure or BIC method.

Instead of calculating the constrained standard error of $\hat{\boldsymbol{\theta}}$, the standard error is calculated by unconstrained standard error estimate. For all the segments except the one that includes the known change-point s, the standard error is calculated as in the regular Joinpoint model. For the segment where s lies, say $[x_{j1}, \ldots, x_{jm}]$, $\mathrm{E}(y|x) = \alpha_{j+1} + \beta_{j+1}x + \lambda I(x \geq s)$ for $x \in [x_{j1}, \ldots, x_{jm}]$, the standard error estimates of $(\hat{\alpha}_{j+1}, \hat{\beta}_{j+1}, \hat{\lambda})$, can be calculated by using

$$[SE(\hat{\alpha}_{j+1}, \hat{\beta}_{j+1}, \hat{\lambda})]^2 = \hat{\sigma}^2 \left(X'_{[x_{j1},\, x_{jm}]} X_{[x_{j1},\, x_{jm}]} \right)^{-1},$$

where

$$X_{[x_{j1},\, x_{jm}]} = \begin{pmatrix} 1 & x_{j1} & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_j & 0 \\ 1 & x_{j+1} & 1 \\ \vdots & \vdots & \vdots \\ 1 & x_{jm} & 1 \end{pmatrix}.$$

Like the regular Joinpoint model, this unconstrained standard error estimate tends to perform better than the constrained standard error estimate, see Kim et al. (2008).

In the JP-Jump model, λ, representing the size of the jump, is estimated simultaneously with other parameters $\alpha_1, \ldots, \alpha_{\kappa+1}$ and $\beta_1, \ldots, \beta_{\kappa+1}$. The variance of $\hat{\lambda}$ can be obtained as well. To compare the rates after and before the jump, we consider the ratio of the after-jump rate and the before-jump rate. We call it the model-based comparability ratio, which can be estimated directly from the JP-Jump model. Specifically, let ρ denote the model-based comparability ratio, calculated as

$$\rho = \frac{\exp(E(y|x = s^+))}{\exp(E(y|x = s^-))} = \frac{\exp(\alpha_{j+1} + \beta_{j+1}x + \lambda)}{\exp(\alpha_{j+1} + \beta_{j+1}x)} = e^{\lambda}, \tag{4}$$

where $x = s^+$ and $x = s^-$ denote the right and left side limits of x at s. The model-based comparability ratio ρ is estimated by $\hat{\rho} = \exp(\hat{\lambda})$, and the standard error of $\hat{\rho}$ can be estimated by using the delta method, which is

$$Var(\hat{\rho}) \approx \left(e^{\hat{\lambda}}\right)^2 Var(\hat{\lambda}),$$

and $SE(\hat{\rho}) = e^{\hat{\lambda}} SE(\hat{\lambda})$. For any $\alpha \in (0,1)$, the $(1 - \alpha) \times 100\%$ approximate confidence interval for ρ can be constructed by $[\hat{\rho} - z_{\alpha/2}SE(\hat{\rho}), \hat{\rho} + z_{\alpha/2}SE(\hat{\rho})]$, where $z_{\alpha/2}$ denotes the upper $\alpha/2$-th percentile of the standard normal distribution.

2.3. The Joinpoint-Comparability Ratio Model (JP-CR)

For this method, we assume that the analyst is provided with a comparability ratio and its variance from an external source. The ratio, denoted by C, represents a ratio of the case counts after the change to before the change. If C is 1, the total net count is not affected by the code change, even if individual records may be impacted. If C is less than 1, fewer events are classified to this cause under the new coding system compared to the previous system.

Conversely, if the ratio is greater than 1, more events are classified to this cause under the new coding system than under the previous one.

Recall that the incidence /mortality rate is denoted by γ. The comparability ratio method specifically uses the following steps:

- Step 1. For data before the coding change, find the Comparability ratio-Modified rate, γ^{CM}, by multiplying the original rate by the comparability ratio, that is,

$$\gamma^{CM} = \gamma \times C.$$

- Step 2. Combine the adjusted data before the coding change and the original data after the coding change. Analyze the trend by applying the regular Joinpoint model to the combined data.
- Step 3. Before graphing the results, convert the fitted rates prior to the coding change derived from Step 2 (denoted by $\hat{\gamma}^{CM}$), back to the original scale by dividing by the comparability ratio, $\hat{\gamma} = \frac{\hat{\gamma}^{CM}}{C}$.

3. Results

3.1. US Mortality With Code Changes From ICD-9 To ICD-10

The International Classification of Diseases is revised periodically to stay current of medical knowledge and advances (Anderson et al. 2001). The most recent code changes occurred between 1998 and 1999 when ICD-10 replaced ICD-9. The coding changes have impacts on some of the major causes of death in the United States. Comparability ratios developed by Anderson et al. (2001) are designed to quantify the impact of the coding change. Although the first year with new code is 1999, the comparability ratios presented by Anderson et al. are based on double-coding the same deaths occurring in 1996 by both ICD-9 and ICD-10. For each cause of death, the comparability ratio, denoted by C, is calculated as

$$C = \frac{D_{ICD-10}}{D_{ICD-9}},$$

where $D_{ICD\text{-}10}$ and $D_{ICD\text{-}9}$ are the number of deaths classified by ICD-10 and ICD-9 respectively. We demonstrate these coding changes using two examples, one with a large comparability ratio (septicemia) and one with relatively small comparability ratio (melanoma).

3.1.1. Septicemia Mortality With Code Changes From ICD-9 To ICD-10

The published comparability ratio for septicemia is 1.1949 (Anderson et al. 2001), which represents nearly 20% more deaths according to the ICD-10 coding paradigm than that of ICD-9. Under ICD-10, septicemia is selected as the underlying cause of death over pneumonia when both are listed on the death certificate; septic shock, coded in ICD-9 in a different category is coded as “Unspecified septicemia” in ICD-10. The mortality rates due to septicemia from 1979 to 2010 for all races and both sexes along with the fitted trends based on the regular Joinpoint model (called JP-Standard model hereafter) are shown in Figure 1. The JP-Standard model does not take into account the code change information and the fitted line indicates there is a large upward trend from 1996 to 2000. The annual percent change (APC) between 1996 and 2000 is 9.95% per year, which is statistically significant at level 0.05. The discrepancy between the observed pattern and the fitted model suggests that the JP-Standard model fails to capture the observed data when there is a sudden change in data.

To account for the coding change, both JP-Jump and JP-CR models are applied to the data. The coding change is placed at 1998.5 to represent that 1998 is coded using ICD-9, and 1999 is coded using ICD-10. Figure 1 shows these two methods have similar results. The location of the joinpoints in the two models are identical, except for the end of the second segment, which ends at 1994 for the JP-Jump model and 1995 for the JP-CR model. Both the JP-Jump and the JP-CR model capture an upward trend from 1994/1995 to 2002; but unlike the JP-Standard model, the upward trend only shows a modest increase. The APC for the last segment (2002–2010 for both models), which is the most important

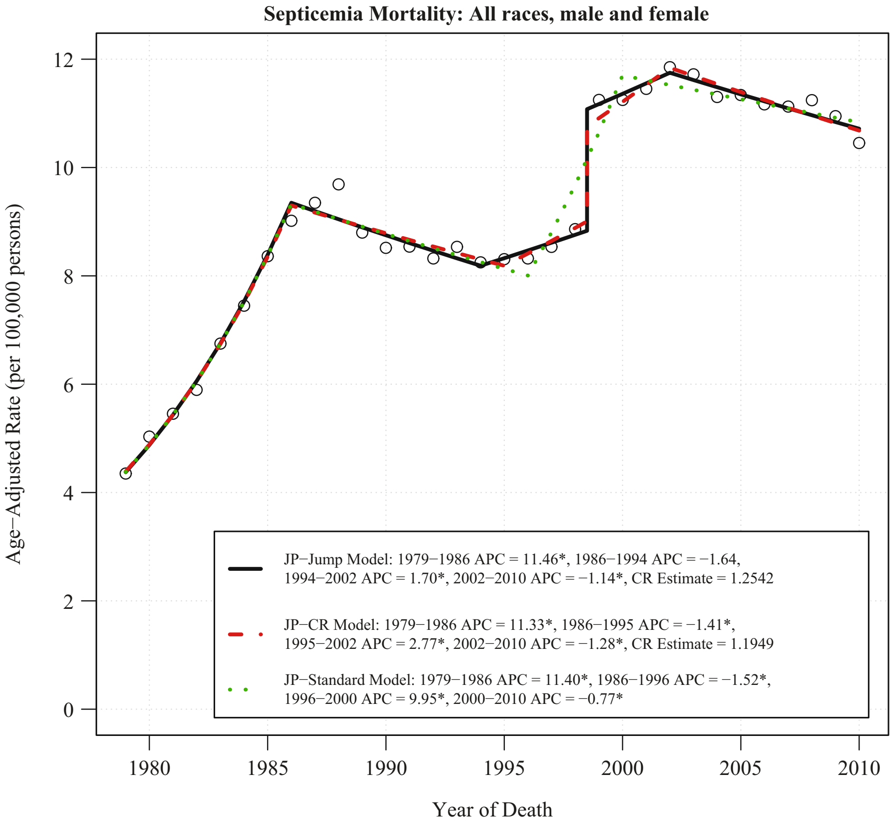


Fig. 1. JP-Jump model, JP-CR model and JP-standard model of septicemia US mortality for all races and both genders, 1979–2010.

Notes: There are three joinpoints for each model. The estimate of the comparability ratio from JP-Jump Model is 1.2542 with standard error = 0.032. The published comparability ratio is 1.1949 with standard error = 0.002. Both estimated and published comparability ratio are statistically different from one at level 0.05. The symbol * is shown if the annual percent change (APC) is significantly different from zero at level 0.05.

segment, is nearly identical: -1.14% per year for the JP-Jump model and -1.28% per year for the JP-CR model; both are statistically significant. The model-based comparability ratio $\hat{\rho}$ estimated from the JP-Jump model is 1.2542 with a standard error (SE) of 0.032, compared with the published comparability ratio C ($C = 1.1949$ with SE $= 0.0042$). Both the estimated and published comparability ratio are statistically different from one at level 0.05.

3.1.2. Melanoma Mortality With Code Changes From ICD-9 To ICD-10

While some causes of death other than cancer have large ICD-9/ICD-10 coding changes, most causes of death due to cancer tend to have relatively small comparability ratios – most have comparability ratios hovering around 0.99 and 1.01 (Anderson et al. 2001), and among the 52 cancer sites published in NCI's Cancer Statistics Review (Noone et al. 2018), only ten sites have comparability ratios that fall outside of the [0.99, 1.01] range. Until now, a common practice in cancer surveillance is to use the regular Joinpoint model (JP-Standard) to fit the data. For the majority of cancers, this common practice works fine due to the small comparability ratio. However, in a few cases it is possible that even a relatively modest comparability ratio can change the overall conclusions about the trends.

Figure 2 shows the mortality due to melanoma for all races, both males and females combined analyzed from 1992 to 2014 by the JP-Jump, JP-CR, and JP-Standard models. The published ICD-9 to ICD-10 comparability ratio for melanoma is 0.9677, with SE $= 0.0032$ and 95% CI $= (0.9614, 0.9741)$. The JP-Standard model finds no joinpoint and shows a flat trend with a non-significant APC of -0.06% per year. The JP-CR model with the relevant comparability ratio (0.9677) finds a joinpoint in 2010 with a significant rise of 0.30% per year prior to 2010 and a significant fall from 2010 to 2014 of 1.43% per year. The JP-Jump Model estimates a similar comparability ratio of 0.9444 (SE $= 0.0116$ and 95% CI $= (0.9218, 0.9671)$) and finds a joinpoint at 2009 with a significant rise of 0.50% per year prior to the joinpoint, and a significant decline of 1.19% per year after the joinpoint. Despite the small comparability ratio size, these model results are qualitatively different when the coding change from ICD-9 to ICD-10 is taken into account; it produced biased trends when the coding change is not accounted for (using the JP-Standard Model). The JP-Jump model is able to pick up approximately the same modest jump as the JP-CR model, which is estimated by using external data.

3.2. Cancer Staging: Summary Staging 2000 and SEER Historical Staging

Analysis of cancer data incidence trends requires consistent staging algorithms across time periods. This can be difficult because staging algorithms change over time to reflect changes in our understanding of disease and changes in prognosis for various subgroups who may be benefitting from new targeted therapies. The National Cancer Institute's Surveillance Epidemiology and End Results (SEER) cancer registry program started in 1973, and has a series of nine registries covering approximately 10% of the US population since 1975 (more registries have been added in later years). While there are various staging systems in use, the SEER Historical Staging System has been consistent since the program's inception. Summary Staging 2000 is another staging system, but it can only consistently code back to cases diagnosed in 1998. Both staging systems code cancers as

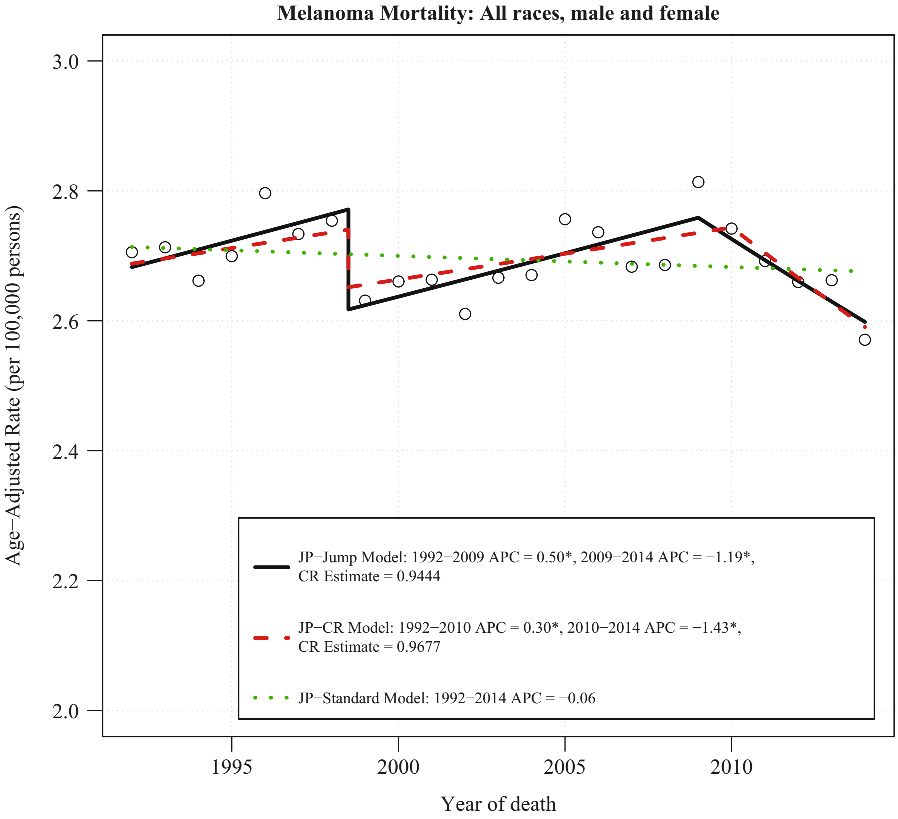


Fig. 2. JP-jump model, JP-CR model and JP-standard model of melanoma US mortality rates for all races and both genders, 1992–2014.

Notes: There is one joinpoint for JP-Jump Model, one joinpoint for JP-CR Model, and no joinpoint for JP-Standard Model. The estimate of the comparability ratio from JP-Jump Model is 0.9444 with standard error = 0.0116. The published comparability ratio is 0.9677 with standard error = 0.0032. Both estimated and published comparability ratio are statistically different from one at level 0.05. The symbol ∗ is shown if the annual percent change (APC) is significantly different from zero at level 0.05.

"local", "regional", or "distant". It is of interest to estimate a single trend using the older Historical Staging from 1975 through 1997 and switch over to the more contemporary Summary Staging 2000 in 1998 using the JP-Jump model or JP-CR model. Cases from 1998 onward have been coded using both staging systems, making it possible to compute a comparability ratio.

Figure 3 plots the age-adjusted incidence rate for all races, female, breast cancer at the "distant" stage. The Historical Staging series are plotted from 1975 to 2014 (the most recent year currently available), while the Summary Staging series are plotted from 1998 to 2014. The large gap between these staging systems is due to inflammatory carcinoma being changed from "distant" to "regional" disease, as the improving prognosis of this cancer subtype. Despite this change in coding, incidence rate trends from the two coding systems appear approximately parallel. In order to apply the JP-CR model, we first find the ratios between the Summary Staging 2000 rates and the SEER Historical Staging rates from 1998 to 2014. The mean of these ratios are CR = 0.8059 and the variance of

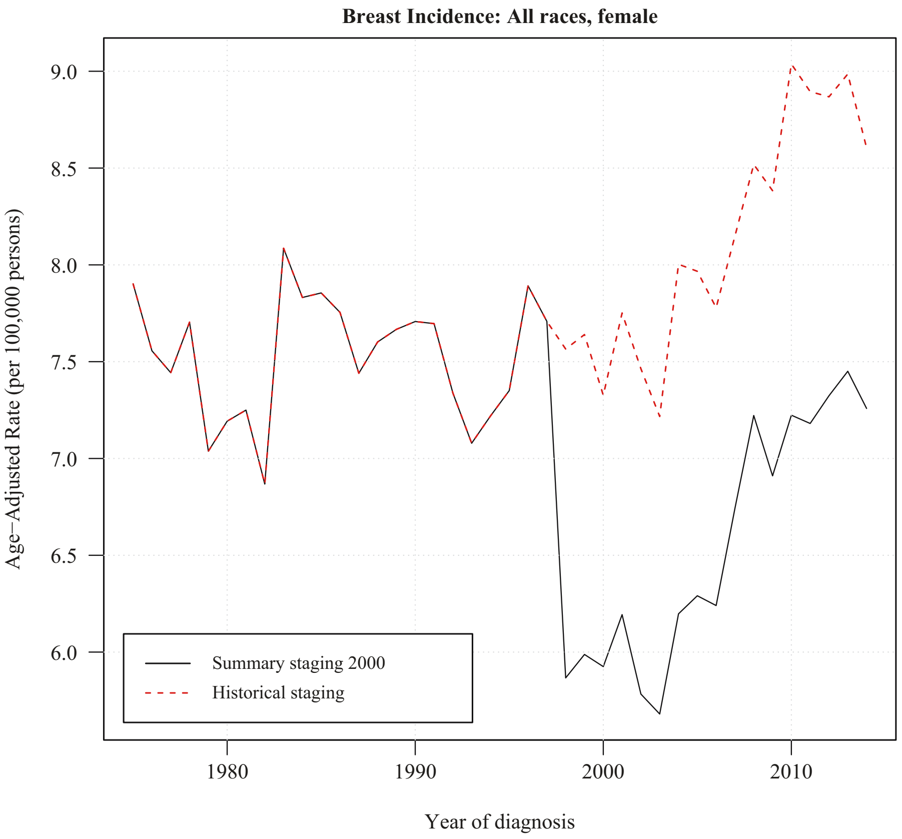


Fig. 3. Data of breast cancer incidence rates at distant stage for all races female. The historical staging series has data from 1975 to 2014. Between 1998–2014, the summary staging 2000 series is plotted against the historical staging.

CR = 0.00056, which represents about 20% reduction of incidence rates in the "distant" stage coded by the Summary Staging 2000 when compared to the Historical Staging. Figure 4 consists of data from Historical Staging before 1998 and Summary Staging 2000 from 1998 forward. Both JP-Jump and JP-CR models are fitted to the data. The estimated comparability ratio from the JP-Jump model is 0.7871, which is close to the pre-determined CR = 0.8059, and significantly different from one. The fitting illustrates a continuing flat trend (APC close to 0, not statistically significant) between 1975 to 2002, a drop in 1998 in the incidence rate due to difference between the Historical Staging and Summary Staging 2000 from 1998, and an increasing trend starting from 2002.

For comparison, the fitted plot from JP-Standard model shows a downward trend with APC = − 8.89% per year between 1996 and 1999, which covers the time point of coding system change in 1998. Both the JP-Jump model and the JP-CR model capture the real trends. The Summary Staging 2000 has APC = 2.05% per year for the JP-CR model and APC = 2.09% per year for the JP-Jump model in 2002–2014 – both statistically significant. A possible reason behind the upward trend of incidence rates

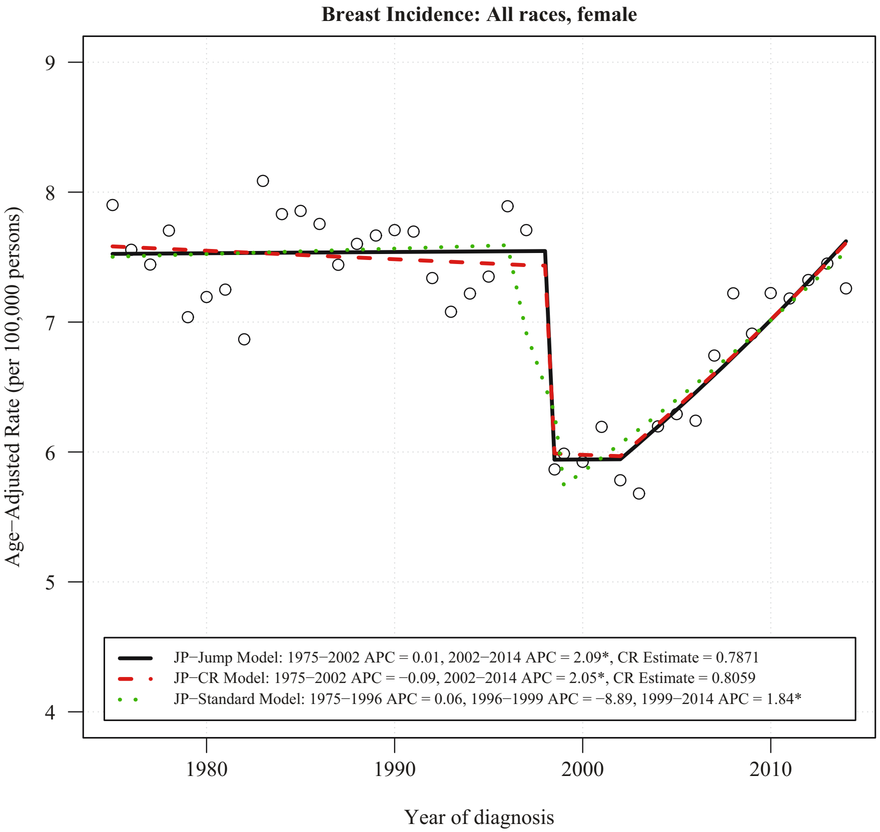


Fig. 4. JP-jump model, JP-CR model and JP-standard model of breast incidence rates for all races female at distant stage using historical staging between 1975–1997 and summary staging 2000 between 1998–2014. Notes: There is one joinpoint for JP-Jump Model, one joinpoint for JP-CR Model, and two joinpoints for JP-Standard Model. The estimate of the comparability ratio from JP-Jump Model is 0.7871 with standard error = 0.0279, while the comparability ratio calculated from original data is 0.8059, with standard error = 0.0237. Both estimated and calculated from original data comparability ratio are statistically different from one at level 0.05. The symbol $*$ is shown if the annual percent change (APC) is significantly different from zero at level 0.05.

starting from 2002 could be due, at least in part, to improvements in diagnostic technology, such as the broad use of pet scans since 2002. These more sensitive scans can find sites of distant disease that would have not been apparent in earlier technologies. The finding is consistent with the report of a statistically significant increase in the incidence of breast cancer with distant staging for women aged 25–39 years (Johnson et al. 2013).

4. Discussion

In this article, we propose two approaches – the JP-Jump and JP-CR models – to model trends in data when there is a coding change. There are several practical issues to consider when selecting the most suitable model.

To analyze trend data when a coding change occurs, assuming a comparability ratio based on double coding using the new and old codes is available, any one of the three models (JP-Jump, JP-CR, or JP-Standard model) may be applicable. In many cases, one may find that all three models produce virtually identical results, especially when the CR from the double coding is small (e.g., between 0.99 and 1.01), denoting little effect of the coding change. If no CR is available, then only the JP-Jump model or the JP-Standard model are appropriate. Which method to use in practice often depends on several factors, such as size of the data cohort, and size and variability of the CR ratio.

In comparing the three methods, the estimates of trends from the JP-Standard model can be biased when coding changes are not taken into account, even if the jumps are modest. The first two methods can yield similar results, but the JP-Jump model has advantages over the JP-CR model because there is no requirement to know the size of the jump prior to applying the model. Even if double coding is available and the ratio between the two codes can be estimated, the JP-CR model may not be suitable for practical use in cases where the statistical variability of the CR is high or the accuracy or applicability of the estimate is questionable. Sampling error or small numbers may result in unstable CRs. Also, demographic and geographic variation in the CR may make the applicability of a CR for all races, or for all states, problematic when analyzing data by demographic groups or states. In contrast, the JP-Jump model has advantages because it is always estimated directly using the data series of interest, that is, it can adapt data to estimate the size of jump and the trend simultaneously, for each subgroup or for the whole group.

However, like the regular Joinpoint model, the JP-Jump model is sensitive to the data and the location of the joinpoints, and results should be interpreted with caution. Estimates of jump sizes, particularly when close to a joinpoint, might be confounded with the slopes before and after the joinpoint itself. Furthermore, the underlying data variability may make estimation of a small or modest jump size impossible. For small subpopulations (e.g., Asian/Pacific Islander, American Indian/Alaska Native, rare cancer sites, or small geographic areas), such situations may occur. In the JP-Jump model, one can check that the jump is statistically different from zero. We note that in cases where the jump size is not statistically significant, the JP-CR model may be the better choice, and that even with statistical significance one should be wary of the data variability as compared to data set size when assessing results of the JP-Jump model. Overall, we suggest applying both models to data sets to consider the best fit for particular scenarios. In many cases where the estimates are similar, the JP-Jump model may be preferred due to its specificity per data set.

The JP-Jump model has a similar form to the regular Joinpoint model. The model fitting and inference on the parameters are derived by modification from those of the regular Joinpoint model. Note that the regular Joinpoint model has a continuous mean function, while the mean function of the JP-Jump model is continuous only at places other than the jump location. Additionally, we assume in the JP-Jump model that the location of the jump is known, and only one such jump is allowed. A direct extension to the JP-Jump model is to allow more than one known jump in response to multiple coding changes over time. The implementation of such a model is straightforward, although with more than one coding change one has to be especially careful about the issue of confounding between the APC of segments and the size of the jump.

In the JP-Jump model (Equation (2)), the errors ϵ_i are assumed to be independent $N(0, \sigma^2)$. Just like the regular Joinpoint model, these assumptions can be relaxed to allow correlated errors and non-constant variance. The model fitting and inference procedures can be extended from the regular Joinpoint model as well.

In summary, trend data with a sudden jump occurs in many situations, particularly in health data, where coding systems and practices change over time. Analyses using the regular Joinpoint model that ignore these jumps could produce misleading conclusions, as shown in the melanoma mortality example. The proposed methods in this article incorporate a sudden discontinuous jump in an otherwise continuous joinpoint model. In particular, the JP-Jump model can estimate the magnitude of the jump, while at the same time can find the change-points of the continuous trend. The methods provide a useful tool for trend analysis.

5. References

Amin, M.B. 2017. *AJCC cancer staging manual*. Eight edition / editor-in-chief, Mahul B. Amin, MD, FCAP; editors, Stephen B. Edge, MD, FACS and 16 others; Donna M. Gress, RHIT, CTR – Technical editor; Laura R. Meyer, CAPM – Managing editor. ed. Chicago IL: American Joint Committee on Cancer, Springer.

Anderson, R.N., A.M. Minino, D.L. Hoyert, and H.M. Rosenberg. 2001. "Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates." *Natl Vital Stat Rep no.* 49(2): 1–32.

Clegg, L.X., B.F. Hankey, R. Tiwari, E.J. Feuer, and B.K. Edwards. 2009. "Estimating average annual per cent change in trend analysis." *Stat Med no.* 28(29): 3670–3682. DOI: https://doi.org/10.1002/Sim.3733.

Hudson, D. 1966. "Fitting segmented curves whose join points have to be estimated." *Journal of the American Statistical Association no.* 61: 1097–1129. DOI: https://doi.org/10.1080/01621459.1966.10482198.

Johnson, R.H., F.L. Chien, and A. Bleyer. 2013. "Incidence of Breast Cancer With Distant Involvement Among Women in the United States, 1976 to 2009." *Jama-Journal of the American Medical Association no.* 309(8): 800–805. DOI: https://doi.org/10.1001/jama.2013.776.

Kim, H.J., M.P. Fay, E.J. Feuer, and D.N. Midthune. 2000. "Permutation tests for joinpoint regression with applications to cancer rates." *Stat Med no.* 19(3): 335–351. DOI: https://doi.org/10.1002/(sici)1097-0258(20000215)19:3 < 335::aid-sim336 > 3.0.co;2-z.

Kim, H.J., B. Yu, and E.J. Feuer. 2008. "Inference in segmented line regression: a simulation study." *Journal of Statistical Computation and Simulation no.* 78(11): 1087–1103. DOI: https://doi.org/10.1080/00949650701528461.

Lerman, P.M. 1980. "Fitting Segmented Regression Models by Grid Search." *Journal of the Royal Statistical Society. Series C (Applied Statistics) no.* 29(1): 77–84. DOI: https://doi.org/10.2307/2346413.

National Center for Health Statistics. 2009. *Comparability of Cause-of-death Between ICD Revisions*. Available at: https://www.cdc.gov/nchs/nvss/mortality/comparability_icd.htm (accessed March 2019).

Noone, A.M., N. Howlader, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D.R. Lewis, H.S. Chen, E.J. Feuer, and K.A. Cronin (eds). 2018. *Cancer Statistics Review, 1975–2015*, National Cancer Institute, Bethesda, MD, U.S.A. Available at: https://seer.cancer.gov/csr/1975_2015/ (accessed March 2019).

Received March 2019
Revised September 2019
Accepted November 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 63–88, http://dx.doi.org/10.2478/JOS-2020-0004

Statistical Challenges in Combining Survey and Auxiliary Data to Produce Official Statistics

Andreea L. Erciulescu[1], *Nathan B. Cruze*[2], *and Balgobin Nandram*[3]

Combining survey and auxiliary data to produce official statistics is gaining interest at federal agencies and among policy makers due to its efficiency. Recent studies have shown the practicality of small area estimation modeling approaches in the context of integrating data from multiple sources to improve estimation at fine levels of aggregation. In this article, agricultural predictions are constructed using a hierarchical Bayes subarea-level model, fit to data available from different sources. Auxiliary data are initially used to complement the survey data and define the prediction space, and then to define covariates for the model. Finally, not-in-sample predictions are constructed using the model output, and benchmarking constraints are imposed on the final set of in-sample and not-in-sample predictions. Unlike most of the studies discussing not-in-sample prediction, this article illustrates a method that uses the data available from multiple sources to define the prediction space. As a consequence, the resulting framework provides a larger set of nationwide predictions as candidate for official statistics, and extrapolation is not of concern. Challenges in developing the methods to combine different data sources are discussed in the context of planted acreage prediction.

Key words: Administrative data; benchmarking; incomplete data; not-in-sample prediction; small area estimation.

1. Introduction

Survey summary statistics at disaggregated levels may not be fit for use as official statistics because the limited amount of information available may result in estimates with high levels of uncertainty. With an increase in available data from auxiliary sources, an increase in needs for official statistics at detailed levels of aggregation and a decrease in allocated budgets, federal agencies have an increased interest in using models in the estimation process. In this article, we consider novel ways of using administrative data in the process of constructing official statistics. Specifically, administrative data are used to complement the survey data and define the set of domains for which predictions are needed. Then,

[1] Westat, 1600 Research Blvd., Rockville M.D., U.S.A. Email: alerciulescu@gmail.com
[2] USDA National Agricultural Statistics Service, Research and Development Division, 1400 Independence Avenue, SW, Washington D.C., U.S.A. Email: nathan.cruze@usda.gov
[3] Worcester Polytechnic Institute, Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609-2247, Massachusetts, 01609, U.S.A. Email: balnan@wpi.edu

Acknowledgments: The findings and conclusions in this preliminary publication have not been formally disseminated by the U.S. Department of Agriculture and should not be construed to represent any agency determination or policy. This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agriculture Statistics Service. The work of Dr. Erciulescu was conducted while she was a Research Associate with the National Institute of Statistical Sciences, working at the National Agriculture Statistics Service. The authors thank the Associate Editor and the two referees for their feedback and questions that led to an improved manuscript.

models that integrate survey and administrative data are used to construct predictions for domains with survey sample sizes as small as zero. This work builds on a series of research studies conducted at the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) to innovate the current methods of setting official statistics for acreage, production and yield at state and substate levels of aggregation. We consider data collected by the USDA's NASS using a probability sample and auxiliary data from other sources, to produce end-of-season county-level and agricultural statistics district-level predictions for planted acreage, where an agricultural statistics district (hereafter, denoted by district) is defined as a group of contiguous counties within a state.

Area-level and subarea-level models are excellent reproducible tools that combine survey data and auxiliary data to produce reliable estimates for areas where survey estimates are available. In the area-level model, introduced by Fay and Herriot in 1979 (FH), the survey estimates, $\hat{\theta}_k$, are modeled using the sampling model,

$$\hat{\theta}_k|(\theta_k, \hat{\sigma}_k^2) \overset{ind}{\sim} N(\theta_k, \hat{\sigma}_k^2),$$

where $\hat{\sigma}_k^2$ are the estimated sampling variances and k $(k = 1, \ldots, m)$ is an index for the small areas. The small area parameters of interest θ_k are estimated using a linking model,

$$\theta_k|(\beta, \sigma_u^2) \overset{ind}{\sim} N(\mathrm{z}_k'\beta, \sigma_u^2), \tag{1}$$

where z_k are area-level covariates with p components, including an intercept, and (β, σ_u^2) is a vector of nuisance parameters. A rich literature is available for the FH model and its extensions, using both frequentist and Bayesian methods. In a hierarchical Bayes analysis, prior distributions are assigned to (β, σ_u^2).

As an extension to the FH model, Fuller and Goyeneche (1998) introduced a subarea-level model (FG) to account for a grouping structure of the subareas into areas. The survey estimates at the subarea level, $\hat{\theta}_{ij}$, are modeled using the sampling model,

$$\hat{\theta}_{ij}|(\theta_{ij}, \hat{\sigma}_{ij}^2) \overset{ind}{\sim} N(\theta_{ij}, \hat{\sigma}_{ij}^2),$$

where $\hat{\sigma}_{ij}^2$ are the estimated sampling variances, j $(j = 1, \ldots, n_i^c)$ is an index for the subareas, i $(i = 1, \ldots, m)$ is an index for the areas, and $n^c = \sum_{i=1}^m n_i^c$ is the total number of subareas. The parameter of interest is the subarea mean θ_{ij}, which is estimated using a hierarchical linking model,

$$\begin{aligned} \theta_{ij}|(\beta, \sigma_u^2, v_i) &\overset{ind}{\sim} N(\mathrm{x}_{ij}'\beta + v_i, \sigma_u^2), \\ v_i|\sigma_v^2 &\overset{ind}{\sim} N(0, \sigma_v^2), \end{aligned} \tag{2}$$

where x_{ij} are subarea-level covariates with p components, including an intercept, and $(\beta, \sigma_u^2, \sigma_v^2)$ is a vector of nuisance parameters. Torabi and Rao (2014) studied the FG model in a frequentist framework and Kim et al. (2018) extended the linking model in Torabi and Rao (2014) to allow for a hierarchical level for parameters β and to remove distributional assumptions in the first hierarchical level. Erciulescu et al. (2018, 2019) studied the FG model using a hierarchical Bayes framework, adopting prior distributions for $(\beta, \sigma_u^2, \sigma_v^2)$.

In the area-level (subarea-level) sampling models, it is assumed that $\hat{\theta}_k(\hat{\theta}_{ij})$ and $\hat{\sigma}_k^2(\hat{\sigma}_{ij}^2)$ are valid estimates available from the survey summary, that is the estimates exist and are in the parameter space (positive total acreage estimates and positive sampling variances). However, for the not-in-sample subareas (domains with missing survey data), inference conducted relies on the linking model's specification. Given (1), a typical choice of estimator for the not-in-sample areas is the synthetic estimator $\mathrm{z}_k'\beta$, see Rao and Molina (2015) for more information on regression synthetic estimation. While one choice for a not-in-sample subarea estimator, given (2), is the synthetic estimator $\mathrm{x}_{ij}'\beta$, a better estimator is the composite estimator $\mathrm{x}_{ij}'\beta + v_i$ (note the contribution of both the subarea-level auxiliary data and the area-level random effect). In a Bayesian approach, the predictions are drawn from the assumed linking model (1) or (2), for area-level or subarea-level, respectively.

Building on the work of Erciulescu et al. (2019), we combine survey and auxiliary data and use a subarea-level model to construct planted acreage predictions for a set of counties defined by the union of all the available data sources. Statistical challenges and breakthroughs in combining data from multiple sources to produce official statistics are discussed throughout the paper. In particular, we identify a common geographic level and time point to combine data from a probability survey with nonprobability data from three administrative sources, the latter lacking uncertainty measures. As in Erciulescu et al. (2019), we treat the auxiliary data as fixed and free of error, but details on potential error sources in these data are available in Erciulescu et al. (2019). Note that Erciulescu et al. (2019) investigated these sources only for predictive power, and used only one at a time in developing the models (to avoid multicollinearity problems). Also, the authors tackled prediction for harvested acreage only for counties with both sample and administrative data available. In this article, we integrate all the data to identify the set of counties with planting activity for a specific crop (or the prediction space), in a given crop season, and to construct a covariate with good predictive power and observations available for all the counties in the prediction space. Challenges in multistage, nationwide prediction for counties with sample sizes as small as zero (not the case in Erciulescu et al. 2019) are addressed using hierarchical Bayes subarea-level models.

Modeling strategies are developed to deal with incomplete data and benchmarking methods are implemented to overcome the challenge of attaining consistency among predictions at nested levels of aggregation. Whereas Erciulescu et al. (2019) developed and compared models for direct estimates scaled by the sample sizes, with a hierarchy for sampling variances and different benchmarking methods, here we adopted the model for the direct estimates, with fixed sampling variances and the ratio adjustment, as a practical method with good performance that allows for prediction for subareas with sample sizes of just one or even zero where suggested by auxiliary information. This outcome was not possible under the model specification pursued in Erciulescu et al. (2019). Moreover, due to the extended prediction space, the possible over-adjustment due to benchmarking is less of a concern for NASS than it was for Erciulescu et al. (2019). As a result, a crop-specific framework of producing predictions is presented, with the potential to increase the number of official statistics constructed using current methodology.

In summary, the major contributions of this article are as follows:

- integration of all the available data to define the prediction space, as well as a covariate with good predictive power;

- modeling strategies to deal with incomplete administrative data and missing at random (MAR) assumption;
- not-in-sample prediction;
- reduction in over-adjustment due to benchmarking; and
- increase in the number of official statistics, given a common criterion.

In Section 2, we introduce different data sources and present a method that combines survey data and administrative data to identify and predict planted acreage for in-sample and not-in-sample subareas of interest for a certain crop, that is, county-level corn, as in the case study illustrated here. In Section 3, modeling strategies addressing different scenarios of available data and the corresponding derived predictors are presented. In Section 4, we present nationwide prediction results for 2015 corn planted acreage, including model efficiency and different contributions of administrative data to produce official statistics. A discussion is provided in Section 5. Additional results on corn, soybean, sorghum and winter wheat are presented in the Appendix (Section 6).

2. Data for Modeling End-of-Season Crop Acreage

County-level survey estimates may be improved using auxiliary information and small area model-based procedures, especially for counties with small sample sizes. Estimation challenges are driven by the needs for multi-stage (county, district, state), nationwide, estimates, constructed using a small amount of survey data. In this section, we describe the sources of data considered to produce small area model predictions for end-of-season crop planted acreage for corn in 2015. Next, we introduce a method that combines survey data and administrative data to identify the 2015 in-sample and not-in-sample counties of interest for corn planted acreage prediction. Finally, we investigate the potential for using auxiliary data as covariates in hierarchical models. The NASS survey data and the auxiliary data available from other USDA agencies on corn planted acreage are combined at the county level for each state.

2.1. NASS Survey Data

The probability sample of interest in this study is the pooled sample from the quarterly crops Agricultural Production Surveys (USDA NASS APS 2018) and their supplement, the County Agricultural Production Surveys (USDA NASS CAPS 2018), and will be denoted hereafter by CAPS. Due to the updates to the list sampling frame and the survey questionnaires, and to the year-to-year changes in planting activity, the set of subareas to be estimated for a given year-commodity combination is not predefined. For example, each survey response includes information on the entire operation (farm or ranch), and for all the sampled commodities with activity in the given season. As a result, the number of known operations in a county may change over time, the number of sampled operations may vary from year to year, and each of the operations may vary the type of crops grown annually. See Appendix A in National Academies of Sciences, Engineering, and Medicine (2017) for more details on NASS's survey design and data collection.

County-level and district-level survey estimates and associated variance estimates are available from the NASS's CAPS summary. The district-level survey data are derived directly from the county-level survey data and, hence, only the county-level data will be used for modeling. The district-level survey data will be used for comparing model predictions to the survey estimates. In the 2015 crop season, NASS sampled 36 states for corn. The 36 states were comprised of 2,837 counties, and NASS produced survey estimates for 2,426 in-sample counties. Survey estimates are not available for the remaining 411 counties; we refer to these counties as not-in-sample with respect to corn. A nationwide map of the end-of-season positive county-level planted acreage survey estimates available for corn in 2015 is shown in Figure 1. The 12 states that were not sampled for corn in 2015 are represented as blank states with a black dot. The counties with zero planted acreage predictions and not-in-sample counties for corn in 2015 are represented in white. Since the range of planted acreages in counties with available sample data is state-dependent and can vary from tens to hundreds of thousands of acres, the county-level map in Figure 1 depicts estimates on the log(10) scale. Dark areas correspond to high acreage intensity regions, in particular the Midwestern corn belt states.

As a result of the NASS survey and publication cycle, state-level planted acreage values are prepublished and considered as fixed targets in the substate-level estimation process. The sum of the county-level survey estimates in a state does not necessarily equal the prepublished state-level value, the latter being the result of an expert assessment of multiple sources of data (including, but not limited to the survey data). Hence, one of the challenges encountered is to attain consistency among estimates constructed for nested levels. To overcome this challenge, we study a benchmarking adjustment applied to the substate-level predictions, for the county-to-district-to-state agreement to hold. More details on the benchmarking adjustment we utilize are presented in Subsection 3.3.

The number of counties and districts vary across the states and across commodities. For 2015 corn, the number of counties within districts ranges from 1 to 32, with a median of 8 and the number of districts within state ranges from 3 to 15, with a median of 9. Because the source of survey data for this study is the survey summary at the county level and district level, we denote the sample size by the number of positive records used to construct the survey summary; a positive record refers to a survey record for which

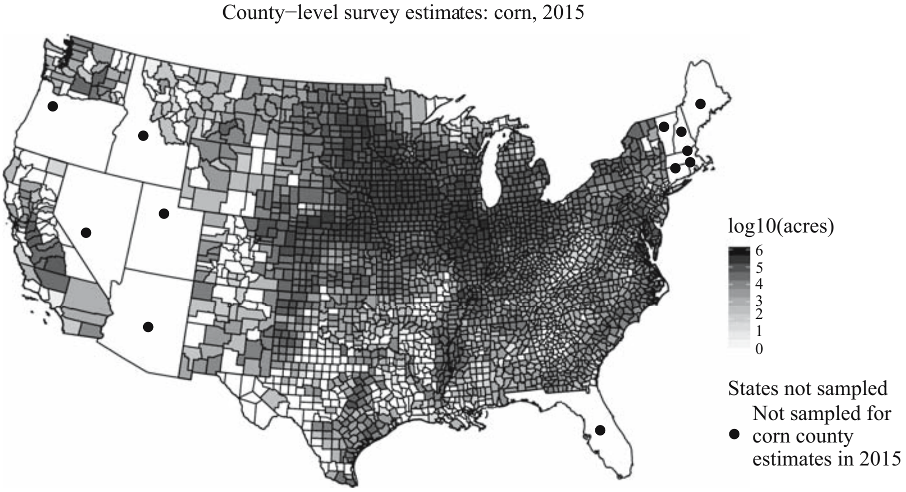

County−level survey estimates: corn, 2015
log10(acres)
6
5
4
3
2
1
0
States not sampled
● Not sampled for corn county estimates in 2015

Fig. 1. Nationwide map of the end-of-season positive county-level planted acreage survey estimates available for corn in 2015 from the NASS CAPS summary, with all non-zero estimates on the log(10) scale.

positive acreage was reported. The county sample size differs from state to state and commodity to commodity. For corn, the county sample sizes range from 1 to 191, with a median of 18 and the district sample sizes range from 1 to 993, with a median of 206; the sample size ranges for Illinois are illustrated on the x-axes in Figure 2.

The estimated coefficients of variation (CVs) for the survey estimates increase as the county sample sizes decrease, and their ranges also differ from state to state and commodity to commodity. For 2015 corn, the CVs of the county-level survey estimates range from 0.07% to 107.66%, with a median of 31.94%, and the CVs of the district-level survey estimates range from 3.27% to 100.70%, with a median of 10.67%. Figure 2 shows the inverse relationship between the CVs of the 2015 corn county-level planted acreages survey estimates in Illinois and the corresponding sample sizes. Similar patterns are observed in other states, and for other commodities.

2.2. *Auxiliary Data*

We explore auxiliary data, available from three USDA agencies: NASS, the Farm Service Agency (FSA) and the Risk Management Agency (RMA). FSA administers US farm programs, such as county-level revenue loss protections (USDA FSA 2019). RMA oversees the Federal Crop Insurance Corporation, which provides crop insurance to participating farmers and agricultural entities (USDA RMA 2019). For this, FSA and RMA collect data from farmers participating in such programs. NASS produces the Cropland Data Layer (USDA NASS CDL 2018), a crop-specific land cover product that uses satellite and FSA ground-reference data to classify crop types in the continental United States (Boryan 2011; USDA NASS 2016a).

The levels and time of availability, and potential sources of error vary by data source (FSA, RMA, NASS), geography and commodity. Combining data from multiple sources and assessing its quality and usability is a challenging effort, often not mentioned in small area studies. For example, the CAPS sample data are collected on farms or ranches that the

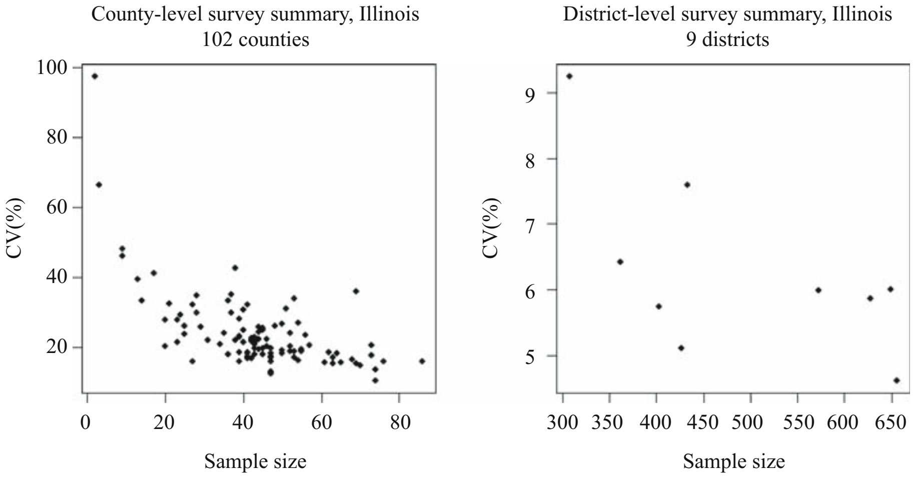


Fig. 2. Plots of CVs of the 2015 survey county-level and district-level estimates of planted acreage of corn in Illinois against corresponding sample sizes.

respondents operate and participation in the FSA and RMA programs is popular, but not compulsory; farmers who choose to participate in either agency's support programs supply data to the FSA and RMA administrative offices voluntarily. However, the definition of farm or ranch and the spatial unit used differ among the three data sources: NASS, FSA and RMA (National Academies of Sciences, Engineering, and Medicine 2017, 96–97). Linking data at a fine scale has been of interest to NASS, but final solutions have yet to be developed. The administrative data of interest for this study are the self-reported corn planted acreage values supplied to FSA and RMA and the acreage values derived from pixels classified as corn, aggregated at the county level and comprising the nonprobability sample data under consideration.

Quantifying the quality of nonprobability sample data has been of interest to many government agencies, but conclusive studies have yet to be published. Parsons (1996) evaluated the quality of FSA acreage totals with respect to coverage. Kennedy et al. (2016) evaluated nonprobability surveys and assumed that the nonprobability samples were drawn as simple random samples from the population and constructed pseudo-weights when constructing domain estimates and associated measures of uncertainty. While we acknowledge potential error sources in the aggregated data, in this study we will assume the nonprobability county-level values from FSA, RMA and CDL as fixed and free of error. In Table 1, we report a summary of the number of counties with data available on corn planted acreage in 2015 from at least one source. Note that the sets of counties with data available from either of the four sources are not mutually exclusive, as depicted in the Venn diagram in Figure 3. After accounting for the 2,726 counties with corn planted acreage identified from the CDL, additional planted acreage activity is identified in only 22 ($= 11 + 3 + 6 + 0 + 1 + 1 + 0$) counties from the CAPS, FSA and RMA (see Figure 3). Hence, our goal is to construct 2015 corn predictions for the total of 2,748 counties. The number of counties with corn planting activity differs across years, states, commodities and data sources.

The county-level quantity of interest is the total planted acreage and the values available from the three sources (FSA, RMA, CDL) of auxiliary data are measurements of the same county-level quantity, that is corn total planted acreage. It is known that all three sources may suffer from downward biases (see Cruze et al. 2019 for a literature review of geography and remote sensing studies). As an attempt to avoid the possible downward bias and obtain a covariate with good predictive power for total county-level acreage, we combine the three sources to construct one set of values indicating the maximum number of available corn planted acreages, reported by volunteers or remotely classified. Let Admin PL denote the constructed variable as such. If all FSA, RMA and CDL values are available, then the maximum value is considered. If only two of the values are available,

Table 1. Counties, in sampled states, with corn planting activity, 2015.

Data source (USDA)	Number of counties
NASS CAPS	2426
FSA	2398
RMA	2230
NASS CDL	2726

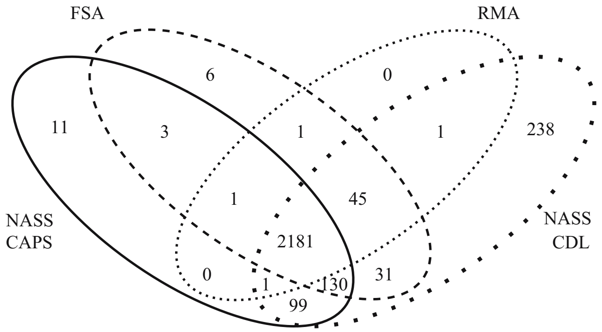


Fig. 3. Counties, in sampled states, with corn planting activity, 2015.

the maximum value is considered. If only one of the values is available, then that value is considered. To investigate the additional contributions of the CDL data, we will also consider an Admin PL variable, as derived from FSA and RMA data only, and present results in Section 4.

2.3. Borrow Information from Multiple Data Sources

As expected, nationwide analysis indicates strong linear relationships between the survey estimates and the administrative data for all the 36 states. For each sampled state, a simple regression model was fit to the survey estimates, using intercept and FSA, RMA, CDL or Admin PL as predictor. Summaries of R^2 values and estimated slope coefficients $\hat{b}$, for all the states, are reported in Table 2 (25%, 50%, 75% quantiles).

In Figure 4 we display the linear fit between the survey estimates and the derived administrative values, Admin PL, and in Figure 5 we display the linear fits between the survey estimates and the values available from each of the three auxiliary sources, FSA, RMA and CDL, respectively. As a result of this analysis, Admin PL will be included as a covariate in the model described in the next section.

3. Modeling Strategies

The proposed model for a given state is a subarea-level model, where the area represents the district, the subarea represents the county and the subarea-level survey variances are treated as fixed and known. Of interest is prediction of planted acreage at the county and district levels. Prediction is conducted state by state and commodity by commodity, for all counties within states identified to have planted acreage activity in the given crop season.

3.1. Hierarchical Bayes Model

Let $i = 1, \ldots, m$ be an index for the m districts in the state under consideration; $j = 1, \ldots, n_i^c$, be an index for the n_i^c counties in district i; and n_{ij} be the sample size of the j^{th} county in the i^{th} district. The total number of counties in the state is $\sum_{i=1}^{m} n_i^c = n^c$ and the state sample size is $\sum_{i=1}^{m} \sum_{j=1}^{n_i^c} n_{ij} = n$.

Let $\hat{\theta}_{ij}$ be the (total planted acreage) survey estimate for county i in district j and $\hat{\sigma}_{ij}^2$ be the corresponding estimated survey variance. For now, assume that county-level covariate

Table 2. Nationwide summaries for linear regression models applied to the data for every sampled state.

	FSA			RMA			CDL			Admin		
	1st Quantile	Median	3rd Quantile	1st Quantile	Median	3rd Quantile	1st Quantile	Median	3rd Quantile	1st Quantile	Median	3rd Quantile
R^2	0.82	0.89	0.92	0.76	0.86	0.91	0.85	0.90	0.93	0.85	0.90	0.93
$\hat{b}$	0.85	0.91	0.99	0.89	0.97	1.17	0.75	0.84	0.91	0.75	0.84	0.89

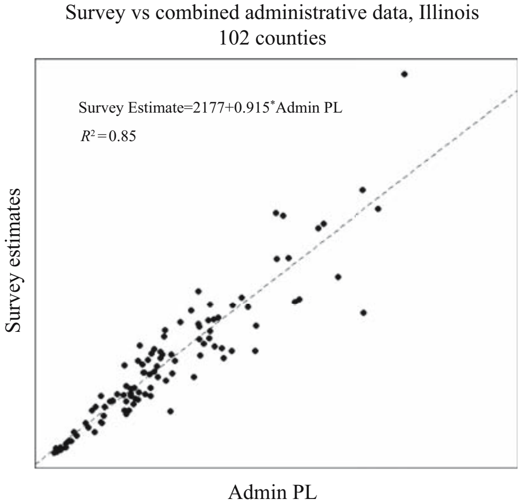


Fig. 4. Plot of survey estimates against derived administrative data values of planted acreage of corn (maximum value from available administrative sources) overlaid with best simple linear regression line.

values x_{ij} are available; a discussion on the availability of such covariates is provided later. Illustrated for one state, one commodity and one parameter, the hierarchical Bayes subarea-level model is

$$\hat{\theta}_{ij}|(\theta_{ij}, \hat{\sigma}_{ij}^2, v_i) \overset{ind}{\sim} N(\theta_{ij}, \hat{\sigma}_{ij}^2), \tag{3}$$

$$\theta_{ij}|(v_i, \beta, \sigma_u^2) \overset{ind}{\sim} N(\mathrm{x}_{ij}'\beta + v_i, \sigma_u^2), \tag{4}$$

$$v_i|\sigma_v^2 \overset{ind}{\sim} N(0, \sigma_v^2). \tag{5}$$

The parameters $(\beta, \sigma_u^2, \sigma_v^2)$ are assumed independent a priori, for which noninformative, proper priors are adopted. The least squares estimates of β are obtained from fitting a simple linear model for the county-level survey estimates against the county-level auxiliary information, and then used as fixed and known parameters in the prior distribution for β. In particular, we adopt a multivariate normal prior distribution for β, with mean and variance denoted by the least squares estimate for the mean and the least squares estimate for the variance, multiplied by 10^3, respectively. By assigning a large

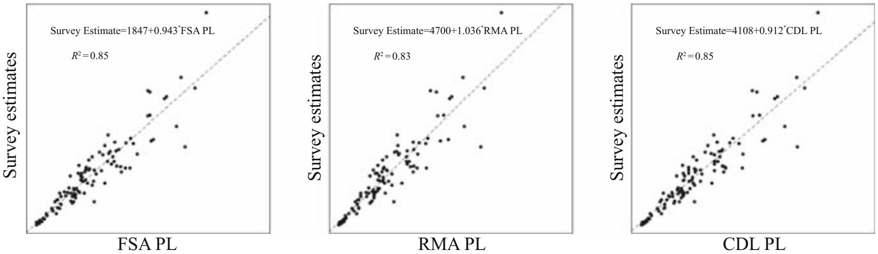


Fig. 5. Plots of survey estimates against administrative data values of planted acreage of corn available from the FSA, RMA, and CDL, respectively, overlaid with data-specific best simple linear regression lines.

prior variance, we adopt a diffuse prior for β. The prior distributions for the model variance components σ_u^2 and σ_v^2 are *Uniform*(0, 10^8) and *Uniform*(0, 10^8), respectively.

The model (3, 4, 5) borrows information from all the counties in a district and from all the districts in the state, while combining auxiliary information available at the subarea level, x_{ij}. The result model predictions are composite predictions, denoted by the weighted average of the subarea survey estimate and the best fitted values, after accounting for the area effect. That is, for a county j, in district i, the posterior mean is a predictor composed of the county-level survey estimator and a composite predictor of county-level synthetic predictor and a district-level effect predictor. The derivation is provided below.

Combining (3) and (4) using Bayes' theorem, we obtain the distribution of θ_{ij} given the data and the nuisance parameters,

$$\theta_{ij}|(v_i, \beta, \sigma_u^2, \sigma_v^2, \hat{\theta}_{ij}, \hat{\sigma}_{ij}^2) \overset{ind}{\sim} N(\gamma_{ij}\hat{\theta}_{ij} + (1-\gamma_{ij})(\mathrm{x}_{ij}'\beta + v_i), (1-\gamma_{ij})\sigma_u^2), \quad (6)$$

where $\gamma_{ij} = \frac{\sigma_u^2}{\hat{\sigma}_{ij}^2 + \sigma_u^2}$.

Integrating out θ_{ij} from (3) and (4), we obtain the conditional distribution of $\hat{\theta}_{ij}$,

$$\hat{\theta}_{ij}|(v_i, \beta, \sigma_u^2, \sigma_v^2, \hat{\sigma}_{ij}^2) \overset{ind}{\sim} N(\mathrm{x}_{ij}'\beta + v_i, \hat{\sigma}_{ij}^2\sigma_u^2). \quad (7)$$

Now, combining (5) with (7) using Bayes' theorem again, we obtain the conditional distribution of v_i,

$$v_i|(\beta, \sigma_u^2, \sigma_v^2, \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\sigma}}_i^2) \overset{ind}{\sim} N(\gamma_i(\bar{\hat{\theta}}_i^{\gamma} - \bar{\mathrm{x}}_i^{\gamma'}\beta), (1-\gamma_i)\sigma_v^2), \quad (8)$$

where $\gamma_{i.} = \sum_{j=1}^{n_i^c}\gamma_{ij}$, $\gamma_i = \frac{\sigma_u^2}{\sigma_v^2 + \sigma_u^2(\gamma_{i.})^{-1}}$, $\bar{\hat{\theta}}_i^{\gamma} = (\gamma_{i.})^{-1}\sum_{j=1}^{n_i^c}\gamma_{ij}\hat{\theta}_{ij}$, $\bar{x}_i^{\gamma} = (\gamma_{i.})^{-1}\sum_{j=1}^{n_i^c}\gamma_{ij}x_{ij}$, $\hat{\boldsymbol{\theta}}_i$ the vector of $\hat{\theta}_{ij}$s and $\hat{\boldsymbol{\sigma}}_i^2$ is the vector of $\hat{\sigma}_{ij}^2$s.

By the conditional mean formula in (6) and (8), it follows that the posterior mean of θ_{ij}, given the data and the nuisance parameters, is

$$\mathrm{E}(\theta_{ij}|\beta, \sigma_u^2, \sigma_v^2, \hat{\theta}_{ij}, \hat{\sigma}_{ij}^2) = \mathrm{x}_{ij}'\tilde{\beta} + \tilde{\gamma}_i(\tilde{\bar{\hat{\theta}}}_i^{\gamma} - \tilde{\bar{\mathrm{x}}}_i^{\gamma'}\beta) + \tilde{\gamma}_{ij}\left\{\hat{\theta}_{ij} - \mathrm{x}_{ij}'\tilde{\beta} - \tilde{\gamma}_i(\tilde{\bar{\hat{\theta}}}_i^{\gamma} - \tilde{\bar{\mathrm{x}}}_i^{\gamma'}\beta)\right\}, \quad (9)$$

where $\tilde{\gamma}_{ij} = \frac{\tilde{\sigma}_u^2}{\tilde{\sigma}_u^2 + \hat{\sigma}_u^2}$, $\tilde{\gamma}_{i.} = \sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}$, $\tilde{\gamma}_i = \frac{\tilde{\sigma}_v^2}{\tilde{\sigma}_v^2 + \hat{\sigma}_u^2(\tilde{\gamma}_{i.})^{-1}}$, $\tilde{\bar{\hat{\theta}}}_i^{\gamma} = (\tilde{\gamma}_{i.})^{-1}\sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}\hat{\theta}_{ij}$, $\tilde{\bar{\mathrm{x}}}_i^{\gamma} = (\tilde{\gamma}_{i.})^{-1}$ $\sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}x_{ij}$, and $\tilde{v}_i = \tilde{\gamma}_i(\tilde{\bar{\hat{\theta}}}_i^{\gamma} - \tilde{\bar{\mathrm{x}}}_i^{\gamma'}\tilde{\beta})$. The estimated variance parameters $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ are constructed as the posterior means for these parameters, that is $E(\sigma_u^2|\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2, \beta, \sigma_v^2, \theta_{ij})$ and $E(\sigma_v^2|\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2, \beta, \sigma_u^2, \theta_{ij})$, respectively.

Note that the posterior mean can be further rewritten as

$$\begin{aligned}\tilde{\theta}_{ij} &= \mathrm{x}_{ij}'\tilde{\beta} + \tilde{\gamma}_i(\tilde{\bar{\hat{\theta}}}_i^{\gamma} - \tilde{\bar{\mathrm{x}}}_i^{\gamma'}\tilde{\beta}) + \tilde{\gamma}_{ij}\left\{\hat{\theta}_{ij} - \mathrm{x}_{ij}'\tilde{\beta} - \tilde{\gamma}_i(\tilde{\bar{\hat{\theta}}}_i^{\gamma} - \tilde{\bar{\mathrm{x}}}_i^{\gamma'}\tilde{\beta})\right\} \\ &= \tilde{\gamma}_{ij}\hat{\theta}_{ij} + (1-\tilde{\gamma}_{ij})\left\{\mathrm{x}_{ij}'\tilde{\beta} + \tilde{v}_i\right\}. \quad (10)\end{aligned}$$

Using Equation (10), note the district-level contribution to the county-level not-in-sample predictions, v_i; for an area-level model, this term would be missing in Equation (10). On the other hand, Equation (10) may be rewritten as

$$\tilde{\theta}_{ij} = \tilde{\gamma}_{ij}\hat{\theta}_{ij} + (1-\tilde{\gamma}_{ij})\,\mathrm{x}_{ij}'\tilde{\beta} + (1-\tilde{\gamma}_{ij})\tilde{\gamma}_i(\tilde{\bar{\hat{\theta}}}_i^{\gamma} - \tilde{\bar{\mathrm{x}}}_i^{\gamma'}\tilde{\beta}), \quad (11)$$

where $\tilde{\gamma}_{ij}$, $\tilde{\gamma}_{i.}$, $\tilde{\gamma}_{i}$, $\tilde{\bar{\hat{\theta}}}_i^{\gamma}$ and $\tilde{\bar{\mathrm{x}}}_i^{\gamma'}$ are defined as for (10). The subarea-level and area-level components to the subarea-level posterior mean are clearly identified in Equation (11).

A discussion on the choice of county-level covariate values x_{ij} is provided in the next subsection, as it depends on the availability of the data. When available, the county-level covariate values, x_{ij}, are Admin PL values constructed as described above, and the model is denoted by M. For comparison, a model with no covariates and a model with Admin PL constructed using only the FSA and the RMA data are also fit, and denoted by M0 and M1, respectively. In addition, the comparison of models M and M1 may be of interest to the agency because the current NASS process of setting official statistics uses FSA and RMA data, but it does not use CDL data directly; see Cruze et al. (2019) for a detailed description of the process.

3.2. *Incomplete Data*

Complete sets of data are needed to define the counties with corn planted acreage activity and for model defined in (3), (4), and (5) to be fitted. One other challenge in combining data from multiple sources is the incomplete availability of the data. For this, we develop modeling strategies to account for three cases of available information for a given county j, in district i:

1. $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$ are available, but x_{ij} is missing,
2. $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$ are available, and
3. $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$ are missing, but x_{ij} is available.

The counties for which data are missing in all of the data sources considered, $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$, are excluded from the prediction set, because there is not enough evidence to conclude that planting activity took place for the specific crop, in the specific crop season. Not-in-sample predictions for these additional counties may be constructed using the methods for the third case above, after imputing covariate values x_{ij} (for example, using the average values available for other counties in the same division or state). However, not having any data to indicate county-level planting activity may lead to severe extrapolation and under-adjustments in the benchmarking step. For the cases with missing data in some of the sources, but available in others, we assume the missing at random (MAR) mechanism.

The first step in the modeling strategies is to impute the missing covariate values x_{ij}, for county j in district i, where survey estimates $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$ are available. For this, we use the x_{ij} values available for the most similar counties in the state. Similarity is defined using the absolute-value norm applied to the available survey estimates,

$$x_{ij} \leftarrow x_{ij'} | j' = arg\ min_k \{ |\hat{\theta}_{ik} - \hat{\theta}_{ij}| \},$$

over all counties k with survey and auxiliary data available. The resulting set of counties n^c with survey and auxiliary data $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$ available denotes all the counties with corn planting activity for the study.

After imputation, the models are fit to the n^c counties for which $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$ are available, using R JAGS (see Plummer et al. 2018), and posterior distributions are constructed using MCMC simulation. To estimate the nuisance parameters and the parameters of interest for the county-level total acreages, we use 3 chains, each of 10,000 Monte Carlo samples, 1,000 burn-in samples and thinned every nine samples.

Convergence diagnostics are conducted for selected states. The convergence is monitored using trace plots, the multiple potential scale reduction factors (values less than 1.1) and the Geweke test of stationarity for each chain (Gelman and Rubin 1992; Geweke 1992). Also, once the simulated chains have mixed, we construct the effective number of independent simulation draws to monitor simulation accuracy.

Using the chains of iterates obtained from the model fit, we construct posterior summaries from the posterior distributions of the nuisance parameters β^r, $(\sigma_u^2)^r$, $(\sigma_v^2)^r$, the county-level parameters of interest θ_{ij}^r and district-level parameters of interest $\theta_i^r := \sum_{j=1}^{n_i^c} \theta_{ij}^r$, where $r = 1, \ldots, R$, and R denotes the total MCMC iterates, after burn-in and thinning, equal to 3,000 in the application study.

In the last step in the modeling strategies, the model output from the complete data fit is used to predict for counties where $(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2)$ is missing but x_{ij} is available. For this, $\left\{\theta_{ij}^r\right\}_{r=1,\ldots,R}$ are drawn from the linking model (4, 5),

$$\theta_{ij}^r | (v_i^t, \beta^r, (\sigma_u^2)^r) \overset{ind}{\sim} N(\mathrm{x}_{ij}'\beta^r + v_i^r, (\sigma_u^2)^r).$$

3.3. Consistency Among Nested Levels

As discussed in the Section 1, NASS publishes the state-level value of corn planted acreage before estimation is conducted at the substate levels. To overcome the challenge of attaining consistency among predictions constructed for nested levels, we consider an external benchmarking adjustment that is timely and practically usable. A detailed discussion of classic benchmarking adjustments is given in Rao and Molina (2015). Studies on different benchmarking adjustments to crop acreage prediction are discussed in Erciulescu et al. (2019). In this section, we illustrate a benchmarking adjustment applied to the model predictions constructed under the different data availability cases, so that the county-level predictions aggregate to the district-level predictions and the district-level predictions aggregate to the prepublished state-level value.

Raking provides a suitable benchmarking adjustment to ensure consistency of substate predictions with state targets. For this study, we use the extension of the classic ratio adjustment given in Erciulescu et al. (2019), and we apply the constraint at the (MCMC) iteration level. This type of benchmarking adjustment is not adopted as part of the prior information or the model, but it facilitates its application to the set of in-sample and not-in-sample counties, in a small amount of time. For this, let the state-level target be denoted by a. Then the relation

$$\sum_{i,j}^{n^{c*}} \tilde{\theta}_{ij}^B = a, \tag{12}$$

needs to be satisfied, where n^{c*} is the total number of counties in the state and $\tilde{\theta}_{ij}^B$ is the final model prediction for county j and district i. Note that $n^{c*} = n^c + (n^{c*} - n^c)$, where n^c is the number of in-sample counties and $(n^{c*} - n^c)$ is the number of not-in-sample counties. The ratio adjustment is applied at the MCMC iteration level as follows

$$\theta_{ij,r}^B := \theta_{ij,r} \times a \times \left(\sum_{k=1}^{m} \sum_{l=1}^{n_k^{c*}} \theta_{kl,r} \right)^{-1}, \tag{13}$$

where $\theta^B_{ij,r}$ is the benchmarking-adjusted iteration, for $r = 1, \ldots, R$. Final county-level and district-level posterior summaries are constructed using the county-level iterates $\theta^B_{ij,r}$ and district-level iterates $\theta^B_{i,r} := \sum_{j=1}^{n_i^{c*}} \theta^B_{ij,r}$. For example, the resulting posterior means (variances) are constructed as Monte Carlo means (variances) of iterates. The county-level and district-level posterior means satisfy the multi-level benchmaking to state-level target a; note that n_i^{c*} is the total number of counties in district i.

From (13), note the importance of correctly specifying the set of counties to be estimated, since a smaller (larger) than the truth number of counties would result in an over-adjustment (under-adjustment) in the predictions.

4. Results

In this section, nationwide prediction results are presented for 2015 corn planted acreage, including a comparison of different models, model efficiency and different contributions of administrative data, serving towards the production of official statistics.

4.1. Model Comparison

Planted acreage data from the four sources summarized in Table 1 are used to define the set of counties to be estimated. For models fit and prediction, we define the set of counties with complete data after implementing the first step in the modeling strategies enumerated in Subsection 3.2. As previously mentioned, we consider three models for comparison: M0, the model fit to the survey data and no covariate; M1, the model fit to the survey data with one covariate derived from FSA and RMA data (directly and imputed, when applicable); and M, the model fit to the survey data with one covariate derived from FSA, RMA and CDL data

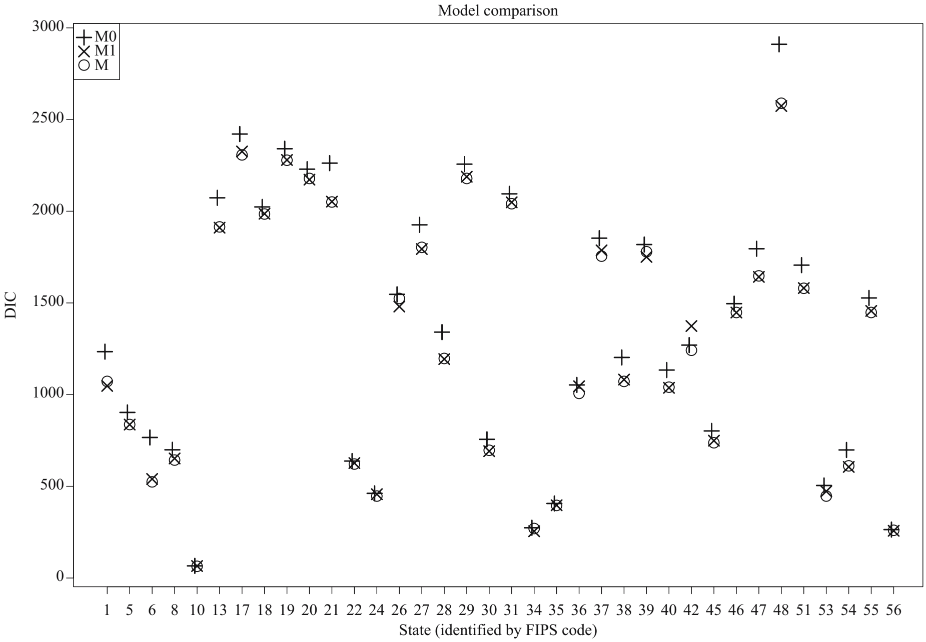


Fig. 6. Deviance information criterion (DIC) for models M0, M1, and M, by state.

Table 3. Summary of estimated factors $\tilde{\gamma}_{ij}$ (%).

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
Model M0	None	60.66	85.69	98.01
Model M1	FSA and RMA	2.67	11.41	44.92
Model M	**FSA, RMA and CDL**	2.42	10.25	40.94

(directly and imputed, when applicable). Note that the survey data modeled in all M0, M1 and M is the same, only the covariate data differ. Also note that various data sources are combined to construct *one* covariate (models M1 and M), therefore avoiding multicollinearity issues (as is the case when multiple covariates would correspond to the data sources).

The goodness of fit for models M0, M1 and M, fit state by state, is evaluated using the Deviance Information Criterion (DIC) and results are presented in Figure 6. The x-axis in Figure 6 illustrates the two-digit Federal Information Processing Standards (FIPS) codes for the 36 states, sampled for corn in 2015. Model comparison is conducted for each state, and not between states. The goodness of fit increases when auxiliary information is incorporated in the model, the best fit being when the Admin PL is defined using FSA, RMA and CDL. Models M1 and M result in similar performance; however, there are other benefits of using the CDL, as discussed in Section 5.

Models M0, M1 and M are further compared with respect to the contribution of auxiliary data to the final model predictions. Three-number summaries (25%, 50%, 75% quantiles) of the estimated factors $\tilde{\gamma}_{ij}$ (%) and $\tilde{\gamma}_i$ (%) defined for (10), are constructed over all the 36 states for which the models are fit and illustrated in Tables 3 and 4. Again, model predictions constructed using M1 and M have similar features. The auxiliary data and their relationship with the survey estimates receive larger weights in the final predictions under model M compared to model M0.

4.2. *Increased Number of County-Level Estimates*

Of great interest is the contribution of administrative data to increasing the number of county-level estimates. A nationwide map of the 2015 corn positive planted acreage county-level model predictions on the log10 scale, using model M, is illustrated in Figure 7. Model predictions are produced for 2,627 counties, of which 2,420 are in-sample counties and 207 are not-in-sample counties. Additionally, 121 model predictions were set to zero, because they corresponded to negative model predictions. Darker areas correspond to higher intensity regions. Not-in-sample predictions are mostly produced for counties located in non-major corn producing states and with small acreage amounts (the maximum not-in-sample model

Table 4. Summary of estimated factors $\tilde{\gamma}_i$ (%).

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
Model M0	None	85.37	92.25	95.48
Model M1	FSA and RMA	46.04	62.13	77.36
Model M	**FSA, RMA and CDL**	47.90	66.35	82.54

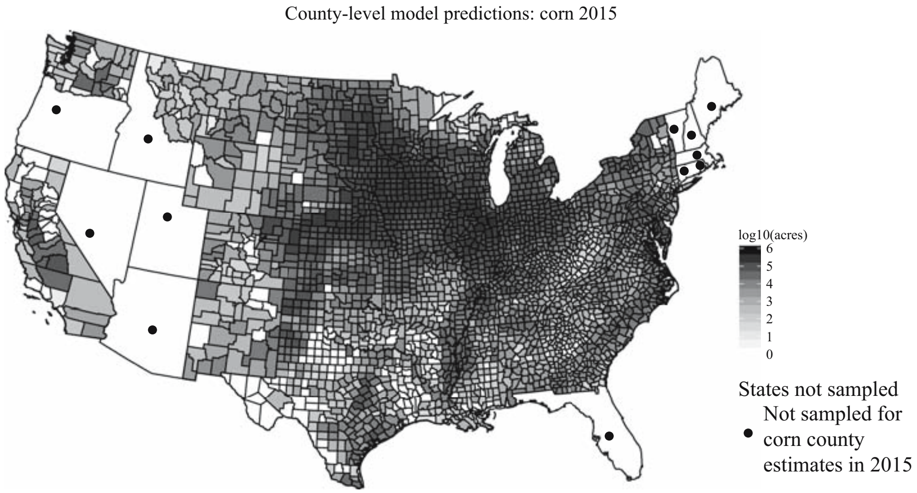

County-level model predictions: corn 2015
log10(acres)
6
5
4
3
2
1
0
States not sampled
● Not sampled for corn county estimates in 2015

Fig. 7. Nationwide map of model-M, positive, predictions of county-level planted acreage of corn in 2015, on the log10 scale; 121 negatively-valued model predictions are set to zero.

prediction is approximately 60% the median of the in-sample model predictions) and large CVs. In contrast, recall that survey estimates are available for 2,426 counties, as illustrated in Figure 1, and under model M1, 2,486 model predictions are produced.

4.3. *Model Efficiency*

Model efficiency comparisons are conducted for the set of counties where both a survey estimate and a model prediction are available. Compared to the survey estimates, the SEs and CVs of the model predictions are lower for most counties and districts. In Figure 8, we illustrate the reduction in CVs for the 2015 county-level estimates of corn planted acreage in Illinois, under model M.

In Tables 5 and 7, we illustrate nationwide results (25%, 50%, 75% quantiles), comparing the county-level survey SEs (CVs) to the model SEs (CVs) for models M1 and M. In Tables 6 and 8 we illustrate nationwide results (25%, 50%, 75% quantiles),

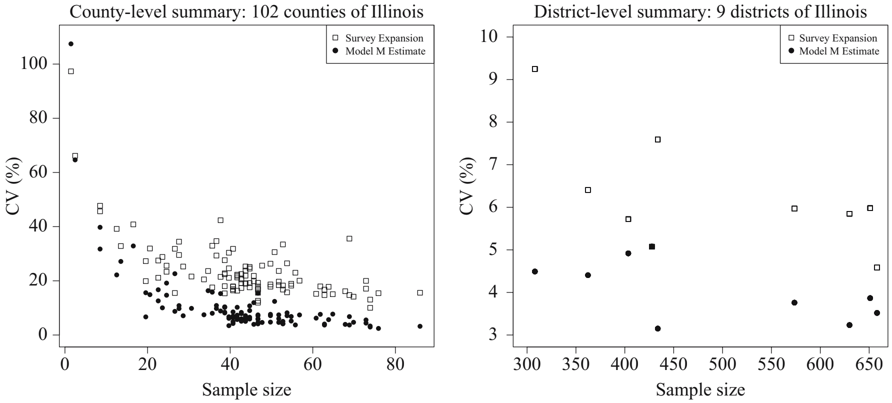


Fig. 8. Plots of CVs of survey estimates and model-M predictions of planted acreage of corn in 2015 against sample size, in Illinois, at the county level and district level.

Table 5. Summaries of standard errors of county-level survey estimates and model predictions (acres) Counties with available survey estimates.

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
Survey		640.90	2719.00	9494.00
Model M1	FSA, RMA	429.40	1233.00	2850.00
Model M	**FSA, RMA and CDL**	429.30	1166.00	2839.00

comparing the district-level survey SEs (CVs) to the model SEs (CVs) for models M1 and M. Comparing a model's performance versus a survey's performance based on precision (relative precision), we observe an increase in precision/relative precision in the range 34–70% (32–72%) in most of the county-level SE (CV) and in the range 27–57% (48–54%) in most of the district-level SE (CV), with slight improvement at the county level for model M versus model M1. We do not see an overall increase in precision at the district level for model M versus model M1 because the districts are composed of both in-sample and not-in-sample counties, and more predictions for not-in-sample counties are constructed under the two different models (M and M1, respectively).

The three-number summaries in Tables 5–8 do not reflect the relative efficiency at the domain (county or district) level. So, we report additional results in Figure 9, in the first row for 2,420 counties with positive survey estimates and model predictions, and in the second row for the corresponding 272 districts (which may include additional model predictions); counties or districts with relative efficiency values greater than 3 are removed to facilitate visualization. The relative SE (CV) is the ratio of the model prediction standard error (coefficient of variation) to the survey estimate standard error (coefficient of variation). Values larger than one for the county-level relative SE are due to the benchmarking adjustments and values larger than one for the district-level relative SE are due to the not-in-sample predictions and to the benchmarking adjustments.

5. Discussion

In this article, we illustrated the contributions of administrative data to produce agricultural official statistics. The methodology developed was illustrated using corn planted acreage, and the results for 2015 were presented. As an external validation exercise, models with specification M1 were fit to data from other years (2014, 2015, and 2016), and for commodities (corn, soybean, and sorghum). Blending survey and administrative data, we produce model county-level and district-level predictions for a set of counties predefined using in-sample data available from the survey summary and not-in-sample data available from administrative sources. The number of positive model

Table 6. Summaries of standard errors of district-level survey estimates and model predictions (acres).

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
Survey		4681.00	12220.00	36400.00
Model M1	FSA, RMA	2597.00	6121.00	15200.00
Model M	**FSA, RMA and CDL**	2958.00	6470.00	15310.00

Table 7. Summaries of CVs (%) of county-level survey estimates and model predictions Counties with available survey estimates.

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
Survey		21.08	31.91	55.42
Model M1	FSA, RMA	5.97	12.60	38.74
Model M	**FSA, RMA and CDL**	5.90	11.84	37.92

predictions is larger than the number of available survey estimates. As another external validation exercise, we compared the model predictions and the corresponding official values, for the counties and districts where both were available, using metrics such as median absolute difference, median absolute relative difference and credible interval coverage. In general, results indicated close agreement between the model predictions and the official values (constructed under the current NASS process).

Our first contribution is a novel use of administrative data to determine the set of subareas with crop-specific planting activity. We encourage similar investigations for other small area estimation applications where small domain characteristics are diverse within the large domains and not-in-sample predictions are of interest, such as agricultural applications (i.e., county-level cash rental rate estimation makes sense only for counties where at least one cash rental contract exists), health applications (i.e., youth smoking prevalence estimation make sense only for domains where at least one youth smoker actually exists) or education applications (i.e., estimation of Native American children aged 5–17 in poverty makes sense only for domains where at least one Native American child aged 5–17 lives).

In order to construct the prediction space, we assume that the data sources considered exhaust the information available on planting activities, for a specific crop, in a specific year. However, exploration of additional sources of data is of interest. When available, such additional information (state-specific, commodity-specific and time-specific) may be used to redefine the set of subareas for which model predictions are to be constructed and to redefine the set of covariates. Also, we acknowledge, but have to ignore the possible errors in administrative planting acreage values. One extension to deal with the possible downward bias in FSA, RMA, and CDL would be to adjust the model to

$$\hat{\theta}_{ij}|\theta_{ij}, \kappa_{ij} \overset{ind}{\sim} N(\kappa_{ij}\theta_{ij}, \hat{\sigma}^2_{ij}),$$

$$\kappa_{ij} \overset{ind}{\sim} \text{Uniform}(1, a_0), \quad \theta_{ij}|v_i, \beta, \sigma^2_u \overset{ind}{\sim} N(\mathrm{x}'_{ij}\beta + v_i, \sigma^2_u),$$

$$v_i|\sigma^2_v \overset{ind}{\sim} N(0, \sigma^2_v),$$

with the same priors adopted for the parameters $(\beta, \sigma^2_u, \sigma^2_v)$, a multiplicative offset κ_{ij} and a prespecified constant a_0, say between 1 and 1.1.

Table 8. Summaries of CV(%) of district-level survey estimates and model predictions.

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
Survey		7.03	10.50	16.04
Model M1	FSA, RMA	3.19	4.58	8.19
Model M	**FSA, RMA and CDL**	3.22	4.73	8.50

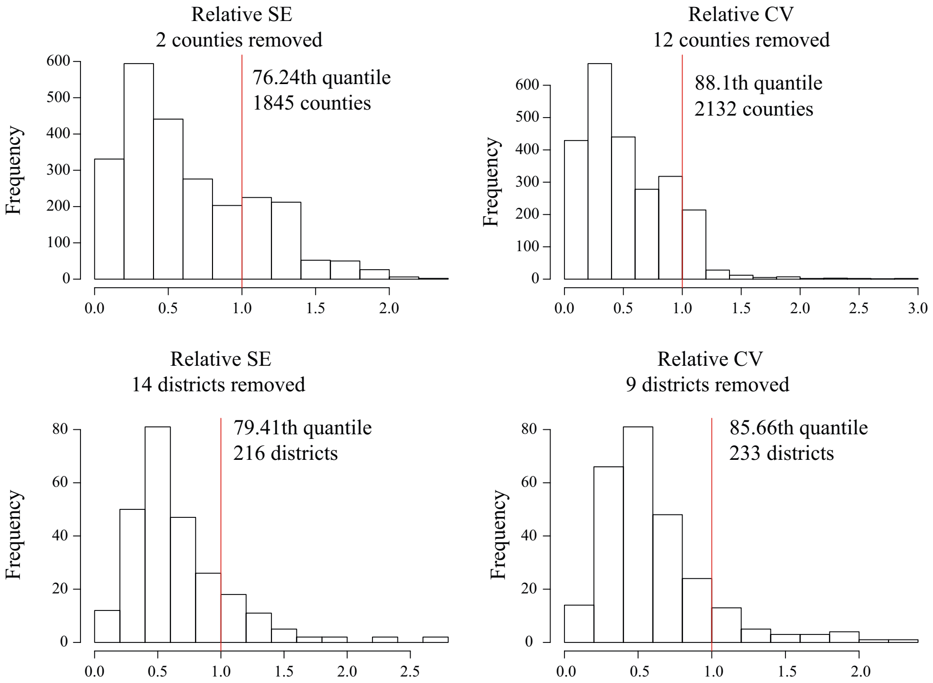


Fig. 9. Histograms of relative standard errors and relative CVs at the county level and district level, model versus survey. Relative efficiency values greater than 3 are removed to facilitate visualization.

For the methodology illustrated, we presented the implicit subarea-level weights associated with the different components of the final prediction. The contribution of administrative data to final predictions was evaluated using the parameter γ_{ij}. Model specifications, using a covariate derived from FSA and RMA data alone (M1), or from FSA, RMA and CDL data (M) are compared. Model M is slightly more efficient than model M1; however, it is important to note that, under model M1, 110 county-level Admin PL values were imputed, while under model M, only 11 county-level Admin PL values were imputed. Alternative strategies for imputation of missing auxiliary values are of interest for future research.

As a consequence of the model specification, in particular the normality assumption in the linking model, predictions are set to zero in some counties because the posterior means were negative. While we acknowledge that other choices of distributions may be considered, for example lognormal (or preferably generalized gamma distribution, lognormal being a special case), we recognize the simplicity of the current specification, especially with respect to prediction and benchmarking at multiple levels of interest. Under a non-normal distribution, the model predictions would need to be back-transformed. This additional operation would have to be performed at the lowest level of aggregation (for our application, the county), and followed by benchmarking adjustments and aggregations to higher levels of interest.

The models were applied separately, for each state, in order to follow with the current NASS process of constructing official statistics; results are communicated to each state individually, and final dissemination follows. One may extend the model to using a three-fold model by including an additional random effect corresponding to the states, and by using the nation-wide data. On the other hand, careful validation may be conducted at the state-level and specific auxiliary data, in addition to the ones considered here, may be incorporated.

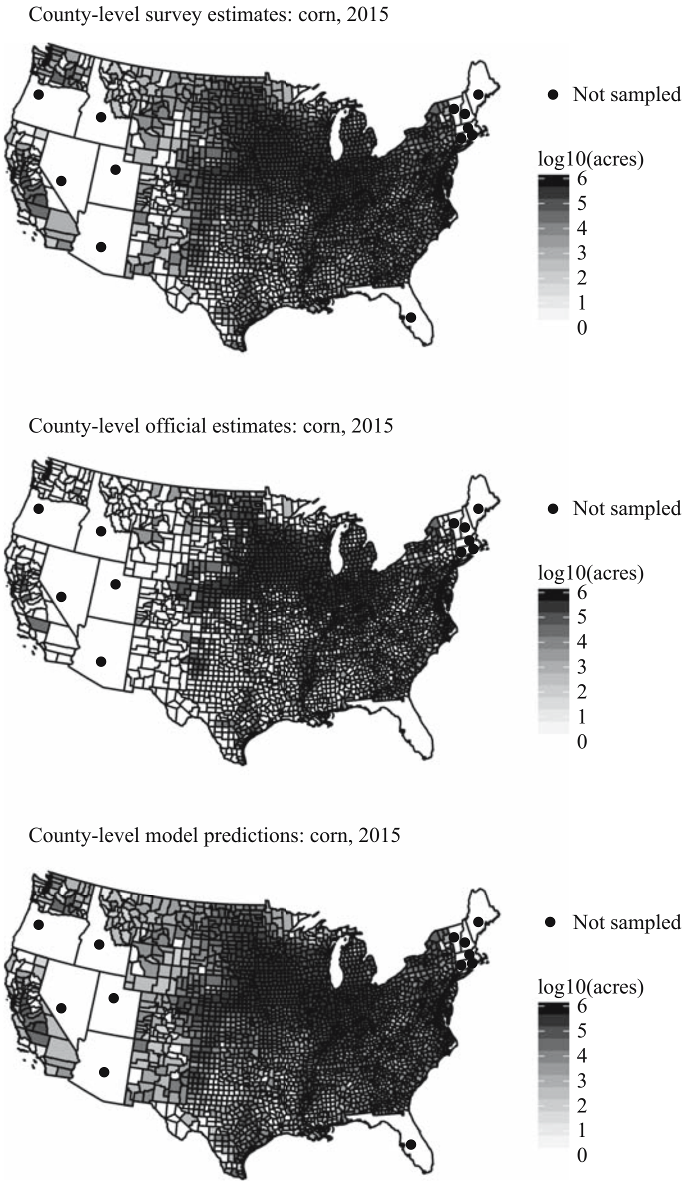
Increasing the number of counties with planted acreage predictions is another important contribution. For corn in 2015, the largest number of not-in-sample predictions happens to be in Texas: 42 out of 184 counties, accounting for approximately 0.7% of the total planted acreage. See the Appendix (Section 6) for additional results on soybean, sorghum and winter wheat. Hence, benchmarking only the set of counties where survey estimates are available would have resulted in over-adjusting the predictions. While the proportion of total acreage accounted for by the not-in-sample counties is small, the predictions play an important role in setting predictions for other variables of interest, such as harvested acreage, production and yield.

Finally, a major contribution of this paper is the operational framework presented, as it applies to any small area estimation application, from data preparation and challenges in dealing with specific features and incompleteness, to constructing a pool of predictions as candidates for official statistics. Addressing challenges associated with the publication process is an ongoing area of interest. The current NASS publication standard is based on the survey summary and on relative properties of the final estimates (the official statistics determined by NASS), for acreage and production; see the National Academies of Sciences, Engineering, and Medicine (2017, 117) for more details. For this application study, we investigate a hypothetical CV-based assessment, consistent with the publication standards at other government agencies (Marker 2015 reported CV-based assessments used by various government agencies). Using a 30% threshold for the county-level CVs across the nation leads to 1,694 candidate county-level planted acreage predictions for publication of corn in 2015; see Figure 10 in the Appendix (Section 6). In contrast, in 2015, NASS published estimates of corn for 1,433 counties, which are available in NASS QuickStats (USDA NASS 2016b). Moreover, in Equation (10), we provided the closed-form expression for the model predictions. Since they are composite predictions of various sources, the nationwide set of model predictions is a candidate for official publication. However, the challenge in constructing fit-for-use official statistics is the need for a publication standard that would permit publication of model predictions. While the current publication standard may be adopted for the model predictions, it would not make use of other properties of the model predictions, such as standard errors or credible intervals. The current NASS publication standard is being revised; see Cruze et al. (2018) for recent research on this topic.

6. Appendix: Increased Number of Reliable Estimates for Other Commodities

For corn and soybean in 2015, the largest numbers of not-in-sample predictions are, respectively, 42 and 70 out of 184 and 122 counties, accounting for approximately, respectively, 0.7% and 11.83% of the total planted acreage in Texas. The largest numbers of not-in-sample predictions for sorghum and winter wheat in 2015 are, respectively, 28 and 38 out of 73 and 154 counties, accounting for approximately, respectively, 5.23% and 12.47% of the total planted acreage in Mississippi and Georgia, respectively.

The county-level maps in Figures 10–13 depict positive survey (CAPS) estimates, official values and model (M) predictions on the log10 scale, for corn, soybean, sorghum and winter wheat, respectively. Dark areas correspond to high intensity regions.



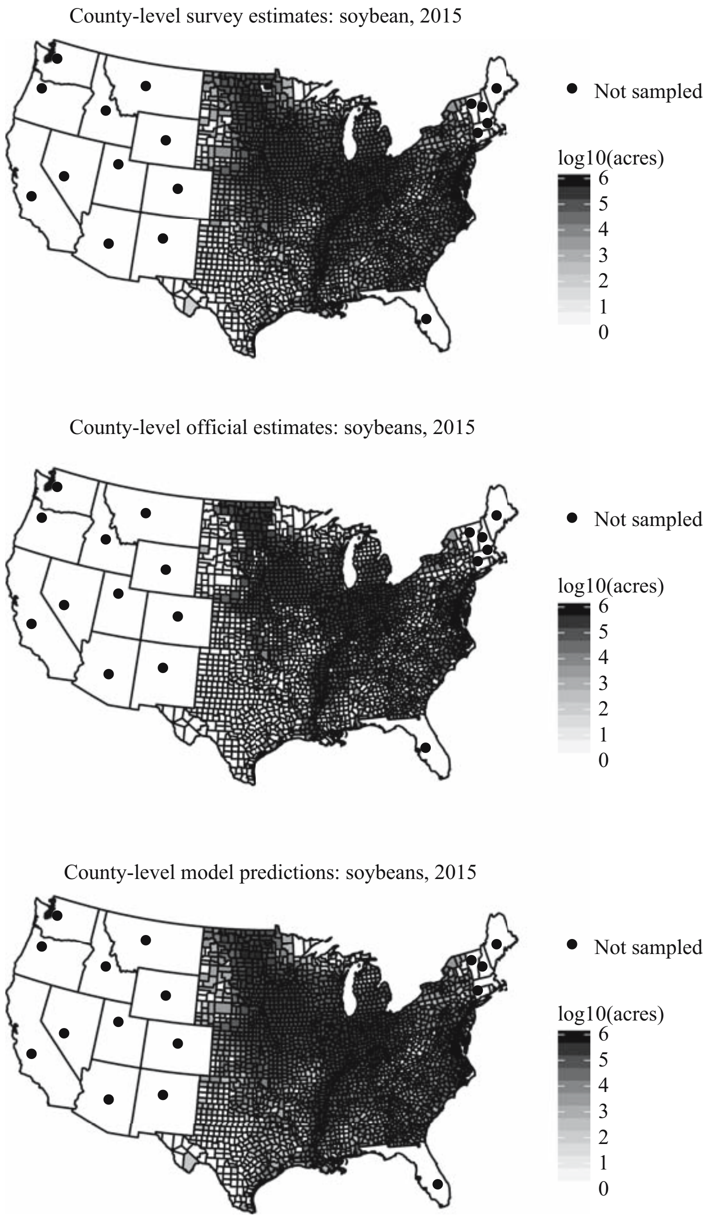
County-level survey estimates: corn, 2015

County-level official estimates: corn, 2015

County-level model predictions: corn, 2015

- 1,433 official values
- 2,426 survey estimates; 1,125 have CVs ≤ 30%
- 2,627 model predictions; 1,694 have CVs ≤ 30%
  - Texas: largest number of not-in-sample predictions, 42 out of 184 counties, accounting for ∼0.7% of planted acreage in the state
  - 121 zero predictions

Fig. 10. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of corn in 2015, on the log10 scale.



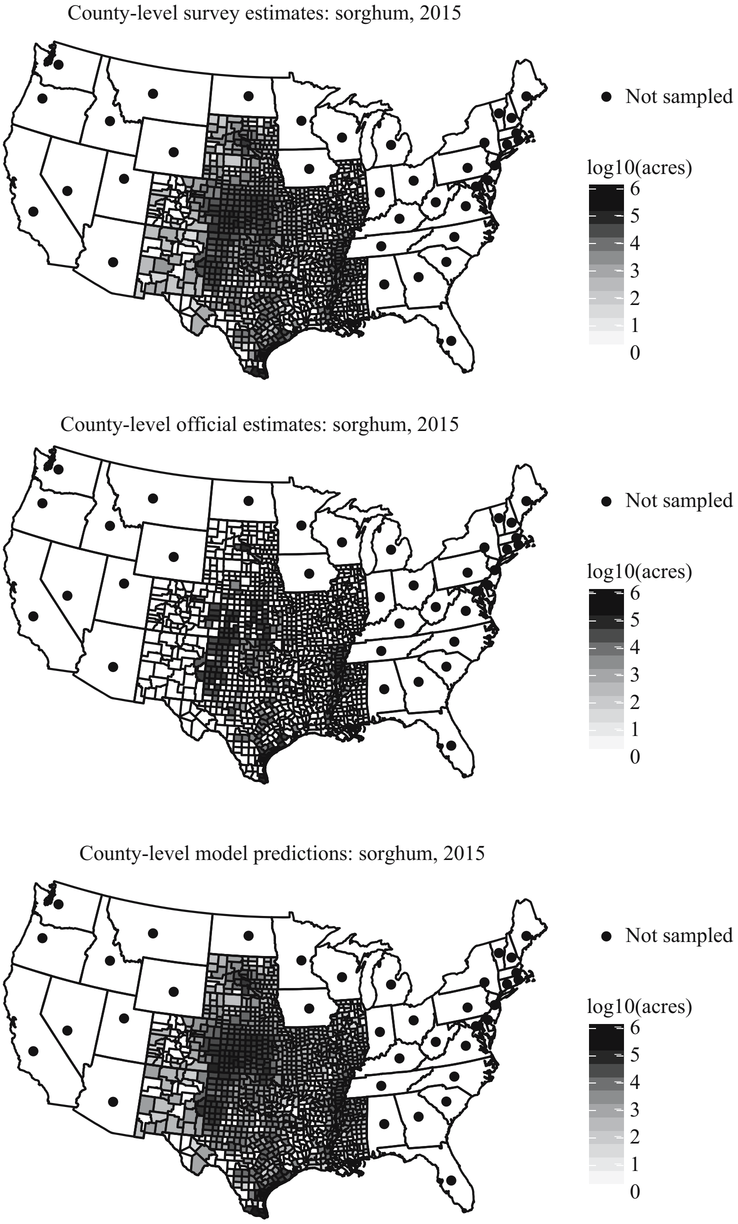
County-level survey estimates: soybean, 2015

County-level official estimates: soybeans, 2015

County-level model predictions: soybeans, 2015

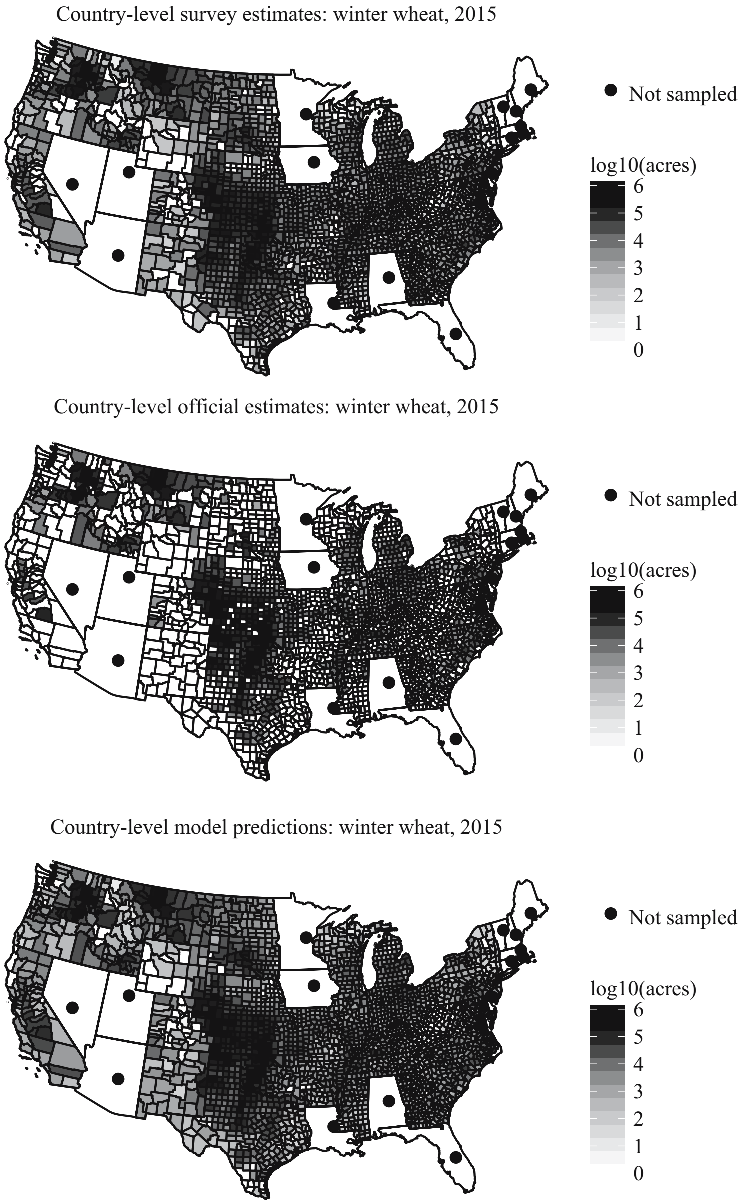
- 1,306 official values
- 2,012 survey estimates; 1,046 have CVs ≤ 30%
- 2,224 model predictions; 1,472 have CVs ≤ 30%
  - Texas: largest number of not-in-sample predictions, 70 out of 122 counties, accounting for ∼11.83% of planted acreage in the state
  - 173 zero predictions

Fig. 11. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of soybean in 2015, on the log10 scale.


County-level survey estimates: sorghum, 2015
● Not sampled
log10(acres)
6
5
4
3
2
1
0
County-level official estimates: sorghum, 2015
● Not sampled
log10(acres)
6
5
4
3
2
1
0
County-level model predictions: sorghum, 2015
● Not sampled
log10(acres)
6
5
4
3
2
1
0

- 218 official values
- 754 survey estimates; 135 have CVs ≤ 30%
- 922 model predictions; 390 have CVs ≤ 30%
  - Mississippi: largest number of not-in-sample predictions, 28 out of 73 counties, accounting for ∼ 5.23% of planted acreage in the state
  - 89 zero predictions

Fig. 12. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of sorghum in 2015, on the log10 scale.



Country-level survey estimates: winter wheat, 2015

Country-level official estimates: winter wheat, 2015

Country-level model predictions: winter wheat, 2015

- 1,049 official values
- 2,191 survey estimates; 697 have CVs ≤ 30%
- 2,417 model predictions; 1,321 have CVs ≤ 30%
  - Georgia: largest number of not-in-sample predictions, 38 out of 154 counties, accounting for ∼12.47% of planted acreage in the state
  - 64 zero predictions

Fig. 13. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of winter wheat in 2015, on the log10 scale.

7. References

Boryan, C., Z. Yang, R. Mueller, and M. Craig. 2011. "Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program." *Geocarto International* 26(5): 341–358. DOI: https://doi.org/10.1080/10106049.2011.562309.

Cruze, N.B., A.L. Erciulescu, B. Nandram, W.J. Barboza, and L.J. Young. 2019. "Producing Official County-Level Agricultural Estimates in the United States: Needs and Challenges." *Statistical Science* 34(2): 301–316. DOI: https://doi.org/10.1214/18-STS687.

Cruze, N.B., A.L. Erciulescu, H. Benecha, V. Bejleri, B. Nandram, and L.J. Young. 2018. "Toward an Updated Publication Standard for Official County-Level Crop Estimates." *Joint Statistical Meetings Proceedings. Government Statistics Section*. Alexandria, VA: American Statistical Association. 1576–1585. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/conferences/JSM-2018/Toward_an_updated_publication_standard_for_official_county-level_crop_estimates.pdf (accessed September 2019).

Erciulescu, A.L., N.B. Cruze, and B. Nandram. 2018. "Benchmarking a Triplet of Official Statistics." *Environmental and Ecological Statistics* 25: 523–547. DOI: https://doi.org/10/1007/s10651-018-0416-4.

Erciulescu, A.L., N.B. Cruze, and B. Nandram. 2019. "Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information." *Journal of the Royal Statistical Society, Series A* 182: 283–303. DOI: https://doi.org/10/1111/rssa.12390.

Fay, R.E. and R.A. Herriot. 1979. "Estimates of income for small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association* 74(366a): 269–277. DOI: https://doi.org/10.1080/01621459.1979.10482505.

Fuller, W.A. and J.J. Goyeneche. 1998. "Estimation of the state variance component." *Unpublished manuscript*.

Gelman, A. and D.B. Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical Science* 7: 457–511. DOI: https://doi.org/10.1214/ss/1177011136.

Geweke, J. 1992. "Evaluating the accuracy of sampling-based approaches to calculating posterior moments." In *Bayesian Statistics 4*, edited by J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford, UK: Clarendon Press.

Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez. 2016. "Evaluating Online Nonprobability Surveys," Pew Research Center. Available at: https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/ (accessed September 2019).

Kim, J.K., Z. Wang, Z. Zhu, and N.B. Cruze. 2018. "Combining Survey and Non-Survey Data for Improved Sub-Area Prediction Using a Multi-Level Model." *Journal of Agricultural, Biological, and Environmental Statistics* 23(2): 175–189. DOI: https://doi.org/10.1007/s13253-018-0320-2.

Marker, D. 2016. "Presentation to National Academy of Sciences Panel on Crop Estimates." *Unpublished presentation*. National Academy of Sciences report. Available at: https://www.nap.edu/catalog/24892/improving-crop-estimates-by-integrating-multiple-data-sources (accessed September 2019).

National Academies of Sciences, Engineering, and Medicine. 2017. "Improving Crop Estimates by Integrating Multiple Data Sources," Washington, DC: The National Academies Press. DOI: https://doi.org/10.17226/24892.

Parsons, J. 1996. "Estimating the Coverage of Farm Service Agency Crop Acreage Totals," *USDA NASS Research Report*, SRB-96-02. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/Survey_Reports/Estimating\%20the\%20Coverage\%20of\%20Farm\%20Service\%20Agency\%20Crop\%20Acreage\%20Totals.pdf (accessed September 2019).

Plummer, M., A. Stukalov, and M. Denwood. 2018. "Bayesian Graphical Models using MCMC," Version 4–8. Available at: https://cran.r-project.org/web/packages/rjags/rjags.pdf (accessed September 2019).

Rao, J.N.K. and I. Molina. 2015. *Small Area Estimation*. 2nd ed. Hoboken: Wiley.

Torabi, M. and J.N.K. Rao. 2014. "On small area estimation under a sub-area level model." *Journal of Multivariate Analysis* 127: 36–55. DOI: https://doi.org/10.1016/j.jmva.2014.02.001.

USDA FSA. 2019. "United States Department of Agriculture Farm Service Agency: ARC/PLC Program." Available at: https://www.fsa.usda.gov/programs-and-services/arcplc_program/index (accessed September 2019).

USDA NASS. 2016a. "CropScape and Cropland Data Layer." Available at: https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php (accessed September 2019).

USDA NASS. 2016b. "QuickStats." Available at: https://quickstats.nass.usda.gov/ (accessed September 2019).

USDA NASS APS. 2018. "Crops/Stocks Agricultural Survey." Available at: https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Crops_Stocks/index.php (accessed September 2019).

USDA NASS CAPS. 2018. "County Agricultural Production." Available at: https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/County_Agricultural_Production/index.php (accessed September 2019).

USDA NASS CDL. 2018. "CropScape and Cropland Data Layers – FAQs." Available at: https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php (accessed September 2019).

USDA RMA. 2019. "United States Department of Agriculture Risk Management Agency: FCIC." Available at: https://www.rma.usda.gov/Federal-Crop-Insurance-Corporation (accessed September 2019).

Received October 2018
Revised May 2019
Accepted September 2019

sciendo

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 89–115, http://dx.doi.org/10.2478/JOS-2020-0005

A Probabilistic Procedure for Anonymisation, for Assessing the Risk of Re-identification and for the Analysis of Perturbed Data Sets

Harvey Goldstein[1] *and Natalie Shlomo*[2]

The requirement to anonymise data sets that are to be released for secondary analysis should be balanced by the need to allow their analysis to provide efficient and consistent parameter estimates. The proposal in this article is to integrate the process of anonymisation and data analysis. The first stage uses the addition of random noise with known distributional properties to some or all variables in a released (already pseudonymised) data set, in which the values of some identifying and sensitive variables for data subjects of interest are also available to an external 'attacker' who wishes to identify those data subjects in order to interrogate their records in the data set. The second stage of the analysis consists of specifying the model of interest so that parameter estimation accounts for the added noise. Where the characteristics of the noise are made available to the analyst by the data provider, we propose a new method that allows a valid analysis. This is formally a measurement error model and we describe a Bayesian MCMC algorithm that recovers consistent estimates of the true model parameters. A new method for handling categorical data is presented. The article shows how an appropriate noise distribution can be determined.

Key words: Additive noise; anonymisation; measurement error; record linkage.

1. Introduction

Providers of data sets for research purposes are typically confronted by a tension between making available useful fit-for-purpose data that retains the fine grain with which the data were obtained, and altering the data sufficiently so that, even without obvious identification information such as name, birth date and address location, an 'intruder' or 'attacker' cannot easily obtain the identity of any given data subject within the data set. A review of procedures for anonymisation of such public use data sets is given by Willenborg and De Waal (2001) and Hundepool et al. (2012) and references therein. For microdata, the disclosure risk concerned is the risk of re-identification. This occurs if an attacker has knowledge about a particular data subject in the microdata and can make an identification by linking that information to the microdata through a set of identifying variables. On the basis of a re-identification, attribute disclosure follows, where

[1] Graduate School of Education, University of Bristol, Bristol, BS8 1JA, UK. Email: h.goldstein@bristol.ac.uk
[2] University of Manchester, Social Statistics, Humanities Bridgeford Street Manchester, M13 9PL, UK. Email: natalie.shlomo@manchester.ac.uk

Acknowledgments: The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for its hospitality during the record linkage programme supported by EPSRC grant EP/K032208/1, where useful feedback was obtained.

information can be learnt about the data subject from the remaining variables in the microdata. To protect against these types of disclosures, we can either protect the identifying variables in the microdata, thus avoiding the risk of re-identification, or protect the sensitive variables, so that even if a re-identification occurs, no new information can be learnt. Often, a combination of both approaches is used to protect the microdata against disclosures. The identifying variables in microdata from social surveys are typically categorical, for example age, sex, place of residence. The identifying variables in microdata from business surveys can be both categorical and continuous, for example industry code and the number of employees.

In this article, we develop an approach that allows us to quantify the disclosure risk when data are perturbed, while at the same time it allows data analysts to fit models that respect the fine grain of the original data. The general idea is to use the addition of random noise to some or all variables in a data set. in which the values of those variables for individuals of interest are also available to an external attacker who wishes to identify those individuals so that they can interrogate their full records in the data set. The idea is that this avoids identification by an attacker via the linking of patterns based on the values of such variables. The noise addition can be carried out on the categorical and continuous identifying variables. Noise generated this way can be removed at the analysis stage if its characteristics are known, and requires disclosure by the statistical agency of the distribution parameters generating the noise. This leads to consistent model parameter estimates, although it will entail a loss of efficiency. In contrast, the usual method of anonymisation, by coarsening the data such as grouping or k-anonymity, would not allow the retrieval of the required model parameter estimates.

There have been many articles that discuss additive random noise as a disclosure control method, see, for example Tendick (1991), Duncan and Mukherjee (2000) and Brand (2002), and references therein. The basic concept is also discussed in some detail by Fuller (1993) and by Winkler (1998). Fuller (1993) points out that the optimum approach is one in which the random noise is added independently for each variable in the data set, and we shall also make this assumption. He treats the case of normal measurement errors and true data that have a multivariate normal distribution (also used as an approximation for discrete data). His method of constructing the perturbed data is designed to provide almost unbiased inferences for linear models and he discusses some of the difficulties for non-linear and non-additive models. For a risk measure, Fuller (1993) derives a probability that any given record in the released data can be identified as the 'correct' one based upon a subset of the variable values that are known to the attacker. In this present article, we propose a similar approach and define a new metric, the h-index, based on the rank distance between the released data and the subset of the variable values known to the attacker.

Another approach for additive random noise is to add correlated random noise (Kim 1986; Little 1993; Ting et al. 2008; Shlomo 2010). Here, the noise that is added is a linear function of the variables to be perturbed. This preserves sufficient statistics in the form of means and covariance matrices, without requiring knowledge of the precise parameter values used to generate the noise, in contrast to Fuller's procedure. The main drawback is that it is restricted to models that can be fitted using sufficient statistics, such as linear regression, and thus excludes, for example, generalised linear models and multilevel models. It also does not allow diagnostics based on residuals, since the production of residual estimates requires knowledge of the noise parameters. Our approach allows for multilevel, generalised linear and non-linear models to be fitted to noise-added data within a more general measurement error modelling framework than that of Fuller (2006), and requires knowledge of the parameters used to generate the noise.

Releasing parameters used to generate the noise is common practice in cryptography literature in computer science in order to be able to decode encryptions, although it is rarely done by statistical agencies. Cox et al. (2011) discuss the need for transparency, whereby a statistical agency releases information about the disclosure control processes used to transform the original data to the masked released data. They distinguish between legitimate users and attackers and advocate controlled release of the parameters used to generate the noise so that legitimate users can carry out statistical inferences. Hence, we will assume here that the parameters for generating noise are known, either released to the data analyst or are in the public domain as is the case for the computer science additive noise approach of Differential Privacy (Dwork 2006). In practice, the noise parameters may also be known to an attacker and in Section 3 we demonstrate that this information does not lead to increased disclosure, thus supporting the argument for their release. It has long been recognized by statistical agencies and data custodians that there is always a trade-off between reducing disclosure risk through statistical disclosure control methods and preserving the analytical properties of the data (Winkler 1998). However, if stochastic perturbation methods are used to anonymise the data, then, as we demonstrate, the statistical analysis is able to account for both the measurement errors and the substantive model of interest. The greater the degree of noise that is added, the lower the statistical efficiency in terms of larger confidence interval estimates for parameters.

Section 2 discusses the technique of anonymisation by adding random noise, thus inducing measurement errors into statistical models, and how the disclosure risk can be quantified through the h-index. Section 3 presents a simulation study on the disclosure risk assessment under the proposed approach. Section 4 describes how the anonymisation by adding random noise can be applied to categorical variables. Section 5 discusses the measurement error model used to retrieve the fine grain of the data to yield consistent parameter estimates, with a more detailed description in the Appendix (Section 10), where the proposed approach is shown to apply generally to complex models including non-linear and multilevel models. Section 6 provides a simulation to demonstrate the measurement error model and Section 7 presents example analyses. Section 8 contrasts the techniques developed in this article with other common approaches of anonymisation, disclosure risk assessment and statistical analysis. Section 9 closes with a discussion.

2. Additive Random Noise

Consider a subset of q variables, y, which can include both identifying and sensitive variables in the data, that are to undergo a statistical disclosure control method. We may also have other variables, say x, that are available to the data analyst, but that may not undergo a statistical disclosure control method. In an extreme case, such variables may not exist, so that effectively the subset q is the complete set of available variables.

We deal first with the case of a set of continuously distributed variables assumed to be multivariate normal (MVN). We introduce random noise m to the subset q of variables in the observed data and, for simplicity, we shall consider the special case in which these variables are independent. We have

$$z = y + m,\, y \sim MVN(\mu, \Omega),\ \ m \sim MVN(0, \Omega_m),\ \ z \sim MVN(\mu, \Omega + \sigma_m^2 I),\ \ \Omega_m\ diagonal.$$

The value of Ω_m will determine the strength of the resistance to attack. In the appendix, we will introduce a more general notation. We note that the assumption of normality for the random noise (measurement errors) is arbitrary, but it is convenient since it allows us to make use of standard results based on multivariate normality.

The attacker scenario is based on the following assumptions:

1. the attacker has knowledge of the perturbation technique of additive random noise being used;
2. the attacker has a copy of the perturbed data set and has a set of the true values for a data subject known to be in the data set. Without loss of generality, we assume that the attacker has the full set of q variables, denoted as y^*, that they intend to match against records in the data set in order to identify a record that matches exactly the data they possess. If this can be done, then the attacker will be able to access any remaining variable values that have not been perturbed associated with the identified record;
3. the original values y are themselves measured without error, although our procedure can be extended to that case straightforwardly; and
4. in the worst case scenario, for each set of y^* belonging to the attacker, there does correspond a single record in the data set and that the original values before perturbation are the true values. From the attacker's perspective, the best case scenario is one in which all their y^* variables also have their true values, and we shall assume that this is the case in our simulations and substantive example. In other words, we ignore the case in which naturally occurring measurement errors are present.

The variables of concern are referred to as identifiers. In fact, these may comprise all released variables, but the data provider may consider that some variables have little disclosive potential and chooses not to perturb them or to do so only with small amounts of noise. In deriving the following formulae, we may treat all variables in the same way. Some may have noise distributions with changing variance, or even zero variance.

We now form a measure of the distance between the attacker's data y^* and all possible values of z in each record in the perturbed data set and rank these distances. We shall then briefly consider how an attacker might be able to improve their chance of detecting the desired record.

A general distance measure can be written in the form

$$D^* \propto (z - y^*)^T W (z - y^*) \tag{1}$$

where in the case of independence, we have the Euclidean distance for each record i in the dataset of size n as follows:

$$D_i^* = \sum_{j=1}^{q} \left(z_{ij} - y_j^*\right)^2, \quad D_i = \sum_{j=1}^{q} \left(y_{ij} - y_j^*\right)^2, \quad i = 1, \ldots, n \tag{2}$$

For a given attacker record we form $r_i^* = rank\left(D_i^*\right)$, and $r_i = rank(D_i)$, and let i^* be the value of i for $r_i^* = 1$, that is the closest record for the attacker in terms of the distance measure. We define $h = r_{i^*} - 1$ which is the difference in ranks between the record identified by the attacker and the rank of the actual closest record, and we refer to it as the h-rank disclosure index for y^*, or simply as the h-index. Thus if $h = 0$ we have the correct match. We note that it is convenient to define our h-index in terms of ranks. It may also be useful to consider the distance measures themselves since these may not, in fact, vary much among the lowest ranked records. However, such information would be somewhat disclosive if available to an attacker and hence not suitable for public release.

We now consider each record, with its true values, in the data set in turn, treating this as y^* and forming h to give a distribution of values of h across the sample. We define $E(h)$ with respect to this distribution, namely the mean value over the sample. Therefore, we need to determine σ_m^2 such that $E(h)$ is large enough (for example taking the value of 3, which coincides with the minimum threshold rule of 3 often used at statistical agencies) to create sufficient unreliability in determining the correct record, thus making the attack not worthwhile for an attacker. In practice, it may be more useful to require

$$pr\left(h < p\right) < \epsilon$$

where a suitable choice might be, say, $p = 3, \epsilon = 0.1$. The range of parameters will ultimately be determined by a policy decision within the statistical agency. In our simulations and examples, we will study the full distribution of h and in particular $pr(\mathrm{h} = 0)$.

The distribution of h will, in general, be a function of y^*. For example, if y^* is a multivariate 'outlier' the attacker is more likely to find the correct match. In the simulation presented in Section 3, we will examine the performance of the procedure with respect to values of percentiles of D. We will not consider the case where we have missing data, except to note that this will contribute to the unreliability of the matching process. In fact, data values may be missing in any given data set, so that our computations in that respect represent a 'best' case scenario.

In principle, an attacker who has access to the noise parameters may be able to utilise this to improve their attack strategy by making use of this information.

Thus, with knowledge of Ω_m rather than utilising z, the attacker could obtain more precision since they would be able to estimate

$$z^* = E\left(y|z\right) = (cov(z))^{-1} cov\left(y\right) \times z \tag{3}$$

The attacker can calculate $cov(z)$ from the set of perturbed data that they possess and can obtain an estimate of $cov\left(y\right)$ by subtracting Ω_m.

Thus, in the case of independent random noise for normal variables, a simple procedure would be to use $z_j^* = z_j R_j$ for variable j, where $R_j = \sigma_y^2 / \left(\sigma_m^2 + \sigma_y^2\right)$ is the 'reliability' of the observed variable. We will investigate any advantage to the attacker that the use of (3) might have in the simulation presented in Section 3.

In summary, the best the attacker can do is to find that set of perturbed records that is closest to their own (unperturbed) record that they wish to match. Hence, the use of the h-index ascertains whether the attacker will be able to identify the correct matching record. The value of h needs to be computed by the data provider. Even if the data analyst and the attacker were the same person, the knowledge of the noise parameters and tools available to the analyst would be of no advantage in their role as attacker, as will be shown in Section 3. The attacker may also, in general, have access to h-values since these may be available in the public domain. Indeed, knowledge of these may often be sufficient to deter a potential attacker from undertaking the attack.

3. A Simulation for Quantifying Disclosure Risk

The purpose of this simulation is to investigate how disclosure risk varies with how far the attacker's record is from the centroid of the distribution of true values across all records, and how any additional information about the joint distribution of variables might be useful to the attacker.

We generate a series of 1,000 simulated data sets, each with 1,000 records, and each being generated with a mean vector of zero, $q = 5$ and $\sigma_m^2 = 0.1$ and Ω has all variances $= 1$ and covariances $= 0.25$. This value of q is chosen since it will typically represent the number of variables that may be available to an attacker under disclosure risk scenarios (Elliot and Dale 1999), but we have varied this number below.

For each true value record known to an attacker, we generate D_i as in (2). We choose 9 true value records representing approximately deciles of the distribution of D_i, to define suitable attacker records y^*. These records are then used to compute the distribution for the h-index as described in Section 2. The distribution of the h-index varies by decile with greater precision of attack at extreme values and we show results for different deciles. The choice of standardised variates simplifies the computations somewhat.

Routines are written in MATLAB. Based on 1,000 simulations we obtain the following results in Table 1 for the cumulative distributions.

From the viewpoint of the attacker, in more than 43% of the cases for any of these deciles and more than 58% in the middle of the distribution, the nearest record is not the true one. For the 10th decile $pr(h > 3) = 0.25$, $pr(h < 2) = 0.63$ and $pr(h > 5) = 0.19$. For the median an attacker has a harder time with $pr(h > 5) = 0.29$. Depending, of course, on the degree of disclosure risk that can be tolerated, it could be argued that this is adequate to make an attack too unreliable to be worthwhile, and this would be a matter for careful consideration by the data provider.

For the weighted distance case in (1) where $W = \Omega$, we obtain essentially similar results. We have also run the simulation with larger sample sizes and we find that at the lowest decile for a greater sample size of 10,000, $pr(h > 3) = 0.60$ and $pr(h > 5) = 0.54$.

We now look at a range of values for Ω and σ_m^2 and different sample sizes. We study just the case for the lowest decile, as defined above, in which we believe that an individual with more extreme values presents a target that is more favourable to an attacker. We present results for $pr(h > 5)$ and for two different sample sizes in Table 2.

Table 1. Cumulative percentage distribution for h for deciles of the distribution of D_i.

	Decile								
h	10	20	30	40	50	60	70	80	90
0	52.2	49.4	43.9	41.3	41.7	43.5	40.5	47.0	56.2
1	62.9	60.7	56.1	53.1	53.1	54.3	53.0	58.3	67.3
2	70.0	65.3	62.0	61.2	60.8	61.8	60.9	64.7	73.8
3	74.7	70.2	68.6	66.0	65.8	66.6	65.8	70.3	77.9
4	78.5	74.4	72.8	70.1	68.7	70.0	69.6	73.8	81.8
5	80.8	77.5	76.5	72.7	71.5	72.1	72.8	76.8	84.3
6	83.1	79.9	78.1	74.3	74.1	74.8	75.5	78.7	86.2
7	84.4	82.4	80.6	76.5	76.1	76.2	78.0	81.2	87.3
8	85.9	83.9	81.7	78.7	78.2	77.8	80.0	83.9	88.9
9	87.7	84.9	83.5	79.5	79.4	80.2	81.8	84.9	90.7
10	89.1	86.4	85.0	81.1	81.2	81.8	83.2	86.6	91.9
11	90.1	87.1	85.4	82.4	83.5	83.2	84.1	87.4	92.9
12	91.1	88.3	86.1	83.8	84.6	84.7	85.3	88.2	93.9
13	91.9	88.8	87.1	85.3	85.6	85.8	86.6	89.1	94.3
14	93.0	89.3	87.8	86.5	86.1	86.6	87.5	90.0	95.0
15	94.2	90.1	88.1	87.4	86.6	87.4	88.0	90.2	95.1
16	94.9	91.1	88.8	88.0	87.5	87.9	88.5	90.9	95.1
17	95.2	91.7	89.3	88.8	88.5	88.5	89.5	91.4	95.6
18	95.7	91.9	89.8	90.0	89.6	89.1	90.1	91.8	95.9
19	96.5	92.6	90.3	90.9	90.0	89.9	90.5	92.4	96.6
20+	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

We see that even with the smaller sample size of 1,000 and moderate proportions of noise (10% of the variance of the true values), we have reasonably high probabilities of *h* exceeding a value of 5 and small probabilities that the nearest record is the correct one.

We now study what effect the use of (3) has, that is, when the attacker makes use of information about the parameters of the noise distribution.

Table 3 shows, for $h = 0, 1, \ldots, 5$, and for the 10th to the 50th percentiles, the percentage distributions of *h* when knowledge of the noise parameters is used as described in (3) under the same simulation conditions as Table 1. In other words, rather than working with the observed (perturbed) values *z*, the attacker uses z^* as defined in (3). Comparing with Table 1, we generally see small increases in the cumulative probability that an attacker selects a record close to the correct one.

We would expect that the probability of the attacker selecting a record close to the correct one will increase with the number of distinct identifiers used. Thus, for example, if there are 10 identifiers and we simulate with the same value for Ω as before, we now need a noise parameter $\Omega_m = 0.34$ rather than $\Omega_m = 0.1$ to obtain approximately the same values for the distribution of *h*. This suggests that careful consideration needs to be given to the likely number of identifiers available to the attacker. We have also varied the size of the covariances from 0.1 to 0.5. However, this only has a small effect on the distribution of *h*, with a decrease in the covariance associated with a slightly higher risk of disclosure. For example, for the 5th percentile the $pr(h = 0)$ is 0.47 for a covariance of 0.5, as opposed to 0.53 with a covariance of 0.1 as in Table 1.

Table 2. Lowest decile estimates for h.

pr(h $>$ 5) for combinations of Ω *and* σ_m^2 where Ω always has unit diagonal elements and equal off-diagonal elements (given by columns) are shown.					
Sample size = 1000					
σ_m^2.	0.1	0.2	0.3	0.4	0.5
0.1	0.15	0.16	0.19	0.23	0.24
0.2	0.45	0.43	0.46	0.50	0.54
0.3	0.58	0.63	0.63	0.65	0.70
0.4	0.73	0.74	0.74	0.76	0.77
Sample size = 5,000					
σ_m^2.	0.1	0.2	0.3	0.4	0.5
0.1	0.41	0.48	0.48	0.50	0.55
0.2	0.72	0.71	0.75	0.77	0.80
0.3	0.84	0.84	0.87	0.88	0.89
0.4	0.90	0.90	0.92	0.90	0.93
pr(h = 0). For combinations of Ω *and* σ_m^2 where Ω always has unit diagonal elements and equal off-diagonal elements (given by columns) are shown.					
Sample size = 1,000					
σ_m^2.	0.1	0.2	0.3	0.4	0.5
0.1	0.56	0.54	0.53	0.49	0.45
0.2	0.27	0.27	0.24	0.23	0.18
0.3	0.16	0.13	0.15	0.12	0.09
0.4	0.10	0.09	0.10	0.08	0.07

4. Treatment of Categorical Variables

Consider for simplicity, a series of q independent binary identifying variables. We assume that one of the categories is small, for example, $\pi = pr\big(y = 1\big) = 0.1$. Thus, if $q = 3$, with $\pi_i = 0.1$, $i = 1, . . ., 3$ then the most favourable vector for an attacker y^*, is $y_i = 1 \ \forall i$.

Suppose now we introduce a simple misclassification, where for each $i|y_i = 0$ independently, we randomly assign $y_i = 1$ with probability 0.1. Thus, for each binary variable we now have $pr(z_i = 1) = 0.19$. The probability that all three variables have value 1, that is, $pr(z_i = 1), \ \forall i = \ 0.19^3 = 0.007$, whereas $pr\big(y_i = 1\big), \ \forall i = \ 0.1^3 = 0.001$.

Table 3. Cumulative percentage distribution of h for deciles of the distribution of D_i where noise parameters are known and expected values of identifiers with noise are used.

	Decile				
h	10	20	30	40	50
0	54.7	48.7	44.3	41.3	45.4
1	67.1	61.8	54.5	55.6	57.0
2	73.1	68.2	62.8	64.1	63.7
3	78.6	72.4	66.7	69.3	68.4
4	81.9	75.3	70.6	73.1	72.4
5	84.5	79.3	73.7	75.4	74.9

Thus, of those identified only 15% are correctly identified. Such procedures for categorical variables have been implemented in the PRAM method (Gouweleeuw et al. 1998).

For multicategory, including ordered and unordered variables, we can alternatively consider the following novel procedure. Coding the categories $j = \{1, \ldots, p\}$ we independently add to the true category code, random noise as follows:

$$m_j \sim N\left(0, \sigma_m^2\right),\ 1 \leq m_j \leq p \tag{4}$$

This results in a truncated normal distribution and the variable enters the calculation of the distance (2) in the same way as the continuous variables. The purpose of the truncation is to avoid easy detection for the extreme category codes. We note that for this procedure, we may still use the h-index as a distance measure based on the observed values with added noise. The attacker is interested in ascertaining the correct code for a categorical variable so that it will be appropriate to utilise a measure based on the distance between the observed value assigned to a categorical variable and its true code.

We could also generalise (4) to allow different variances for each category. We note that the noise is simply added to the category codes $\{1, \ldots, p\}$, irrespective of whether this is an ordered or unordered variable. The MCMC steps described in Section 5 and in the appendix show how we can then draw from the posterior distribution of the true (unknown) category codes. To avoid the potential objection that it may be confusing to release categorical variables (with added noise) as continuously distributed, we also describe in the appendix how a rounding of the noise-added values to integer values can also be used, although this will result in less efficient estimates. The advantage of this procedure is that it is completely general and has the simplicity that we can deal with all variables in the model estimation, whether continuous or categorical, in a similar fashion.

5. Fitting Models with Known Noise or Measurement Error

In common with other approaches (Fuller 2006), we adopt a model-based approach for fitting data with measurement error, and more specifically, we adopt a Bayesian approach. Bayesian procedures for fitting models with measurement errors have been proposed early in the literature (see Richardson and Gilks 1993) and these have been further developed to fit multilevel data structures and allow models that include interaction and power terms. The noise that is added in our procedure has the characteristics of measurement error and can be treated as such. An outline of a general algorithm with details specific to data anonymisation is given in the appendix, and further details of the estimation algorithm can be found in Goldstein et al. (2017). Other methods for estimation of models with measurement errors are also available, such as the simulation-extrapolation method, SIMEX (Delaigle and Hall 2008) and moment-based estimators described by Fuller (2006), and these can also be used. The Bayesian model procedure that we describe has the advantage that it is a fully specified probabilistic model that is readily generalised to handle complex data structures, including multilevel and generalised linear models without approximations, with straightforward computation of interval estimates. We assume uniform priors for all the model parameters to be estimated.

Full estimation details are given in the appendix, and can be summarised as follows. For ease of exposition, we assume a single level linear model with just a single predictor variable that has added noise. The case of several predictors with independent added noise, the case where we have a generalised linear model, and the multilevel case follow straightforwardly. Here, we assume multivariate normality for the noise and discuss the modifications needed to fit generalised linear models and multilevel models in the appendix.

We define the true values of the variable with added noise as X_1 and those variables without added noise as X_2, and $X = [X_1\, X_2]$ where X_2 is known. We note that our procedure is fully general and includes the case where X_2 is null and all variables are perturbed.

We define the joint model – the noise or measurement error model (MEM) in two parts, (5a) and (5b) and the model of interest (MOI) (5c):

$$x_1 = X_1 + \gamma_1 \tag{5a}$$

$$X_1 = X_2^T \alpha + \gamma_2 \tag{5b}$$

$$Y = X\beta + e \tag{5c}$$

where $\gamma_1 \sim N\left(0, \sigma_{\gamma_1}^2\right)$, $\gamma_2 \sim N\left(0, \sigma_{\gamma_2}^2\right)$, $e \sim N\left(0, \sigma_e^2\right)$.

Thus, (5a) defines a simple additive measurement error model, in which the observed value consists of an independent random variable added to the unknown true value. Equation (5b) expresses the relationship between the true values for those variables measured with error and those measured without error, here assumed to be linear, and (5c) is the substantive model of interest expressed in terms of true values. This is a standard formulation and further details can be found in Goldstein et al. (2017). Instead of additive measurement error as in (5a) we could develop our procedure in terms of multiplicative measurement error (Hwang 1986), for example writing $x_1 = X_1 e^{\gamma_1}$. This is further discussed by Goldstein et al. (2017) and may be useful for certain non-normally distributed variables. We note that the MOI (5c) may contain functions of the X_1, such as interaction or power terms. Lower case variables define observed and upper case true values, and we assume the residual terms in (5a) – (5c) are independent.

The appendix details the MCMC steps required to fit this model. In brief, this involves the following steps:

1) update the true values using a Metropolis step, conditionally on the current values of the other parameters;
2) update the α parameters using a Gibbs step, conditionally on current values of the other parameters;
3) update the β parameters using a Gibbs step, conditionally on current values of the other parameters; and
4) update the variance and covariance parameters, conditionally on current values of the other parameters.

In the appendix, we also discuss the following extensions and the difficulties associated with them. Where we have added noise in the response variable, the general effect of correcting these is to shrink the estimate of the residual variance and hence inflate the standard errors for the parameters. Where the response is categorical, notably binary, a further step in the algorithm is involved. In the appendix, we also discuss how to deal with variables that have been treated with other SDC techniques, for example truncation to handle large outliers, and how to handle perturbed categorical variables that are rounded to the nearest integer value.

Where the original variables are also subject to other measurement errors with known distributions, the fitted model will be based on the total measurement error. We also note that our proposed approach could be adapted to the case of sampling from a finite population without assuming a distribution for *y*, which would require modification of the estimation of the parameters of the underlying ‘true’ model. However, we do not pursue this and leave it for future work.

After fitting a suitable model as described above, the original variable scales and relationships are fully recovered, albeit with a loss of efficiency. With very large data sets, this may not be an important issue. The loss of efficiency, in terms of interval estimates or standard errors associated with parameter estimates, can be estimated for any proposed model to be fitted to the perturbed data, given the noise parameters. Thus, for example, in the simple regression case where independent normal noise with a common variance σ_m^2 has been added to the set of predictors *X*, we can use, as a simple overall measure for the relative efficiency, the determinantal ratio $\left(\left|X^TX\right|/\left|X^TX+n\sigma_m^2\right|\right)$ where *X* includes those with and those without measurement errors and *n* is the sample size. The data provider would be able to supply such estimates. However, perhaps of more use will be estimates of the inflation of standard errors, for some typical models, associated with individual parameters. These could be provided alongside the released data, or possibly requested from the data provider by a data analyst, with respect to any given fitted model. We illustrate the effects on standard errors in our example analysis in Section 7.

6. A Simulation of a Measurement Error Model

We carry out a simple simulation from the following model in order to illustrate that we can readily recover the signal from noisy data for both a binary and continuous predictor. The model we simulate from is given by

$$y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+e_i,\qquad e_i\sim N(0,1),\quad \beta_0=\beta_1=\beta_2=1 \tag{6}$$

$$\begin{pmatrix}x_1\\ x_2^*\end{pmatrix}\sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}\right),\quad x_2=0\,if\,x_2^*\le 0;\quad x_2=1\,if\,x_2^*>0.$$

Independent noise with variance $\sigma_m^2=0.2$ is added to x_1, x_2. We carry out two sets of 100 simulations with 1,000 records. The results are given in Table 4.

We see negligible bias, no more than 0.5% for the adjusted estimates and in all cases the 95% confidence interval overlaps the true value.

Table 4. Estimates from addition of noise (standard errors in brackets). MCMC burn in = 500 iterations = 500. Simulations from model (6) sample size = 1,000, number of simulations = 100.*

Parameter	Simulation parameters	Noisy data no adjustment	Noisy data adjusted	Noisy data adjusted % bias
β_o	1.0	0.974 (0.004)	0.997 (0.003)	−0.3
β_1	1.0	0.887 (0.002)	1.004 (0.003)	0.4
β_2	1.0	1.051 (0.005)	1.003 (0.005)	0.3
σ_e^2	1.0	1.0	1.0	0

*The standard error estimates are the standard deviations computed from the MCMC chains.

7. Example Analyses

Our first example illustrates the use of a measurement error model for two-level data, where we have added noise and used the procedures in the appendix to estimate the true model parameters. The data will be referred to as the Tutorial data set (Goldstein et al. 1993). The response is a normalised examination score taken at age 16 by 4,059 students in 65 schools in Inner London. The predictor variables are a standardised reading test score taken at age 11 (x_1) before pupils attended their secondary school, and the binary variable gender (x_2). The true model of interest is a 2-level random intercept model as follows:

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + u_j + e_{ij}, \quad u_j \sim N\left(0, \sigma_u^2\right), \quad e_{ij} \sim N\left(0, \sigma_e^2\right) \tag{7}$$

The subscript i indexes students and j schools. This is thus a simple 'random intercept' model (Goldstein 2011). To illustrate our procedure, we first perturb the true variable values by adding normally distributed noise with mean 0 and variance $\sigma_m^2 = 0.2$ independently to the reading score and gender. For gender, perturbed values less than 0 are set to 0 and values greater than 1 are set to 1. The purpose of this first example is to show how adjusting for the added noise produces consistent estimates rather than to explore a range of values against disclosure risks.

Table 5 shows the results from fitting the model before adding noise, fitting the model with added noise but without adjusting for measurement error, and fitting the model adjusting for measurement error. We have used the known parameters of the noise addition to the two covariates in the analysis and the response variable is not perturbed. We implement the Bayesian estimation procedures described in the appendix with uniform priors for the parameters.

Table 5. Tutorial data set. Estimates from addition of noise (standard errors in brackets). MCMC burn in = 500 iterations = 500. Standard errors in brackets using MCMC chain standard deviation estimates. Normal noise variance 0.2.

Parameter	Original model	Noisy data no adjustment	Noisy data adjusted
β_o	−0.097 (0.050)	−0.082 (0.044)	−0.093 (0.042)
β_1	0.559 (0.013)	0.460 (0.012)	0.549 (0.014)
β_2	0.174 (0.033)	0.109 (0.030)	0.155 (0.035)
σ_u^2	0.097 (0.021)	0.102 (0.022)	0.100 (0.020)
σ_e^2	0.563 (0.013)	0.614 (0.013)	0.562 (0.014)

We note that the adjusted estimates are close to those using the original data, whereas ignoring the measurement error produces estimates with considerable biases.

Our second example uses a data set from a 1982 survey of the sugar cane farm industry in Queensland, Australia that was used in Chambers and Dunstan (1986). It illustrates both the measurement error model and the computation of the h-index for disclosiveness, and suggests how these can be integrated into the release of such data. In order to compare our approach for compensating for the measurement error in our analyses with the case of adding correlated random noise that preserves sufficient statistics, we use a linear regression model. We note that under more complex models, such as generalised linear models and multilevel models, the correlated additive random noise approach would not provide valid results.

The model of interest has the sugar cane yield receipt as response (y) and predictors are region (x_1, Northern $= 1$, Southern $= 0$), sugar cane harvest (x_2, continuous in tonnes) and cost (x_3, in Australian dollars). There are 333 farms in the data set with no missing data.

The model to be fitted is the linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i \quad e_i \sim N\left(0, \sigma_e^2\right) \tag{8}$$

We have added noise to the data in two different ways:

- Correlated Gaussian noise as described in Shlomo and De Waal (2008) and summarised here:

  Generate multivariate random noise: $\left(\varepsilon_1, \ldots, \varepsilon_q\right) \sim MVN\left(\mu', \Sigma\right)$ where $\mu' = \left(\frac{(1-d1)}{d2}\mu_{1,\ldots,}\frac{(1-d1)}{d2}\mu_q\right)$, $d_1 = \sqrt{\left(1-\delta^2\right)}$ and $d_2 = \sqrt{\delta^2}$ and $\delta \in [0, 1]$ is the perturbation parameter determined by the statistical agency. The vector μ and matrix Σ are the original mean and covariance matrix of the data. Then, for each separate variable, calculate the linear combination as follows: $z_j = d_1 y_j + d_2 \varepsilon_j$, $j = (1, \ldots, q)$;
- Independent Gaussian noise is added to each variable as described in Section 2.

In both cases, we add two levels of noise: the variance of the noise distribution is 0.05 and 0.17 times the variance of the corresponding variance of the true values for each variable.

First, we present the results of fitting (8) to each of the noisy data sets in Tables 6a and 6b. Then, we present the results of computations on disclosiveness in Tables 7a and 7b. We note that for correlated noise, the model is simply fitted to the observed data after a single draw and the reported standard errors are those obtained analytically from the model fit. For independent additive noise, we use the procedures described in the appendix where reported standard errors are obtained empirically using the standard deviations resulting from the MCMC chains. Table 6a is based on the case where noise is added to the predictors only, and Table 6b is based on the case where noise is added to both the predictors and the response variable.

In Table 6a, we see the negative effect of releasing a draw of correlated noise to users without providing the parameters of the noise distribution when only the predictor variables intended for the statistical modelling have added correlated random noise. Users will obtain biased estimates with very large standard errors. This is not the case in

Table 6a. *Sugar cane farm data. Results of fitting (8) to 5% and 17% relative variance noisy data. Standard errors in brackets. Noise added to predictors only.*

Parameter	True data with no added noise	Correlated noise with 5% of variance added	Independent noise with 5% of variance added	Correlated noise with 17% of variance added	Independent noise with 17% of variance added
β_o	−7.89 (1.70)	−6.01 (2.25)	−5.71 (1.70)	−3.56 (3.35)	−9.16 (2.27)
β_1	11.35 (1.37)	10.13 (1.83)	10.03 (1.38)	14.20 (2.68)	13.05 (1.83)
β_2	0.022 (0.0007)	0.022 (0.0009)	0.022 (0.0007)	0.021 (0.0014)	0.021 (0.0011)
β_3	0.00013 (0.00004)	0.00008 (0.0006)	0.00010 (0.00004)	0.00015 (0.00008)	0.00022 (0.000070)
σ^2	146.2 (11.23)	254.02	150.1 (11.63)	530.2	170.8 (17.86)

Table 6b. *Sugar cane farm data. Results of fitting (8) to 5% and 17% relative variance noisy data. Standard errors in brackets. Noise added to predictors and response.*

Parameter	True data with no added noise	Correlated noise with 5% of variance added	Independent noise with 5% of variance added	Correlated noise with 17% of variance added	Independent noise with 17% of variance added
β_o	−7.89 (1.70)	−7.33 (1.71)	−8.95 (2.88)	−7.56 (1.81)	−7.04 (4.39)
β_1	11.35 (1.37)	11.82 (1.41)	11.91 (2.26)	11.36 (1.43)	11.38 (3.64)
β_2	0.022 (0.0007)	0.022 (0.0007)	0.021 (0.0014)	0.022 (0.0007)	0.022 (0.0031)
β_3	0.00013 (0.00004)	0.00011 (0.0004)	0.00023 (0.000086)	0.00014 (0.0004)	0.00016 (0.00019)
σ^2	146.2 (11.23)	148.01	180.7 (28.32)	150.46	179.8 (73.3)

Table 7a. Sugar cane data. Values of h-index for records at different multivariate distance percentiles. Random correlated noise addition on all variables as percentage of true variances.

Distance percentile	*h*-index 5% of true variance	*h*-index 17% of true variance	*pr*(h = 0) 5% of true variance	*pr*(h = 0) 17% of true variance
5	11.2	23.2	0.11	0.04
10	17.2	29.8	0.05	0.05
50	17.5	37.9	0.08	0.04
90	2.7	10.7	0.43	0.17
95	0.0	0.5	0.98	0.78

Table 6b, in which all variables, including the response variable, have correlated noise added to them. In that case, we obtain similar parameter estimates and standard errors to the model run on the original true data. Given that it is generally unknown what types of analysis will be carried out on the released perturbed data, it is essential that users obtain the noise parameters and be able to analyse the data under measurement errors using the procedures described in the appendix. We see in Tables 6a and 6b that under our proposed procedures we obtain unbiased parameter estimates, taking into account the measurement error regardless of which variables have been perturbed. We also see some considerably increased standard errors as greater amounts of noise are added. For example, under the 17% option for the variance, the variance of the noise for the response variable was set at 411, which is nearly three times the true residual variance as estimated in Tables 6a and 6b. This governs the size of the standard errors. Even when the variances are chosen at the 5% option, we still see an increase in the standard errors.

Table 7a presents the values of the *h*-index for individual records chosen to represent both the centre and extremes of the data distribution for different amounts of random noise on the correlated noise addition on all four variables (predictors and response variable): receipts, region, harvest, and costs. Similarly, Table 7b presents the independently added random noise. To avoid skewness in these variables, the distances were calculated on standardised variables, although given the nature of this data, some skewness remains.

From both Tables 7a and 7b, it is clear that in terms of their multivariate distance from the data centroid, the *h*-index values show a high level of disclosure protection for both the 5% and 17% variance options except for the far right tail due to the skewness of the data in the sugar farms data set. In the latter case, the data provider may well decide

Table 7b. Sugar cane data. Values of h-index for records at different multivariate distance percentiles. Random independent noise addition on all variables as percentage of true variances.

Distance percentile	*h*-index 5% of true variance	*h*-index 17% of true variance	*pr*(h = 0) 5% of true variance	*pr*(h = 0) 17% of true variance
5	7.0	16.0	0.12	0.06
10	12.6	23.0	0.07	0.04
50	12.7	27.8	0.08	0.03
90	1.7	6.8	0.38	0.15
95	0.0	0.2	1.00	0.89

to group (truncate) the very large values, as we describe in the discussion below. Comparing correlated random noise addition in Table 7a with the independent random noise addition in Table 7b, we see quite similar results for $pr(h = 0)$, although the value of h overall does tend to be greater for the correlated noise. The 17% variance option offers greater disclosure protection than the 5% variance option, as expected. However, this may still not be satisfactory and the data provider would then have a choice of increasing the amount of added noise, adopting a truncation procedure to reduce the skewness, or making the noise variance a monotonically increasing function of the true value, that is, adding more noise as the true value increases. These options will be pursued in future research.

This example demonstrates the importance of considering both the disclosure aspects and the issue of model parameter bias in the data release in order to provide an appropriate balance between disclosure risk and data utility.

8. Similarity to Other Methods

Standard procedures for anonymisation are typically based on a distinction that is made between primary individual identifiers such as name and birth date, quasi-identifiers such as age, gender and ethnic group, and sensitive identifiers such as disease diagnostic categories that may have disclosive values. The primary identifiers are typically removed from the data. The coarsening technique of k-anonymity and extensions such as l-diversity and t-closeness consider two stages. In the first stage, the values for the quasi-identifiers are coarsened to the extent that in the final data set there are at least k records that have identical identifier values for any given set of identifier values (a k-anonymity data cell). In the second stage, the simplest form of l-diversity requires that for any quasi-identifier there is at least l distinct value for each k-anonymity data cell. For t-closeness, the difference (using a suitable measure) between the distribution of the l values in the data cell and the distribution in the whole sample should not exceed a threshold t. These techniques may also lead to the suppression of certain variable values to ensure the required anonymity level (see, for example, the Statistical Disclosure control package (SdcMicro 2019)). One of the drawbacks of such methods is that they may remove or degrade too much of the identifying data that is of importance to the data analyst.

We make no distinction between identifying or sensitive variables that may be available to an attacker, and all variables could be incorporated in the distance computation in (1) or (2) if needed. Our proposed method is essentially probabilistic rather than one that guarantees a given level of anonymity, as in the standard k-anonymisation methods. The key element, however, is that from the data analyst's viewpoint there is no data coarsening with potentially enhanced quality for the inferences. Our approach is applicable to all statistical models, in particular the family of generalised linear models, for which measurement error methods can be applied to compensate for errors in the data. In addition, established procedures for diagnostics can also be utilised since the estimation, as described in Section 5 and in the appendix, provides the necessary parameters for the model of interest.

Other approaches that focus on estimating the risk of re-identification of a data subject withnoise-added data, are found in Winkler (1998), based on probabilistic record linkage and in Reiter and Mitra (2009), and Shlomo and Skinner (2010) on using probabilistic modelling to estimate a probability of re-identification whilst accounting for the perturbation. Reiter and Mitra (2009) also account for attacker knowledge. These procedures estimate a probability of a correct match. The probabilistic modelling estimates the probability of a correct match based on distributional assumptions of the data, while the probabilistic record linkage estimates the probability of a correct match using a naïve Bayes approach that is similar to the h-index based on a distance metric. Probabilistic modelling relies heavily on assumptions and, at best, we may obtain an unbiased estimate for the total number of high-risk records rather than a risk measure for an individual record. Winkler (1998), in particular, concluded that even moderate amounts of additive noise, where some of the analytical properties of the data are preserved, may still have considerable disclosure risks.

Arima and Polettini (2019) propose a method for inference under perturbed data that has similarities to our own as described in Section 5. They develop it in the context of small area estimation, in which the predictor variables have been masked using the post-randomisation method (PRAM) (Gouweleeuw et al. 1998). They apply a Bayesian algorithm to the data measured at the aggregate small area level, where categorical data are approximated by a multivariate normal distribution and the adequacy of the approximation is a function of the number of individual records within an area. Our own procedure primarily operates at the level of individual records and does not involve such approximations, although it can also be used to deal with data aggregated to higher levels. Woo and Slavkovic (2014) also discuss adjustments to logistic regression with variables subjected to PRAM.

Adding noise is similar in some ways to the "fully synthetic" data approach, in which data is generated from a series of predictive distributions (Rubin 1993; Reiter 2005). This approach relies on the structure of the synthetic data corresponding sufficiently closely to the structure of the real data, so that valid analyses are possible, and thus involves an additional set of considerations (Reiter 2005). Our procedure does not rely at all on the use of synthetic data. Nevertheless, if a partially synthetic data set is released our procedure could be applied in that setting.

9. Discussion

In our simulation in Section 3, we have shown that the disclosure protection provided by additive random noise addition increases with sample size and therefore, disclosure risk will generally be of concern for sample surveys with relatively small sample sizes. For example, we see that with a sample size of only 1,000, in the case most favourable to an attacker that we have explored in our simulation, the probability that the nearest record is the correct record is less than half. If an attacker has access to the noise parameters and utilises this information to obtain an estimate of the covariance matrix for the true values, our simulation suggests that there is only a small increase in the probability of selecting the correct record, and that this is of little practical importance. If the number of identifiers available to the attacker increases, then this can enhance the chance of a successful attack. We have shown that we require rather larger amounts of noise as the number of identifiers available to the attacker increases, so that a realistic assessment is needed on the number of identifiers that may be available to an attacker, and this is an area for further exploration. On the basis of our simulation, it appears that the disclosure risks are relatively insensitive to correlations between the identifiers. However, we would caution that our simulations are limited and based on an assumption of multivariate normality. Further research needs to explore more general cases and we would suggest that one responsibility of a data provider is to provide estimates for disclosure probabilities for their own data based on the distributions observed in the data.

Our results are based on the assumption that the attacker has exact knowledge of who is in the data set and a particular individual's exact data values. In the absence of any other information available to the attacker, this is a worst case scenario and indeed in a sampling context, response knowledge is not assumed known, especially for microdata arising from social surveys. Therefore, disclosure risks may be considerably less than our reported findings. If an attacker has some random data points from the population, they will first have to check if the target data subject is in the data set and that depends on the sampling fraction, which is generally small. For business surveys with large sampling fractions and take-all strata, these assumptions may not be relevant. Often, a data attacker will have no pre-existing individual data and may focus on trawling the data set to discover an 'interesting' record, for example an individual or business with an unusual combination of values. Having identified such a data subject, they may then attempt to identify the real entity in the population using other variables in the data record. Our procedure is also relevant to such an attack so long as the noise has been applied to the variables in question.

For the extremes of the distribution it would be useful for disclosure purposes to apply measurement errors with larger variances. For example, if noise is added to a variable such as income, we might wish to make the variance parameter of the noise distribution a function of income itself; for example by adding noise generated with smaller variances in lower quintiles of the original income variable and then increasing the variance of the noise for upper quintiles. Such a function could be non-linear or a step function, in which all true value greater than a specified (absolute) value have additional noise added. Nevertheless, the release of the information describing such a functional relationship will generally be informative for individual records and so, such information could be disclosed only to the accredited data analyst. In terms of the measurement error algorithm described in Section 5 and in the appendix, the value of the variance would simply need to be updated at each MCMC iteration, together with the estimated current value for the true value of the variable. This is an area of further research to be pursued.

There is an interesting contrast with the *k*-anonymity criterion that is often used as a measure of disclosiveness. If we have 2-anonymity, this implies that an attacker is able to identify two individual records matching their own information, so that choosing either of them at random means that there is a probability of 0.5 that it is the correct one. The *h*-index, however, only yields a single individual with a probability of about 0.5 and, thus, provides less information to the attacker than in the case of 2-anonymity. Indeed, an attacker may be quite content with the information that they can access 2 or perhaps even 5 records containing the one that is sought. By contrast, with the *h*-index procedure, in our most favourable case, the probability of the sought-for individual being one of the two nearest, is just under 70% and one of the five nearest just over 80%. Thus, it could be argued that this is sufficient to deter an attacker and hence suitable in terms of protecting against disclosure. In practice, careful attention needs to be paid to the amount of noise required to satisfy disclosure concerns and this is an area for further research.

The generality of our procedure for compensating for measurement error described in Section 5 and in the appendix is that it makes no assumptions about the final model to be fitted and can be a generalised linear model or a multilevel model with some or all of the variables perturbed, and the procedure allows a full range of exploratory analyses. A particular advantage of our procedure is that, when fitting a model of interest, we will obtain chains of imputed 'true' values that can be used for general model checking and data display purposes: for example the means of these can be plotted to study their (joint) distributions. Likewise, in contrast to previous work, it does not assume any particular distribution for the true values, at least for those used as covariates in the substantive model of interest. Nevertheless, when data are being anonymised for release it is important that both the disclosure risk and the consistency of subsequent parameter estimation are considered. This will principally be the responsibility of the data provider, who will need to balance the requirements of confidentiality with the needs of data analysts. We consider that our proposal provides a suitable framework for doing this, as we illustrate in our second example in Section 7.

Our procedure can be contrasted with procedures based on the production of fully synthetic data simulated from estimates of the structure of the real data where exploratory analyses are recommended prior to the choice of a small number of models to be fitted to the real data within a secure environment. Such procedures not only rely on good estimates of the real data structure, they also rely on exploratory analyses converging on the appropriate set of final models, and this is by no means guaranteed. We note, however, that in some cases where we wish only to fit a linear regression model, a more tailored procedure such as that using correlated noise (Shlomo 2010) may provide superior estimates. It would be possible, in principle, to develop our procedure for those cases in which the correlated noise procedure is used. However, we have not pursued this, since, as we point out, a general procedure that can be used with any subsequent analysis model is generally preferable. With our proposed procedure, exploratory analyses would generally be carried out using the measurement error methods proposed in Section 5.

For a response variable with added noise, the situation is more complex. However, for normally distributed responses, we can carry out an analysis using the observed values to obtain consistent estimates. In this case, we can obtain a consistent estimate for the residual variance by subtracting the (known) variance for the measurement errors in the response from the final estimated residual variance. The drawback, as illustrated in our second example in Section 7 on the sugar cane farms data set, is that where the proportion of variance explained is high, this will lead to large standard errors. In this case, we can either add less noise to any variables that a given data analyst wishes to use as response variables, or no noise at all. In our example, this has little effect on the disclosiveness and therefore, in practice will often be acceptable. For categorical response variables, as with continuous response variables, when data are released it may be acceptable to provide any variable(s) to be treated as a response variable with its true values or with just a small amount of added noise. Where a categorical response variable has added noise, it is rounded to the nearest category value, for example 0 or 1 in the case of binary data, and these values are then used as responses as described in the appendix. Since, in general, different data analysts may wish to treat different variables as response variables, this implies that there should be a close liaison between the data provider and the analyst so that appropriate amounts of noise can be attached to each variable.

In some cases, we can reduce disclosure risk by first transforming one or more variables. This will often arise with skewed distributions with long tails where, for example, a logarithmic transformation will create fewer extreme values. In such cases, the noise will be added to the transformed variable. In an analysis where the original variable is required in the model of interest, then for the likelihood term associated with the model of interest, the back-transformed value of the proposed value will be used. This is also described in the appendix.

A key issue, of course, is the requirement that the data provider supplies to the data analyst the necessary parameters used to generate the noise. Since the degree of privacy established needs to be published and the degree of privacy established is a function of these parameters, in a weak sense aspects of the noise will become publicly available. This, however, is not seriously disclosive since a guarantee of h-level disclosure is only weakly informative about the noise parameter values themselves. Furthermore, we have also shown in Table 3 that even where the attacker has access to the noise parameters this does not materially improve the probability of disclosure. Thus, a sensible precaution is to release the noise information only to accredited data analysts under secure conditions and avoid the release to an external malicious attacker. It is also worth mentioning that an attacker who has access to h-values may be deterred from attempting an attack when they realise the low chances of succeeding.

Whilst our proposed procedure for additive noise described in Section 2 can provide general protection against attacks, a data provider may wish to guard against specific aspects of disclosiveness in the data, such as the presence of one or two very large outliers, where the use of truncation on the perturbed data may be adequate. As long as the rules used by the data provider are made available, the analyst typically will be able to take account of these additional constraints in the analysis. As discussed in the appendix, it is possible to incorporate judicious groupings of data values within the estimation algorithm, and this allows the data provider some freedom in deciding where to coarsen particular data ranges. Further research will explore such possibilities, and in particular investigate the trade-off between increased security and reduced efficiency.

The protection method of perturbing a few variables in released data and other common approaches such as truncation and grouping that we have presented here are all standard statistical disclosure control (SDC) methods implemented by statistical agencies (Willenborg and De Waal 2001; Hundepool et al. 2012 and references therein). The computer science definition of Differential Privacy (DP) also assumes a worst case scenario that the attacker knows who is in the data set and does not take into account any protection afforded by sampling. The perturbation mechanism in the DP setting is also additive random noise where the noise is generated using, for example, the Laplace distribution and the parameters of the noise distribution depend on a privacy budget and the ‘sensitivity’ defined as the maximum distance of two neighbouring data sets that differ by only one data subject. For more details on DP, (see Dwork 2006; Dwork and Roth 2014). However, DP is related to output perturbation where every query is perturbed and it is less relevant in our case, where we release microdata with only a few variables perturbed, coarsened or truncated as is the norm in SDC practices at statistical agencies. One advantage of the DP framework is that the noise distribution does not need to be secret and can be freely released, thus removing one potential threat. However, this does not appear to be an insurmountable problem in our proposed approach, as we have shown in the simulation in Section 3. See Charest (2010) and Rinott et al. (2018) for examples of statistical inference under DP.

Data sets are often supplied with weights that may incorporate aspects of sample design or bias correction procedures. This poses particular problems for our procedure described in Section 5 and the appendix, as it does in general for models utilising Bayesian methods. In a recent paper, Goldstein et al. (2018) showed how weights could be incorporated in a Bayesian model for handling missing data. Goldstein et al. (2017) extend the model for missing data to handle measurement errors, and the procedures for handling weights given by Goldstein et al. (2018) can be extended in a straightforward fashion to this extended model. It should also be noted that the ability of the extended model to handle both measurement errors and missing data allows our procedures to deal with the case where there are missing values.

There are practical considerations to be taken into account if our procedures are to be implemented. Not least of these is the need to provide easy-to-use software to perform the appropriate analysis on the noisy data and accompanying training materials. The software routines written in MATLAB (2017), used for the present article, are not optimised for either speed or user accessibility. They are, however, available by request from the first author.

Finally, as we pointed out in the introduction in Section 1, it is important to recognise that there is always a trade-off between reducing disclosure risk and increasing the complexity and efficiency of any resulting analysis. The more noise that is added, the lower the statistical efficiency. In practice, the balance between disclosure risk and analytical efficiency can be tailored to individual data users through a secure environment and the implementation of automatic procedures for noise dataset generation according to given specifications of disclosure risk and statistical estimation efficiency. The safer the environment of the data analyst, in general the less noise will be needed, and likewise the level of noise could be tailored to the sensitivity of the data.

10. Appendix: Random Noise Addition and Model Estimation with Measurement Errors

The following exposition is for a single level linear model with a single predictor variable that contains noise (measurement error) with known parameters. The case of several predictors with added noise, the case where we have a generalised linear model and the multilevel case follow straightforwardly. We assume multivariate normality for the added noise and where we have categorical or count variables or continuous variables for which a normalising transformation exists (Goldstein et al. 2009). Then the appropriate extra steps are inserted into the MCMC algorithm to enable a random draw from the underlying normal distributions. The following estimation steps are based on those described by Goldstein et al. (2017). We assume uniform priors for all the model parameters described below.

Define the true values of the variable with measurement error (noise) as X_1 and those without measurement error as X_2, and $X = [X_1\, X_2]$.

Define the joint model – the measurement error model (MEM) in two parts, (1a) and (1b) and the model of interest (MOI) (1c) – see derivation below.

$$x_1 = X_1 + \gamma_1 \tag{A1a}$$

$$X_1 = X_2^T \alpha + \gamma_2 \tag{A1b}$$

$$Y = X\beta + e \tag{A1c}$$

For the priors we have

$$p(\alpha) \propto 1$$

$$p\left(\beta\right) \propto 1$$

$$p(\delta) \propto 1$$

$$\sigma_{\gamma_1}^{-2} \sim gamma\left(\epsilon, \epsilon\right)$$

$$\sigma_{\gamma_2}^{-2} \sim gamma\left(\epsilon, \epsilon\right)$$

$$\sigma_{e}^{-2} \sim gamma\left(\epsilon, \epsilon\right)$$

where n is the sample size, $\epsilon = 0.001$, the $\gamma_{1i}, \gamma_{2i}, e_i$ are obtained by subtraction, and $\gamma_1 \sim N\left(0, \sigma^2_{\gamma_1}\right)$, $\gamma_2 \sim N\left(0, \sigma^2_{\gamma_2}\right), e \sim N\left(0, \sigma^2_e\right)$. We note that the MOI (A1c) may contain functions of the X_1, such as interaction or power terms. Lower case variables define observed and upper case true values, and we assume the residual terms in (A1a)-(A1c) are independent. A Metropolis step for the true value is used for record i, where a new value is proposed. If we denote this by X_{1i}, the joint log likelihood for (A1a), (A1b) and (A1c) is

$$-\left\{1.5log2\pi + \log\left(\sigma_{\gamma_1}\sigma_{\gamma_2}\sigma_e\right) + \frac{0.5(x_{1i} - X_{1i})^2}{\sigma^2_{\gamma_1}} + \frac{0.5\left(X^T_{1i} - X^T_{2i}\alpha\right)^2}{\sigma^2_{\gamma_2}} + \frac{0.5\left(\tilde{y}_i\right)^2}{\sigma^2_e}\right\} \tag{A2}$$

where $\tilde{y}_i = y_i - X_i\beta$ and only the final three terms in (A2) are required in the Metropolis step.

For a proposal distribution we can use

$$p\left(X_1 | x_1\right) \sim N\left(x_1 R, R(1 - R)\sigma^2_{x_1}\right) \tag{A3}$$

where R is the reliability $= \frac{var(X_1)}{var(x_1)}$. Model (A1) is similar to the formulation by Richardson and Gilks (1993) where they have a 'gold standard' validation sample that provides the information contained in (A1a). In the present case, of course, the values of the noise variances are known to the data analyst. The remaining steps for the parameters in the model of interest are standard and can be found in Goldstein et al. (2017).

We can readily extend this model to the case where we have multilevel data with random effects following the exposition in Goldstein et al. (2017).

For the case where we have more than one variable with measurement error, we can propose the set of values defined independently for each variable or look at the joint proposal distribution in the case where correlated noise has been used, namely $(X_1|x_1) \sim MVN\left(X_1\Omega_{x_1}^{-1}\Omega_{X_1}, \Omega_{X_1} - \Omega_{X_1}\Omega_{x_1}^{-1}\Omega_{X_1}\right)$; although the use of correlated noise generally would seem to be unnecessary and serves only to complicate the analysis. Further details can be found in Goldstein et al. (2017).

For discrete variables where we have misclassification errors, there is an analogous procedure (Goldstein and Browne 2016), but this becomes complicated when there are multiple categories. Instead, we introduce a novel procedure as follows.

For the discrete variables in X_1 in (A1a), we now have the set of category codes $(0, \ldots, p - 1)$, as discussed in Section 5. Associated with such a variable we will have $p - 1$ dummy variables D_1. The proposal distribution for the Metropolis step is conveniently chosen as the observed distribution across the categories, based on choosing the nearest integer, or alternatively the 'true' distribution could be made available by the data provider for use as the proposal distribution. The log-likelihood contribution is then given by one of the following, depending on the observed value m_{ij}

$$-\left\{0.5log2\pi + \log(\sigma_m) + \frac{0.5\left(m_{ij} - j^*\right)^2}{\sigma_m^2}\right\} \quad \text{if } 0 < m_{ij} < p - 1$$

$$\log\left(\int_{-\infty}^{0} \phi(f)df\right), \quad f \sim N\left(j^*, \sigma_m^2\right) \quad \text{if } m_{ij} \leq 0$$

$$\log\left(\int_{p-1}^{\infty} \phi(f)df\right), \quad f \sim N\left(j^*, \sigma_m^2\right) \quad \text{if } m_{ij} \geq p - 1$$

where j^* is the proposed value and m_{ij} is the observed 'noisy' value truncated at zero and $p - 1$.

In some cases we may wish to present perturbed categorical data to a data analyst only as a set of discrete values, for example as the nearest integer to the 'noisy' value. Thus, for example for a perturbed value in the range $(-\infty, 1.5)$ we would report the value as 1 and generally as j if it is in the interval $(j - 0.5, j + 0.5)$. For the likelihood contribution for a proposed value j^* we now have the likelihood contributions

$$\int_{-\infty}^{0.5} \phi(f)df \quad \text{if } m_{ij} = 0$$

$$\int_{m_{ij}-0.5}^{m_{ij}+0.5} \phi(f)df \quad \text{if } 0 < m_{ij} < p - 1$$

$$\int_{p-1.5}^{\infty} \phi(f)df \quad \text{if } m_{ij} = p - 1$$

For each proposed category for variable X_1 we will have a corresponding entry of '1' for the dummy variable in the model of interest, that is, in D_1. This set of dummy variables will enter the MOI as predictors with the response vector corresponding to (A1b), where the default link function is the multivariate probit as described in Goldstein et al. (2009). If we wish to allow actual measurement errors for continuous predictors, as well as the imposed anonymisation categorical measurement errors, it will be convenient to propose true values for the former in a separate step, conditional on all the current categorical predictor values. Where we have imposed anonymisation measurement errors for continuous variables, as well as actual measurement errors, we may simply add the variances for the former to the corresponding diagonal terms of the actual measurement error covariance matrix.

Standard errors as quoted in the tables are the standard deviations computed from the MCMC chains in the usual way.

As mentioned in Section 5, in some cases we may have additional constraints on the data values. For example, if the perturbed value x_1 is constrained to be no larger than a chosen value, we may have

$$if(x_1 > C_1),\ set\, x_1 = C_1$$

where the value C_1 does not occur in the dataset. Then, whenever $x_1 = C_1$, we would sample a value from the upper tail of the normal distribution defined by the current values from (A1b) and use this in the Metropolis step. Likewise, where we apply truncation to a continuous response variable, an extra step will be introduced into the algorithm that samples from the tail area of the normal distribution conditional on current parameter values. A similar procedure could be used more generally where a grouping of values takes place. However, care will be needed to ensure that there is not too much loss of efficiency associated with this.

For a response variable with added noise, the situation is more complex, but for normally distributed responses we can carry out an analysis using the observed values to obtain consistent estimates of the fixed coefficients. We can obtain a consistent estimate for the residual variance by using the observed variance during the chain sampling and subtracting the (known) variance for the measurement error in the response, say σ_{δ}^2, from the final estimated residual variance. As we show in our example, this may result in a large increase in the standard errors when the measurement error variance is large relative to the residual variance. When data are released, it may be acceptable to provide any variable(s) to be treated as a response with the true values or just a small amount of noise, and as we show in our example in Section 8, this may not be too disclosive.

For categorical responses with misclassification or measurement errors, a further modification is required. Thus, for example, in the case of a binary response where normally distributed noise has been added as in (4), if we choose, as above, to round the observed value to the nearest integer (0, 1), denoted by y_i, then for the likelihood for the model of interest we can write

$$\delta_1 = \Pr\left(obs = 1|true = 0\right) = \int_{0.5}^{\infty} \phi_\delta(t)dt), \quad \Pr\left(obs = 1|true = 1\right) = 1 - \int_{0.5}^{\infty} \phi_\delta(t)dt)$$

and this leads to

$$\Pr\left(y_i = 1|X\beta\right) = \delta_1 + (1 - 2\delta_1)\int_{-X\beta}^{\infty} \phi(t)dt, \;\; \phi_\delta \sim N\left(0, \sigma_\delta^2\right) \tag{A4}$$

Thus for the β parameters, since σ_δ^2 is known, we will have a Metropolis step for each one in turn using the observed (0, 1) values, subject to $\delta_1 + (1 - 2\delta_1)\int_{-X\beta}^{\infty} \phi(t)dt < 1$.

Routines to implement the models described in this appendix have been written in MATLAB (2017) and details can be obtained from the first author.

11. References

Arima, S. and S. Polettini. 2019. “A Unit Level Small Area Model with Misclassified Covariates.” *Journal of the Royal Statistical Society*, Series A. Early View: DOI: https://doi.org/10.1111/rssa.12468.

Brand, R. 2002. *Microdata Protection Through Noise Addition*. In *Inference control in statistical databases*, 97–116. Berlin, Heidelberg: Springer.

Chambers, R.L. and R. Dunstan. 1986. “Estimating Distribution Functions from Survey Data.” *Biometrika* 73: 597–604. DOI: https://doi.org/10.2307/2336524.

Charest, A.-S. 2010. “How Can we Analyse Differentially-private Synthetic Datasets?” *Journal of Privacy and Confidentiality* 2: 21–33. DOI: https://doi.org/10.29012/jpc.v2i2.589.

Cox, L., A.F. Karr, and S.K. Kinney. 2011. “Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act.” *International Statistical Review* 79(2): 160–183. DOI: https://doi.org/10.1111/j.1751-5823.2011.00140.x.

Delaigle, A. and P. Hall. 2008. “Using SIMEX for Smoothing-Parameter Choice in Errors-in-Variables Problems.” *Journal of the American Statistical Association* 103(481): 280–287.

Duncan, G.T. and S. Mukherjee. 2000. “Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks Through Additive Noise.” *Journal of the American Statistical Association* 95(451): 720–729.

Dwork, C. 2006. “Differential Privacy.” In *ICALP 2006*, edited by M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, 4052: 1–12. Lecture Notes in Computer Science. Heidelberg: Springer.

Dwork, C. and A. Roth. 2014. “The Algorithmic Foundations of Differential Privacy.” *Foundations and Trends in Theoretical Computer Science* 9: 211–407. DOI: https://doi.org/10.1561/0400000042.

Elliot, M.J. and A. Dale. 1999. “Scenarios of attach: the data intruder’s perspective on statistical disclosure risk.” *Netherlands Official Statistics* 14: 6–10. Available at: https://www.researchgate.net/profile/Ton_De_Waal/publication/255565237_Exact_ disclosure_in_a_super-table/links/0c960539450114582d000000.pdf#page=6 (accessed February 2020).

Fuller, W.A. 1993. “Masking procedures for microdata disclosure limitation.” *Journal of Official Statistics* 9: 383–406. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/masking-procedures-for-microdata-disclosure-limitation.pdf (accessed February 2020).

Fuller, W.A. 2006. *Measurement error models*. Chichester: John Wiley and Sons.

Goldstein, H. 2011. *Multilevel Statistical Models*. Chichester: Wiley.

Goldstein, H. and W. Browne. 2016. “Multilevel models: current developments”. Wiley Online Library. Available at: www.wileyonlinelibrary.com/ref/stats.

Goldstein, H., W.J. Browne, and C. Charlton. 2017. “An MCMC procedure for handling measurement and misclassification errors alongside missing data in multilevel multivariate generalised linear models with an application to a study of Australian youth.” *Journal of Applied Statistics*. DOI: https://doi.org/10.1080/02664763.2017.1322558.

Goldstein, H., J. Carpenter, and M. Kenward. 2018. “Bayesian models for weighted data with missing values: a bootstrap approach.” *Journal of the Royal Statistical Society*, Series C. DOI: https://doi.org/10.1111/rssc.12259.

Goldstein, H., J. Carpenter, M. Kenward, and K. Levin. 2009. “Multilevel models with multivariate mixed response types.” *Statistical Modelling* 9(3): 173–197. DOI: https://doi.org/10.1177/1471082X0800900301.

Goldstein, H., J. Rasbash, M. Yang, G. Woodhouse, H. Pan, D. Nuttall, and S. Thomas. 1993. “A Multilevel Analysis of School Examination Results.” *Oxford Review of Education* 19(4): 425–433.

Gouweleeuw, J., P. Kooiman, L.C.R.J. Willenborg, and P.P. de Wolf. 1998. “Post Randomisation for Statistical Disclosure Control: Theory and Implementation.” *Journal of Official Statistics* 14: 463–478. Available at: https://pdfs.semanticscholar.org/cd28/1be11657b944b74169b8fe35dddb91d558e8.pdf (accessed February 2020).

Hundepool, A., J. Domingo-Ferrer, L. Francono, S. Giessing, E. Schulte-Nordholt, K. Spicer, and P.P. de Wolf. 2012. *Statistical Disclosure Control. Wiley Series in Survey Methodology*. Chichester: John Wiley and Sons.

Hwang, J.T. 1986. “Multiplicative Errors-in-Variables Models with Applications to Recent Data” Released by the U.S. Department of Energy. *Journal of the American Statistical Association* 81(395): 680–688. DOI: https://doi.org/10.1080/01621459.1986.10478321.

Kim, J.J. 1986. *A Method for Limiting Disclosure in Micro-data Based on Random Noise and Transformation*. ASA Proceedings of the Section on SRM: 370–374. Available at: http://www.asasrms.org/Proceedings/papers/1986_069.pdf (accessed February 2020).

Little, R.J.A. 1993. “Statistical analysis of masked data.” *Journal Official Statistics* 9: 407–426. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf (accessed February 2020).

MATLAB. 2017. Available at: https://www.mathworks.com/products/matlab.html (accessed October 2017).

Reiter, J.P. 2005. “Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study.” *Journal of the Royal Statistical Society,* Series A 168(1): 185–205. DOI: https://doi.org/10.1111/j.1467-985X.2004.00343.x.

Reiter, J.P. and R. Mitra. 2009. “Estimating Risks of Identification Disclosure in Partially Synthetic Data.” *Journal of Privacy and Confidentiality* 1(1): 99–110. DOI: https://doi.org/10.29012/jpc.v1i1.567.

Richardson, S. and W.R. Gilks. 1993. “Conditional independence models for epidemiological studies with covariate measurement error.” *Statistics in Medicine* 12: 1703–1722. DOI: https://doi.org/10.1002/sim.4780121806.

Rinott, Y., C. O’Keefe, N. Shlomo, and C. Skinner. 2018. “Confidentiality and Differential Privacy in the Dissemination of Frequency Tables.” *Statistical Sciences* 33(3): 358–385. DOI: https://doi.org/10.1214/17-STS641.

Rubin, D.B. 1993. “Discussion Statistical Disclosure Limitation.” *Journal of Official Statistics* 9: 461–468. Available at: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf (accessed February 2020).

SdcMicro. 2019. Available at: http://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf.

Shlomo, N. 2010. “Measurement Error and Statistical Disclosure Control.” In *PSD 2010: Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and E. Magkos, 6344: 118–126. Springer LNCS. DOI: https://doi.org/10.1007/978-3-642-15838-4_11.

Shlomo, N. and T. de Waal. 2008. “Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure.” *Journal of Official Statistics* 24(2): 1–26. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/protection-of-micro-data-subject-to-edit-constraints-against-statistical-disclosure.pdf (accessed February 2020).

Shlomo, N. and C.J. Skinner. 2010. “Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata.” *Annals of Applied Statistics* 4(3): 1291–1310. DOI: https://doi.org/10.1214/09-AOAS317.

Tendick, P. 1991. “Optimal Noise Addition for Preserving Confidentiality in Multivariate Data.” *Journal of Statistical Planning and Inference* 27(3): 341–353. DOI: https://doi.org/10.1016/0378-3758(91)90047-I.

Ting, D., S. Fienberg, and M. Trottini. 2008. “Random orthogonal matrix masking methodology for microdata release.” *International Journal on Information and Computer Security* 2(1): 86–105. DOI: https://doi.org/10.1504/IJICS.2008.016823.

Willenborg, L. and T. de Waal. 2001. “Elements of Statistical Disclosure Control in Practice.” *Lecture Notes in Statistics*, 155. New York: Springer-Verlag.

Winkler, W.E. 1998. “Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata.” *Research in Official Statistics* 1: 87–104. Available at: https://www.census.gov/srd/papers/pdf/rrs2005-09.pdf.

Woo, Y.M.J. and A. Slavkovic. 2014. “Generalized Linear Models with Variables Subject to Post Randomization Method.” *Italian Journal of Applied Statistics* 24(1): 29–56. Available at: http://sa-ijas.stat.unipd.it/sites/sa-ijas.stat.unipd.it/files/2%20Woo%20pg%20(1).pdf (accessed February 2020).

Received May 2018
Revised July 2019
Accepted August 2019

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 117–136, http://dx.doi.org/10.2478/JOS-2020-0006

Can Interviewer Evaluations Predict Short-Term and Long-Term Participation in Telephone Panels?

Oliver Lipps[1] *and Marieke Voorpostel*[1]

Interviewers often assess after the interview the respondent's ability and reluctance to participate. Prior research has shown that this evaluation is associated with next-wave response behavior in face-to-face surveys. Our study adds to this research by looking at this association in telephone surveys, where an interviewer typically has less information on which to base an assessment. We looked at next-wave participation, non-contact and refusal, as well as longer-term participation patterns. We found that interviewers were better able to anticipate refusal than non-contact relative to participation, especially in the next wave, but also in the longer term. Our findings confirm that interviewer evaluations – in particular of the respondent's reluctance to participate – can help predict response at later waves, also after controlling for commonly used predictors of survey nonresponse. In addition to helping to predict nonresponse in the short term, interviewer evaluations provide useful information for a long-term perspective as well, which may be used to improve nonresponse adjustment and in responsive designs in longitudinal surveys.

Key words: Attrition; postsurvey adjustment; longitudinal weights.

1. Introduction

It is of central importance in longitudinal surveys that respondents participate repeatedly. Only by collecting multiple observations can we assess change over time accurately. Hence, a large body of literature has examined the causes of attrition, as well as strategies to prevent and correct for it (e.g., Couper and Ofstedal 2009; Lipps 2012a; Schonlau et al. 2010). One strategy to improve our understanding of what guides continued participation that has received only little research attention is the use of assessments that interviewers provide after completing an interview. Such assessments consist of questions at the end of the questionnaire that address the interviewer's impression of the respondent's willingness to participate and the quality of the responses. The few studies that use interviewer evaluations to predict later-wave response are mostly based on data from face-to-face interviews (but see Barrett et al. 2006, for a combined telephone and face-to-face approach). They indicate that negative assessments of the interviews and the respondents are associated with subsequent nonresponse in face-to-face surveys (e.g., Plewis et al. 2017).

The extent to which interviewers' evaluations predict subsequent participation in longitudinal household studies conducted by telephone remains unknown. One could

[1] FORS c/o University of Lausanne, 1015 Lausanne, Switzerland. Emails: oliver.lipps@fors.unil.ch and marieke.voorpostel@fors.unil.ch

argue that in the absence of face-to-face contact, telephone interviewers have less information on which to base their evaluations. On the other hand, face-to-face contact with the respondent may also increase the use of stereotypes, decreasing the validity of such evaluations. Our study is the first to assess the relationship between interviewer evaluations and subsequent participation in a longitudinal household panel conducted by telephone.

Interviewer evaluations are useful not only to improve our understanding of what determines (repeated) survey participation, but also to correct for nonresponse. Variables based on the interviewers' assessments of the responsiveness and carefulness of the respondents are rarely included in longitudinal weights, although they have the potential to improve nonresponse adjustments (Peytchev and Olson 2007). In addition, with increasing possibilities for responsive and adaptive design procedures (Groves and Heeringa 2006; Chun et al. 2017; Schouten et al. 2017; Wagner 2008), being able to determine which respondents are more likely to drop out has great potential benefits in terms of spending more resources on these respondents and better tailoring fieldwork efforts to keep them in the longitudinal study. Consequently, we distinguished between nonresponse due to non-contact and due to refusal when looking at next-wave participation, which is an informative distinction for responsive design procedures.

Another contribution this study makes is that we extended our focus from only looking at short-term (wave-to-wave) participation to also including longer-term participation *patterns*. Attrition studies so far have exclusively examined the predictive quality of interviewer evaluations on response at the next wave (Kalton et al. 1990; Plewis et al. 2017; Watson and Wooden 2004). Yet, as longitudinal surveys increasingly include online questionnaires in their designs, interviewer evaluations collected at every wave become less standard. It then becomes useful to have insight in the extent to which interviewer evaluations are able to predict subsequent response patterns, on top of other commonly used predictors.

2. Background

There is a long history of collecting interviewer evaluations and observations in survey research (Feldman et al. 1951). These assessments generally aim to capture two aspects: the interviewer's assessment of the current or future *reluctance* of the respondent to participate (responsiveness), and the interviewer's assessment of the *ability* of the respondent to complete the survey task, which includes the quality of the responses (carefulness).

Interviewer evaluations of the respondent's carefulness help to assess the quality of the responses provided. Barrett and colleagues (Barrett et al. 2006) found that interviewers' perceptions of respondents' performance were valid indicators of item nonresponse and frequency of "don't know" answers in a survey of persons with mental and physical disabilities. However, Kirchner et al. (2017) pointed out that this association is not surprising, as interviewers base their evaluations among other things on item nonresponse during the interview. Kaminska et al. (2010) used interviewer evaluations to measure respondents' reluctance and cognitive ability in the European Social Survey. They showed that reluctance was associated with satisficing, but that this relationship was explained by lower cognitive ability. Peytchev and Olson (2007) demonstrated for the US National

Election Study that inclusion of interviewer assessments improved nonresponse adjustments.

A second strength of interviewer evaluations of both responsiveness and carefulness is that they help to predict (continued) survey participation. In cross-sectional studies, interviewer evaluations have been used to predict final participation after each contact (Eckman et al. 2013). The aim of such contact-based evaluations is an ad hoc tailoring to accommodate sample members' possible concerns, which may change across contacts. A number of studies using longitudinal surveys have shown that interviewers' evaluations of respondents are predictive of cooperation at the next wave of data collection, in addition to commonly used predictors. A recent study by Plewis et al. (2017) used interviewer evaluations at the fourth wave of the UK Millenium Cohort Study to predict response behavior in the subsequent wave, which took place four years later. They found that the interviewer's assessment of the likelihood that the respondent would participate in the future, the difficulty the respondent had answering the questions, the enjoyment of and cooperation during the interview all predicted both non-contact and refusal in the next wave. Moreover, they found that dropout of such "difficult" respondents caused bias, suggesting that in order to reduce nonresponse bias, there is benefit to directing additional resources toward keeping such respondents in the panel (Plewis et al. 2017). Kalton et al. (1990) found for the American Changing Lives study that interviewer ratings of the respondent's understanding of the questions, cooperation and the enjoyment of the interview were positively associated with participation in the second wave. Lepkowski and Couper (2002) confirmed this for the National Election Studies. In their study examining sample attrition between the first two waves of the Household, Income and Labour Dynamics in Australia (HILDA), Watson and Wooden (2004) found that the interviewer's evaluation of the respondent's cooperativeness, suspicion of the study, and required assistance to complete the interview were all associated with attrition in the second wave. In a study using 14 waves of the HILDA panel, Perez and Baffour (2018) showed that the interviewer's evaluation of the respondent's suspicion of the study, question understanding, and the respondent's cooperativeness were all associated with personal characteristics that were precursors of panel attrition.

To our knowledge, only two studies, both conducted using the British Household Panel Survey (BHPS), assessed the extent to which interviewers' evaluations are predictive of participation in the longer term. Uhrig (2008) showed that the interviewer's judgement of poor cooperation of the respondent affected both later wave refusal and non-contact, although his models only controlled for interview characteristics and fieldwork operations, and not for any other predictors of nonresponse. Laurie et al. (1999) showed that poor cooperation by the respondent as reported by the interviewer in the first wave was associated with nonresponse at the fourth wave. These findings indicate that interviewer evaluations have predictive power that goes beyond participation in the next wave.

Research findings so far have shown that interviewers' evaluations in face-to-face surveys help predict later response behavior in longitudinal studies, even if commonly used predictors are taken into account. This means that if interviewers' evaluations correlate with the survey variable, they have the potential to improve nonresponse adjustments and need to be included in appropriate models. We examined whether we find this association also in telephone panel surveys. Compared with face-to-face surveys, interviewers in telephone surveys have less information on which to base their evaluations. There is a higher level of anonymity in telephone surveys (Block and Erskine 2012). Telephone interviewers are not in the home of the respondent, cannot engage in nonverbal communication and their interviews tend to be of a shorter duration, limiting the total time of interaction with the respondent. For our study we might expect, on the one hand, that as a result it would be harder for interviewers to evaluate respondents, limiting predictive power of such evaluations. On the other hand, there may also be less noise in the evaluations, as there is less information (e.g., on the home or physical appearance of the respondent) that may lead to the development of stereotypes. Kirchner et al. (2017) assessed the validity of interviewer evaluations in telephone surveys. They based their expectations on the continuum model of impression formation, which suggests that impressions initially are formed based on observed characteristics (stereotyping), but that one can move beyond this way of processing by using actual behavior to update these pre-existing notions. They found no evidence of stereotyping. Rather, interviewers based their assessments on the quality of the data provided and on other behavior of the respondent. In general confirming the validity of assessments that interviewers are able to make in telephone interviews.

In this study, we first explored the bivariate relationship between interviewer evaluations and participation *patterns* to see if there was indeed a longer-term association. Then we controlled for a rich set of predictors of nonresponse to see if these relationships persist. Next, we analyzed interviewer evaluation predictive power on participation in the next wave in the same way. We expected the interviewer to be better able to predict participation at the next wave than in the long term. Nevertheless, we expected the interviewer evaluation to have a nonzero longer-term predictive power. In addition, we expected that interviewers are better at anticipating subsequent refusal than non-contact (Uhrig 2008; Plewis et al. 2017; Lipps 2012a).

We assessed the contribution that interviewer evaluations make above other commonly used predictors. If, after controlling for such predictors, interviewer evaluations do not have any explanatory power, there is no need to collect such data for the purpose of predicting participation. In line with most previous studies, we took into account commonly used socio-demographic characteristics, as well as the likelihood that the household would move. These covariates are commonly included in studies linking interviewer evaluations to longitudinal survey participation. Our study controlled for a number of additional variables not included in previous studies that also capture the respondent's survey language competence, engagement, characteristics of the participation history (i.e., number of waves in the panel) and indicators of response quality (i.e., proportion of "don't know" answers). Finally, we also added measures of the social engagement of the respondent (Groves et al. 2009).

3. Data and Methods

3.1. Data

We used data from the Swiss Household Panel (SHP) (Tillmann et al. 2016). The SHP is an ongoing large-scale, nationwide, annual, centralized computer-assisted telephone interview (CATI) panel survey that started in 1999 with a sample of 5,074 households and added a refreshment sample in 2004, consisting of 2,538 households. Both samples were randomly drawn from the telephone register and cover the Swiss residential population. A second refreshment sample, drawn at random from the Swiss population register, started in 2013, consisting of 3,989 households. Households were contacted via landline and mobile numbers. Households that split up or moved remain in the study as long as they reside in Switzerland. New household members are included in the study. In addition, households that refused in a certain wave are re-approached for a number of waves in an effort to keep them in the sample. Each year, the household reference person is asked to first complete the household roster using a grid questionnaire. All listed household members of at least 14 years old are then approached to complete individual questionnaires. As household members could only participate if the household reference person participated, we focused only on participation at the household level, measured by whether the household reference person completed the grid questionnaire. Interviewer evaluations are available for all individual interviews in every wave since 2004. Upon completion of each individual questionnaire, the interviewer answers a number of questions on the impression the interviewer had of the respondent and the interview.

We have two different analytical samples (for details, see Subsection 3.5, Methods): a sample to investigate the predictive power of the interviewer evaluations for long-term participation patterns (Sample 1), and a sample to investigate the predictive power of the interviewer evaluations for wave-to-wave response (Sample 2). For the long-term model based on Sample 1, we used data on subsequent participation from all household reference persons who completed an individual questionnaire in 2004 (N = 4,394). For the short-term model (Sample 2), we use all households for which we have at least one interviewer evaluation, which implies that the reference person completed an individual questionnaire at least once between 2004 and 2016. We then looked at participation in the next wave, including data from 2005 to 2017. We disregarded observations in which a household became ineligible (left the country, all members institutionalized or deceased) or was no longer approached for other reasons (no valid landline or mobile telephone number available, written refusal, not contactable or refusal for several waves in a row) and were left with 10,336 households and 61,844 observations. After dropping observations with any missing values on one of the covariates, 60,298 observations (from 10,185 households) remained in our short-term analytical Sample (2). The distributions of the participation patterns and the interviewer evaluations differed only slightly between the full sample and the analytical sample (where observations with any missing covariates were dropped). For example, while the next wave grid is completed by 92.9% of the households in the full sample, this is the case for 93.1% in the analytical sample. For the interviewer evaluation variables, the biggest relative difference occurs for ‘easy to convince the sample member’, which is true of 96.3% in the full sample and 96.6% in the analytical sample. We thus abstain from imputing missing covariates.

3.2. *Dependent Variables*

We used household-level information on participation to construct the categories of our dependent variables. Participation in the survey implied that the household reference person completed the grid questionnaire in a given wave. Non-participation was either the result of non-contact or refusal at the household level. The distinction between these two states is often not clear-cut, as noncontact can be a form of hidden refusal (Stoop 2005). We coded the outcome in a given wave as refusal when the final disposition code was refusal. Whether or not to assign to refusal or noncontact is less clear when an interviewer managed to make contact with the household, the household reference person did not give a hard refusal, but the interview had not taken place by the end of the six-month fieldwork period. We decided to code these cases as a final non-contact (about half of all final non-contacts) because the contact with the household could have been with another household member. In addition, the outcome was coded as final non-contact if the household had a valid telephone number (landline or mobile number), but the interviewer could not establish any contact with the household. We cannot completely rule out the possibility that some non-contacts were hidden refusals, but these cases should be few. In addition, there is no information about the mode of contact (landline or mobile number). In the analytical sample for the short-term response (Sample 2), of the 6.9% nonresponse, 5.0% were refusals and 1.9% non-contacts, meaning that non-contacts were relatively rare.

After assigning participation statuses to the observations, we constructed two dependent variables for the two analyses in this study: for the long-term model (Sample 1), we used participation patterns. For each household reference person who completed an individual questionnaire in 2004, we established patterns of participation (grid completion) in subsequent waves, hence one pattern for each reference person. We distinguished between five exhaustive and mutually exclusive participation patterns that distinguish between full participation, immediate dropout and three different irregular patterns. Our choice of patterns acknowledges the fact that the determinants of participation and therefore the process of attrition differs across respondents. Some respondents participate mostly loyally, others drop out definitely sooner or later, and still others participate infrequently (see Lugtig 2014):

1. The household is either highly committed to the survey or developed a habit of taking part (Lugtig 2014) and participated at every wave in which it was eligible (full participation, 53.9%),
2. The household refused at least once, for example due to a temporary “shock” caused by a life-changing event (Lemay 2009), but participated at the most recent wave in which it was eligible (there may have been non-contacts at other waves) (refusal and re-entry, 9.1%),
3. The household never refused, but at least once no contact could be established, probably caused by a temporary absence (see Lepkowski and Couper 2002, who show that refusal and non-contact have different determinants), and the household participated in the most recent wave in which it was eligible (non-contact and re-entry, 2.6%),
4. The household participated at least in one later wave, but dropped out in or before the most recent wave in which it was eligible, which may be the result of panel fatigue (Lemay 2009) (participation and dropout, 30.1%),
5. The household did not participate in any later wave, probably because it experienced participation negatively (Lemay 2009) (immediate dropout, 4.3%).

For the short-term model (Sample 2), we used next-wave participation status as the dependent variable, distinguishing participation, refusal, and non-contact. We accounted for the multilevel structure of the data with observations nested in households.

3.3. Independent Variables

The main independent variables concerned the interviewer evaluation of the respondent's *ability* and *reluctance,* which we will use to predict next-wave response and response patterns. Ability was measured with the question "Was the respondent's understanding of the questions. . .?" with responses including good (2), fair (1), and poor (0). Three questions measured reluctance: "In general, what was the respondent's attitude toward the interview?" (Friendly and cooperative (3), cooperative but not particularly interested (2), impatient and restless (1), and hostile (0)); "How difficult was this case to get?" (Somewhat easy (0), somewhat difficult (1), and very difficult (2)); and "Do you expect this respondent to participate in the next wave?" (Absolutely (3), probably yes (2), maybe (1), and no (0)). We dichotomized the separate indicators of ability and reluctance, each coded 1 if the respondent was fully cooperative or able (the respective first categories), and 0 otherwise.

The distribution of these indicators in the short-term analytical sample (Sample 2) was as follows (in brackets for the long-term analytical Sample 1 for the year 2004): 92.9% (89.9%) fully understood the questions well, 96.4% (94.5%) were friendly and cooperative, 96.6% (93.3%) were somewhat easy to get, and 84.3% (66.0%) were expected to absolutely participate in the next wave. The interviewer evaluation indicators correlated only weakly with each other with an absolute correlation coefficient between .25 ('respondent friendly' and 'respondent understands questions well', in the short-term analytical sample) and .36 ('respondent easy to get' and 'respondent will participate in next wave', in the short-term analytical sample). We therefore included all four indicators in the models.

3.4. Control Variables

We included the following control variables known to be associated with attrition (Voorpostel 2010; Voorpostel and Lipps 2011):

- Geographical mobility: ownership of the house (yes/no), degree of intention to move during the coming year (0–10), nationality (Swiss/from a neighboring country, that is, Germany, Austria, Liechtenstein, Italy, and France / from another country), whether the respondent has lived in Switzerland for at least 14 years (yes/no).
- Demographic characteristics: age in categories (14–39, 40–49, 50–59, 60–69, 70+), survey language competence (first language, second language, other), education (less than high school level, equivalent to high school level, more than high school level), gender, partner status (living with partner, does not live with partner, no partner), presence of children aged up to seven years in the household (yes/no), number of household members eligible for an interview.
- Social engagement/inclusion/participation history/income: member of a club (yes/no), political interest (0–10), trust in people (0–10), number of waves, equivalized household income.

Table 1. Descriptive statistics independent variables analytical samples for long-term patterns (2004) and for wave-to-wave sample (2005–2017). Data: SHP 2004–2017.

	2004		2005–2017	
Variables	Mean	Std dev.	Mean	Std dev.
Independent variables: Interviewer evaluations				
Respondent is friendly [0,1]	0.945	0.227	0.964	0.187
Respondent understands questions well [0,1]	0.899	0.302	0.929	0.258
Respondent is difficult to convince [0,1]	0.067	0.249	0.034	0.181
Respondent will repeat next wave [0,1]	0.660	0.474	0.843	0.364
Control variables: geographical mobility				
Respondent owns residence [%]	0.464	0.499	0.518	0.500
Intention to move in next 12 months [0 = not at all, . . . , 10 certainly]	1.337	2.895	1.185	2.677
Nationality: Swiss [%]	0.896	0.306	0.906	0.292
Nationality: from a neighbouring country [%]	0.060	0.238	0.058	0.234
Nationality: from another country [%]	0.044	0.205	0.036	0.187
In Switzerland for more than 14 years [%]	0.948	0.221	0.974	0.159
Control variables: demographic characteristics				
Age [14–39 years] [%]	0.292	0.455	0.210	0.408
Age [40–49 years] [%]	0.249	0.433	0.225	0.418
Age [50–59 years] [%]	0.196	0.397	0.217	0.412
Age [60–69 years] [%]	0.140	0.347	0.176	0.380
Age [70+ years] [%]	0.123	0.329	0.172	0.378
Survey language is first language [%]	0.954	0.210	0.971	0.167
Survey language is second language [%]	0.039	0.194	0.026	0.159
Survey language is not first or second language [%]	0.007	0.085	0.003	0.051
Education level: less than high school equivalent [%]	0.163	0.369	0.147	0.354
Education level: high school equivalent [%]	0.556	0.497	0.510	0.500
Education level: more than high school [%]	0.282	0.450	0.343	0.475
Gender: Male [%]	0.365	0.481	0.383	0.486
Partner: yes, living together [%]	0.644	0.479	0.651	0.477
Partner: yes, but not living together	0.100	0.301	0.097	0.295
Partner: no [%]	0.256	0.436	0.253	0.435
Children under seven years in household [%]	0.132	0.338	0.102	0.302
Number of interview eligible household members	2.047	0.967	2.063	0.967
Control variables: social engagement				
Member of a club [%]	0.497	0.500	0.409	0.500
Political interest [0 = not at all, . . . , 10 very interested]	5.690	2.836	5.625	2.794
Trust [0 = can't be too careful, 10 most people can be trusted]	5.623	2.507	6.193	2.273
Number of waves [1, . . . , 6], [1, . . . , 18]	3.666	2.447	8.040	4.656
Equivalised household income [Sfr.]	117829	95868	131026	112500
Response quality variables				
Proportion of don't knows [%]	0.008	0.013	0.007	0.013
Proportion of refused items [%]	0.002	0.008	0.002	0.007
Survey year [2004], [2004, . . . , 2016]	2004	0	2010.569	3.837
Sample (observations)	4,394	60,298		
Sample (households)	4,394	10,185		

- Response quality in current wave: proportion of "don't know" answers, proportion of refused answers.
- Survey-related variables: survey year (only in the short-term analysis).

Table 1 presents the descriptive statistics of all variables in the study for both analytical samples.

3.5. *Methods*

We analyzed the relationship between interviewer evaluations and survey participation separately for participation patterns and wave-to-wave participation. For both, we first assessed the *bivariate* relationship between interviewer evaluations and participation outcomes, followed by logistic multinomial models controlling for a number of covariates. For the bivariate analyses, we compared mean evaluation scores by participation pattern and by next-wave participation and used Chi-square tests to test whether participation differed by interviewer's assessment of the respondent's ability and reluctance. The logistic multinomial model predicting participation patterns used participation in all eligible waves after 2004 as the reference category. For the model predicting next wave participation, the reference category was participation in the next wave. We included covariates (see measures) from 2004 to predict the participation pattern afterwards, and wave-specific measures from 2004–2016 to predict participation in the next wave. As we had multiple observations for each household, we employed a logistic multinomial random intercept model in the fourth analysis. Although households were crossed with interviewers, we did not include interviewers as a level of analysis in any of the models, because the actual outcome depends on a random interviewer in the *next* wave (short-term model) or on several random interviewers in subsequent waves (long-term model). Although the interviewer-respondent assignment is random, there may be very small selection effects due to different shifts worked by interviewers. We tested if interviewer evaluation heterogeneity (e.g., by systematically providing better evaluations) provided different results by using interviewer-centered evaluations as alternative predictors. Differences compared with our original evaluation variables were only marginal.

The equation of the model is presented below (see Haynes et al. 2005, 9–10). Suppose the outcome variable Y_{it} has $J = 3$ categories ($1 =$ response $=$ reference category, $2 =$ non-contact, $3 =$ refusal), then the probability for household i in wave t to not being contacted ($j = 2$) or to refuse ($j = 3$) rather than to respond ($j = 1$) given a set of control variables X_{it} can be estimated as:

$$\pi_{itj} = \Pr(Y_{it} = j|X_{it}) = \frac{e^{X_{it}\beta_j}}{\sum_{k=1}^{J} e^{X_{it}\beta_k}}, \quad \mathrm{j} = 2, 3$$

corresponding to the following multinomial logit model:

$$\log\left(\frac{\pi_{itj}}{\pi_{it1}}\right) = X_{it}^{'}\beta_j, \quad \mathrm{j} = 2, 3$$

If we allow for household-specific random effects α_{ij} and let Z_{ij} denote a vector of coefficients for the random effects, then the model extends to:

$$\log\left(\frac{\pi_{itj}}{\pi_{it1}}\right) = X_{it}^{'}\beta_j + Z_{ij}^{'}\alpha_{ij}, \quad \mathrm{j} = 2, 3$$

The random effects α_{ij} capture non-observable household effects that are assumed to come from a multivariate normal distribution with mean zero and variance-covariance matrix Σ. We first estimated the model for participation patterns before turning to next wave participation status. The next section presents the results of our analyses.

4. Results

4.1. Interviewer Evaluations and Participation Patterns

Our first research question concerned the extent to which interviewer evaluations in 2004 were able to predict five different subsequent participation *patterns* in the panel (see definition in Subsection 3.2). We first note that the participation patterns are different for all four (dichotomized) evaluation criteria, with significant (1% level) chi^2-values ranging between 61.0 for the ‘easy to convince’ criterion, and between 16.7 and 19.3 for the other three criteria (cross-tabulations not shown). Table 2 presents for each evaluation criterion the mean score by participation pattern.

The majority of the participants, regardless of participation pattern, were evaluated as able and willing to complete the survey task. Respondents who participated in every wave in which they were eligible, but also respondents who could not be contacted at least once but re-entered later, had the most positive evaluations. Respondents who dropped out immediately received the lowest evaluation for “respondent will participate next wave”, but not for the other items. Respondents who refused but re-entered were least likely to be evaluated as friendly and easily convinced. This showed that a less positive evaluation was predictive of refusal at a later wave, but these respondents were not necessarily lost to the study. Overall, the interviewers gave ratings that were more positive to respondents with full participation, or non-contact and re-entry, and more negative ratings to respondents who refused in subsequent waves.

Table 2. Mean scores (in 2004) by participation pattern (2004–2017). Data: SHP 2004–2017.

Interviewer evaluations:	Full participation	Refusal and re-entry	Non-contact and re-entry	Participation and dropout	Immediate dropout	Total
Respondent is friendly	0.956	0.910	0.965	0.939	0.921	0.945
Respondent understands questions	0.912	0.905	0.939	0.871	0.884	0.899
Respondent is easy to convince	0.957	0.875	0.948	0.914	0.878	0.933
Respondent will repeat survey	0.680	0.653	0.748	0.625	0.619	0.660
N	2369	400	115	1321	189	4394

In the next step, we analyzed whether interviewer evaluations were predictive of participation patterns after controlling for common covariates. We modeled the same long-term patterns using a multinomial logistic model controlling for all covariates. The first part of Table 3 presents the beta coefficients of the evaluation covariates of the resulting model (see Table 3a). We listed the full model in the Appendix, Section 6 (Table A3).

Since our main interests are the *ceteris paribus* differences in participation pattern by evaluation, we calculated average marginal effects of a more positive evaluation for the different patterns (see Table 3b). Average marginal effects show the *ceteris paribus* increase of the probability of a participation pattern for a positive evaluation compared with a less positive evaluation. For example, the average marginal effect of .142 for the 'easy to convince' item indicates that the likelihood of full participation was 14.2 percentage points higher for the easier to convince than for the difficult to convince, of refusal and re-entry 9.4 percentage points lower, of non-contact and re-entry (an insignificant) 0.6 percentage points lower, and of immediate dropout (an insignificant) 2 percentage points lower, holding all other variables constant. Irrespective of the 'friendliness' of the respondent or his or her evaluation of the likelihood to repeat the next wave, the patterns do not change. For the 'understands questions' item the predicted probability to refuse and re-enter was 3 percentage points higher if the interviewer gave a positive account of the respondent's understanding, relative to a bad understanding. To assess the part of the marginal effects that is due to confounding predictors, we compared the marginal effects to those calculated with no covariates except the evaluations (see Table 3c). As it turns out, we found somewhat larger effects of the 'easy to convince' item and a (5%) significant effect of the 'understands questions' item on participation and dropout. However, overall, the interpretation remains largely the same. The results in this section show that the interviewer's assessment of respondents' understanding and, in particular, reluctance, added to the prediction of participation patterns in subsequent waves, even after controlling for other common predictors of survey participation.

4.2. *Interviewer Evaluations and Short-Term Participation Status*

In the last part of our analysis, we investigated the association between the interviewer's evaluation of the respondent *in each wave* (i.e., across all years and not only in 2004) and the next wave participation *status*. Thus, we focused on short-term participation only, and examined the extent to which interviewer evaluations of separate attributes contributed to predicting nonresponse. Similar to the participation patterns, we first checked bivariate distributions of the interviewer evaluations and the three response outcomes participation, refusal, and non-contact in the next wave (cross-tabulations not shown). Again, all chi^2-values were significant (1% level) for all four evaluation criteria, ranging between 463.8 for the 'easy to convince' criterion, 360.9 for the 'repeat next wave' criterion, 140.0 for the 'friendliness' criterion and 75.7 for the 'understands' criterion. Table 4 presents the mean score by participation outcome for each evaluation criterion.

Again, the majority of the participants were evaluated as able and willing to complete the survey task. Next-wave participants had the most positive evaluations, followed by not contacted respondents. Refusing respondents were least likely to receive positive

Table 3. Multinomial logistic model including all control variables (a, b), without control variables (c), modeling long-term response patterns (base category: full participation).

a) Beta coefficients: full model

Interviewer evaluations:	Refusal, and re-entry	Noncontact, and re-entry	Participation, and dropout	Immediate dropout
Respondent is friendly	−0.65*	−0.01	−0.18	−0.45
Respondent understands questions	0.46*	0.18	0.04	0.35
Respondent is easy to convince	−1.09**	−0.55	−0.40*	−0.78**
Respondent will repeat survey	0.20	0.39	−0.04	0.06
Constant	−0.20	−1.94	1.28**	0.53

b) Average marginal effects (dy/dx w.r.t. interviewer evaluations): full model

Interviewer evaluations:	Full participation	Refusal, and re-entry	Noncontact, and re-entry	Participation, and dropout	Immediate dropout
Respondent is friendly	.071	−.054	.004	−.007	−.013
Respondent understands q.	−.033	.030*	.003	−.011	.011
Respondent easy to convince	.142**	−.094**	−.006	−.020	−.022
Respondent repeats survey	−.009	.016	.009	−.018	.002

c) Average marginal effects (dy/dx w.r.t. interviewer evaluations): only evaluations

Interviewer evaluations:	Full participation	Refusal, and re-entry	Noncontact, and re-entry	Participation, and dropout	Immediate dropout
Respondent is friendly	.037	−.049	.005	.015	−.008
Respondent understands q.	.025	.029*	.010	−.067*	.004
Respondent easy to convince	.190**	−.099**	−.004	−.051	−.035*
Respondent repeats survey	.009	.013	.009	−.029	−.002

Data: SHP 2004–2017, N = 4,394 households in 2004.
$p < 0.01$ (**), $p < 0.05$ (*).

Table 4. Mean evaluation scores (2004–2016) by participation (2005–2017). Data: SHP 2004–2017.

Interviewer evaluations:	Participation	Refusal	Noncontact	Total
Respondent is friendly	0.966	0.924	0.963	0.964
Respondent understands questions	0.931	0.889	0.928	0.929
Respondent is easy to convince	0.970	0.899	0.940	0.966
Respondent will repeat survey	0.850	0.723	0.802	0.843
N	56149	2990	1159	60298

evaluations on all items. This showed that a less positive evaluation was associated with next wave non-contact and especially refusal.

Next, we used multivariate logistic random intercept models to control for commonly used covariates and to accommodate the clustering of the data on the household level. The first part of Table 5 presents the beta coefficients of the evaluation covariates

Table 5. Multinomial logistic random intercept model including all control variables (a, b), without control variables (c), modeling wave-to-wave participation (participation in next wave (base category), refusal, noncontact).

a) Beta coefficients: full model

Interviewer evaluations:	Refusal	Noncontact
Respondent is friendly	−0.18	0.36
Respondent understands questions	0.22*	0.09
Respondent is easy to convince	−0.75**	−0.77**
Respondent will repeat survey	−0.37**	−0.21*
Constant	−.069**	−1.27**
Variance (observation level)	1.42**	
Variance (household level)	2.20**	

b) Average marginal effects (dy/dx w.r.t. interviewer evaluations): full model

Interviewer evaluations:	Participation	Refusal	Noncontact
Respondent is friendly	.003	−.009	.006*
Respondent understands questions	−.011*	.009**	.001
Respondent is easy to convince	.060**	−.042**	−.017**
Respondent will repeat survey	.022**	−.018**	−.004

c) Average marginal effects (dy/dx w.r.t. interviewer evaluations): only evaluations (and survey year)

Interviewer evaluations:	Participation	Refusal	Noncontact
Respondent is friendly	.005	−.010	.006
Respondent understands questions	−.004	.000	.004
Respondent is easy to convince	.069**	−.052**	−.017**
Respondent will repeat survey	.024**	−.021**	−.003

Data: SHP 2004–2017, N = 10,185 households, 60,298 observations.
$p < 0.01$ (**), $p < 0.05$ (*).

of the results (see Table 5a). Again, we included the complete model in the Appendix (Table A5).

When looking at the coefficients, all evaluation items, except for whether or not the respondent was friendly, were associated with next-wave participation, refusal or non-contact.

Also for the short-term model, we calculated average marginal effects of participating in the next wave based on the full model for changes of the four interviewer evaluations (see Table 5b). The average marginal effects showed a significant effect for friendliness: a higher respondent friendliness slightly (at a 5% significance level) increased the predicted probability for next-wave non-contact. A better question understanding meant a 1.1 percentage point *lower* participation on the 5% significance level, and a 0.9 percentage point *higher* refusal, net of other covariates in the model. With regard to difficulty to convince the respondent, the probabilities varied the most: Respondents who were easy to convince exhibited a 6.0 percentage points higher participation rate, a 4.2 percentage points lower refusal rate, and a 1.7 percentage points lower non-contact rate than more reluctant respondents. Again, to assess the part of these marginal effects that is due to confounding predictors, we compared them to average marginal effects calculated with no covariates except the evaluations and the survey years (see Table 5c). In particular, we investigated the counterintuitive effect in which respondents who understood the questions well were, surprisingly, slightly more likely to refuse in the next wave when all variables were included in the model. Looking at the average marginal effects with no covariates except the evaluations and the survey year, the coefficient of question understanding lost significance. It became significantly positive on refusal only after adding the other covariates. For the other evaluation criteria, we did not observe such a change in the size or direction of the coefficients when covariates were dropped from the model.

5. Conclusion

We set out to assess the extent to which interviewers' evaluations are predictive of response patterns and dropout in longitudinal telephone surveys. Prior studies have shown that interviewer assessments are associated with continued participation in the context of face-to-face interviews (Plewis et al. 2017; Kalton et al. 1990; Lepkowski and Couper 2002). Our study adds to this knowledge by extending it to telephone interviews, a setting in which the interviewer has less information on which to judge the respondent. We can draw the following conclusions.

First, our study showed that even in the absence of face-to-face contact, interviewers' assessments of respondents were predictive of subsequent response patterns, which is in line with the findings from Kirchner et al. (2017). We found that when the interviewer evaluated the respondent as capable with minimal levels of reluctance, respondents were thereafter more likely to become loyal participants, with possible non-contacts in between. In particular, when the interviewer judged the respondent as easy to convince, the respondent was more likely to participate in subsequent waves, rather than to refuse or drop out altogether.

Second, the interviewers' evaluations helped to predict short-term participation, distinguishing participation, refusal, and non-contact. As expected, interviewers were better able to predict refusal than non-contact. If the respondent was easy to convince and the interviewer judged it likely he or she would return in the next wave, respondents were more likely to participate and less likely to refuse.

An important finding was that the evaluation of the interviewer added to the explanation of next wave or later nonresponse, even if a large set of commonly used predictors of survey participation were taken into account. The ability and reluctance of the respondent as assessed by the interviewer was not fully captured by characteristics such as educational attainment, political interest or civic engagement, nor by past or current survey behavior.

Our study suggests, in line with Peytchev and Olsen (2007), that the use of information provided by interviewers may help to improve longitudinal weights designed to reduce bias from selective attrition. However, for interviewer evaluations to make good adjustment weights, they should not only be able to predict attrition, but also potential research variables (Little and Vartivarian 2005). Future studies on weighting should explore this further.

Our study is also relevant for responsive design development. Interviewers' assessments may be one of the criteria on which to base decisions on how to allocate fieldwork effort to minimize attrition. For example, respondents who are evaluated as difficult to convince in a given wave can, in the next wave, be offered a higher incentive, a specially tailored newsletter, or be assigned an interviewer who is experienced in refusal conversion. Also, interviewers could be incentivized with additional bonuses for difficult cases. The results of our study can be used to identify respondents for whom special treatment would be most beneficial to improve continued participation in the panel study. Since non-contacts are difficult to anticipate, other measures such as a better timing of the call should be envisaged for this group (Lipps 2012a, 2012b).

There were some limitations to this study that can be addressed in future research. For example, we had no information on whether households used a landline or a mobile telephone to answer the survey request. Since there are differences between landline and mobile telephone surveys with regard to the mechanism generating nonresponse, as well as conducting the interview (Kennedy 2010), future studies should distinguish mobile and landline devices, where possible. There are probably additional measurement errors due to inter-interviewer variability, since interviewer evaluations vary in the extent to which they accurately measure objective characteristics such as gender, ethnicity, and dwelling (e.g., Casas-Cordero et al. 2013; Sinibaldi et al. 2013). In addition, the four evaluation items were signficantly skewed towards positive evaluations. We encourage survey methodologists to design and test more elaborated items. To improve the prediction of subsequent response behavior, these items should be based primarily on the evaluated difficulty to convince the respondent to participate, and the evaluated likelihood that the respondent will repeat the survey, and less on the friendliness or degree of question understanding. As our analyses show, these latter items measure respondent characteristics that may not directly predict subsequent participation. However, for the moment, we showed that at least two of the four interviewer evaluations that were considered provide useful additional information at very low cost. A well-designed battery of interviewer evaluations should become an integral part of at least the first wave of every large-scale panel survey. Designing and conducting appropriate experiments remains to be done in future research.

6. Appendix

Table A3. Multinomial logistic model including all control variables modeling long-term response patterns (base category: full participation).

Interviewer evaluations	Refusal, and re-entry	Noncontact, and re-entry	Participation, and dropout	Immediate dropout
Respondent is friendly	− 0.65*	− 0.01	− 0.18	− 0.45
Respondent understands questions	0.46*	0.18	0.04	0.35
Respondent is difficult to convince	− 1.09**	− 0.55	− 0.40*	− 0.78**
Respondent will repeat survey	0.20	0.39	− 0.04	0.06
Owner of house	0.14	− 0.07	0.05	− 0.19
Degree of willingness to move [0..10]	− 0.02	− 0.02	− 0.01	0.03
Swiss (ref: from another country)	0.23	− 0.03	− 0.50*	− 0.30
From a neighbouring country (ref: from another country)	0.21	0.25	− 0.64**	− 0.46
Lives in Switzerland for 14 years or more	0.30	0.35	0.30*	− 0.41
Age: 40–49 (ref: 14–39)	0.22	− 0.58*	− 0.12	− 0.59**
Age: 50–59 (ref: 14–39)	0.25	− 0.92**	− 0.30**	− 0.41
Age: 60–69 (ref: 14–39)	0.17	− 1.67**	− 0.25	− 0.41
Age: 70+ (ref: 14–39)	0.26	− 1.65**	− 0.01	− 0.31
Survey language is second language (ref: first)	0.17	0.31	− 0.05	− 0.53
Survey language is neither first nor second	− 0.34	− 13.62**	0.34	0.55
Education equivalent to high school (ref: low)	− 0.08	− 0.08	− 0.16	− 0.16
Education more than high school (ref: low)	− 0.48**	− 0.05	− 0.26*	− 0.36
Male (ref: female)	0.13	0.11	0.23**	0.22
Lives with partner (ref: no partner)	− 0.13	0.34	0.05	0.43
Partner but not living together (ref: no partner)	− 0.32*	0.37	0.18*	0.15
Child under 7 in the household (ref: no child)	0.08	− 0.82	− 0.29*	− 0.31
Number interview eligible household members	0.09	0.06	0.13**	0.12
Member of a club	− 0.14	− 0.10	− 0.17*	− 0.01
Political interest [0..10]	− 0.06**	− 0.11**	− 0.06**	− 0.12**
Trust in people [0..10]	− 0.02	0.03	− 0.04**	− 0.03
Number of waves in the panel	− 0.17**	− 0.07	− 0.15**	− 0.26**
Equivalised household income	− 0.00	− 0.00	− 0.00*	0.00
Proportion of refused answers	− 2.66	− 6.86	6.66*	14.78**
Proportion of don't know answers	10.06	4.79	3.50	6.25
Constant	− 0.20	− 1.94	1.28**	0.53

Data: SHP 2004–2017, N = 4,394 households in 2004, $r^2 = .055$.
$p < 0.01$ (**), $p < 0.05$ (*).

Table A5. Multinomial logistic random intercept model including all control variables modeling wave-to-wave participation (participation in next wave (base category), refusal, noncontact).

Interviewer evaluations	Refusal	Noncontact
Respondent is friendly	− 0.18	0.36
Respondent understands questions	0.22*	0.09
Respondent is difficult to convince	− 0.75**	− 0.77**
Respondent will repeat survey	− 0.37**	− 0.21*
Owner of house	0.11*	− 0.27**
Degree of willingness to move [0..10]	0.02*	0.04**
Swiss (ref: from another country)	− 0.07	− 0.45**
From a neighbouring country (ref: from another country)	− 0.23	− 0.12
Lives in Switzerland for 14 years or more	− 0.13	− 0.28
Age: 40–49 (ref: 14–39)	0.15*	− 0.33**
Age: 50–59 (ref: 14–39)	0.30**	− 0.81**
Age: 60–69 (ref: 14–39)	0.10	− 1.56**
Age: 70+ (ref: 14–39)	0.54**	− 2.34**
Survey language is second language (ref: first)	0.13	− 0.22
Survey language is neither first nor second	0.73*	0.80
Education equivalent to high school (ref: low)	− 0.15*	− 0.18
Education more than high school (ref: low)	− 0.40**	− 0.40**
Male (ref: female)	0.17**	0.39**
Lives with partner (ref: no partner)	− 0.22*	0.40**
Partner but not living together (ref: no partner)	− 0.28**	0.42**
Child under seven in the household (ref: no child)	− 0.23**	− 0.25*
Number interview eligible household members	0.06	− 0.13**
Member of a club	− 0.19**	− 0.31**
Political interest [0..10]	− 0.05**	− 0.04**
Trust in people [0..10]	− 0.03**	− 0.06**
Number of waves in the panel	− 0.09**	− 0.06**
Equivalised household income	− 0.00	0.00
Proportion of refused answers	7.31**	2.80
Proportion of don't know answers	8.83**	2.68
Survey year 2004	− 0.24*	− 0.79**
Survey year 2005	− 0.87**	− 0.65**
Survey year 2006	− 0.37**	− 0.53**
Survey year 2007	− 0.50**	− 0.85**
Survey year 2008	− 0.88**	− 0.67**
Survey year 2009	− 1.58**	− 0.60**
Survey year 2010	− 1.23**	− 0.47**
Survey year 2011	− 1.29**	− 0.36*
Survey year 2012	− 0.70**	− 0.24
Survey year 2013	− 0.37**	0.01
Survey year 2014	− 0.38**	− 0.25
Survey year 2015	− 0.18*	0.31*
Constant	− .069**	− 1.27**
Variance (observation level)	1.42**	
Variance (household level)	2.20**	

Data: SHP 2004–2017, N = 10,185 households, 60,298 observations. p < 0.01 (**), p < 0.05 (*).

7. References

Barrett, K., M. Sloan, and D. Wright. 2006. "Interviewer Perceptions of Interview Quality". *In Proceedings of the American Statistical Association, Survey Research Methodology Section*, 4026–4033. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2006/Files/JSM2006-000644.pdf (accessed February 2020).

Block, E.S. and L. Erskine. 2012 "Interviewing by Telephone: Specific Considerations, Opportunities, and Challenges." *International Journal of Qualitative Methods* 11(4): 428–445. DOI: https://doi.org/10.1177/160940691201100409.

Casas-Cordero, C., F. Kreuter, Y. Wang, and S. Babey. 2013. "Assessing the Measurement Error Properties of Interviewer Observations of Neighbourhood Characteristics", *Journal of the Royal Statistical Society:* Series A 176: 229–249. DOI: https://doi.org/10.1111/j.1467-985X.2012.01065.x.

Chun, A., B. Schouten and J. Wagner. 2017. JOS Special Issue on Responsive and Adaptive Survey Design: "Looking Back to See Forward-Editorial". *Journal of Official Statistics* 33(3): 571–577. DOI: http://dx.doi.org/10.1515/JOS-2017-0027.

Couper, M.P. and M.B. Ofstedal. 2009 "Keeping in Contact with mobile Sample Members." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 183–203. New York: Wiley.

Eckman, S., J. Sinibaldi, and A. Möntmann-Hertz. 2013. "Can Interviewers effectively rate the Likelihood of Cases to cooperate?" *Public Opinion Quarterly* 77(2): 561–573. DOI: https://doi.org/10.1093/poq/nft012.

Feldman, J.J., H. Hyman, and C.W. Hart. 1951. "A Field Study of Interviewer Effects on the Quality of Survey Data." *Public Opinion Quarterly* 15(4): 734–761. DOI: https://doi.org/10.1086/266357.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. 2nd Ed., *Wiley Series in Survey Methods*. Hoboken, NJ: John Wiley & Sons.

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for actively controlling Survey Errors and Costs." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(3): 439–457. DOI: https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Haynes, M., M. Western, and M. Spallek. 2005. "Methods for categorical longitudinal Survey Data: Understanding Employment Status of Australian Women." Paper prepared for the HILDA Survey Research Conference, University of Melbourne, Australia, 29–30 September, 2005. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwjfrIGR1sbnAhUkShUIHf1PAQgQFjABegQIBRAB&url=https%3A%2F%2Fwww.researchgate.net%2Fpublication%2F43458841_Methods_for_Categorical_Longitudinal_Survey_Data_Understanding_Employment_Status_of_Australian_Women&usg=AOvVaw26JHyUI5MCUOF5FqgSI1Nc (accessed February 2020).

Kennedy, C.K. 2010. *Nonresponse and Measurement Error in Mobile Phone Surveys. Doctoral thesis, University of Michigan*, Ann Arbor. Available at: https://deepblue.lib.umich.edu/handle/2027.42/75977 (accessed February 2020).

Kalton, G., J. Lepkowski, G.E. Montanari, and D. Maligalig. 1990. Characteristics of second Wave Nonrespondents in a Panel Survey. In *Proceedings of the American Statistical Association, Survey Research Methodology Section*, 462–467. Alexandria, VA: American Statistical Association.

Kaminska, O., A.L. McCutcheon, and J. Billiet. 2010. “Satisficing among reluctant Respondents in a Cross-National Context.” *Public Opinion Quarterly* 74(5): 956–984. DOI: https://doi.org/10.1093/poq/nfq062.

Kirchner, A., K. Olson, and J.D. Smyth. 2017. “Do Interviewer Postsurvey Evaluations of Respondents’ Engagement measure who Respondents are or what they do?” A Behavior coding Study. *Public Opinion Quarterly*: DOI: https://doi.org/10.1093/poq/nfx026.

Laurie, H., R. Smith, and L. Scott. 1999. “Strategies for reducing Nonresponse in a longitudinal Panel Survey.” *Journal of Official Statistics* 15(2): 169–282. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/strategies-for-reducing-nonresponse-in-a-longitudinal-panel-survey.pdf (accessed February 2020).

Lemay, M. 2009. *Understanding the Mechanism of Panel Attrition. Unpublished Doctoral thesis, University of Maryland*, College Park, MD, U.S.A.

Lepkowski, J.M. and M.P. Couper. 2002. “Nonresponse in the second Wave of longitudinal Household Surveys.” In *Survey nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J. Little, 259–272. New York: Wiley.

Lipps, O. 2012a. Using Information from Telephone Panel Surveys to predict Reasons for Refusal. *Methods, data, analyses* 6(1): 3–20. Available at: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-314544 (accessed August 2019).

Lipps, O. 2012b. “A Note on improving Contact Times in Panel Surveys.” *Field Methods* 24(1): 95–111. DOI: https://doi.org/10.1177/1525822X11417966.

Little, R.J. and S. Vartivarian. 2005. “Does weighting for Nonresponse increase the Variance of Survey Means?” *Survey Methodology* 31(2): 161–168. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=2ahUKEwiTkPGq2MbnAhXHUBUIHWLjCq4QFjACegQIBRAB&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.692.507%26rep%3Drep1%26type%3Dpdf&usg=AOvVaw1fkEbnvfviCWWInhD3jfs6 (accessed February 2020).

Lugtig, P. 2014. “Panel Attrition: Separating Stayers, fast Attriters, gradual Attriters, and Lurkers.” *Sociological Methods & Research* 43(4): 699–723. DOI: https://doi.org/10.1177/0049124113520305.

Perez, F.P. and B. Baffour. 2018. “Respondent mental Health, mental Disorders and Survey Interview Outcomes.” *Survey Research Methods* 12(2): 161–176. DOI: https://doi.org/10.18148/srm/2018.v12i2.7225.

Peytchev, A. and K. Olson. 2007. “Using Interviewer Observations to improve Nonresponse Adjustments: NES 2004.” In *Proceedings of the American Statistical Association, Survey Research Methodology Section*, 3364–3371. Alexandria, VA: American Statistical Association. Available at: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1149&context=sociologyfacpub (accessed February 2020).

Plewis, I., L. Calderwood, and T Mostafa. 2017. “Can Interviewer Observations of the Interview Predict future Response?” *Methods, data, analyses* 11(1): 29–44. DOI: https://doi.org/10.12758/mda.2016.010.

Schonlau, M., N. Watson, and M. Kroh. 2010 "Household Survey Panels: How much do following Rules affect Sample Size?" *Survey Research Methods* 5(2): 53–61. DOI: https://doi.org/10.18148/srm/2011.v5i2.4665.

Schouten, B., A. Peytchev, and J. Wagner. 2017. *Adaptive Survey Design*. Boca Raton: CRC Press.

Sinibaldi, J., G. Durrant, and F. Kreuter. 2013. "Evaluating the Measurement Error of Interviewer observed Paradata." *Public Opinion Quarterly* 77: 173–193. DOI: https://doi.org/10.1093/poq/nfs062.

Stoop, I. 2005. *The Hunt for the Last Respondent, Nonresponse in Sample Surveys*. The Hague: Social and Cultural Planning Office of the Netherlands. Available at: https://dspace.library.uu.nl/bitstream/handle/1874/2900/full.pdf (accessed February 2020).

Tillmann, R., M. Voorpostel, U. Kuhn, F. Lebert, V-A. Ryser, O. Lipps, B. Wernli, and E. Antal. 2016. "The Swiss Household Panel Study: Observing social Change since 1999." *Longitudinal and Life Course Studies* 7(1): 64–78. DOI: http://dx.doi.org/10.14301/llcs.v7i1.360.

Uhrig, N.S.C. 2008. *The Nature and Causes of Attrition in the British Household Panel Survey. ISER Working Paper 2008-05*. Colchester: ISER, University of Essex. Available at: https://www.researchgate.net/profile/Sc_Uhrig/publication/242116089_The_Nature_and_Causes_of_Attrition_in_the_British_Household_Panel_Survey/links/0a85e537c6cd161804000000/The-Nature-and-Causes-of-Attrition-in-the-British-Household-Panel-Survey.pdf (accessed February 2020).

Voorpostel, M. 2010. Attrition Patterns in the Swiss Household Panel by Demographic Characteristics and Social Involvement. *Swiss Journal of Sociology* 36(2): 359–377. Available at: https://serval.unil.ch/resource/serval:BIB_78D83DD99A6F.P001/REF (accessed February 2020).

Voorpostel, M. and O. Lipps. 2011. Attrition in the Swiss Household Panel: Is change associated with drop-out? *Journal of Official Statistics* 27(2): 301–318. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/attrition-in-the-swiss-household-panel-is-change-associated-with-drop-out.pdf (accessed February 2020).

Wagner, James. 2008. *Adaptive Survey Design to Reduce Nonresponse Bias*, *Dissertation, University of Michigan*.

Watson, Nicole, and Mark Wooden. 2004. Sample attrition in the HILDA survey. *Australian Journal of Labour Economics* 7(2): 293–308. Available at: https://businesslaw.curtin.edu.au/wp-content/uploads/sites/5/2016/05/AJLE-v7n2-watson.pdf (accessed February 2020).

Received January 2019
Revised July 2019
Accepted September 2019

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 137–149, http://dx.doi.org/10.2478/JOS-2020-0007

Notes on Use of the Composite Estimator: an Improvement of the Ratio Estimator

Kung-Jong Lui[1]

This article discusses use of the composite estimator with the optimal weight to reduce the variance (or the mean-squared-error, MSE) of the ratio estimator. To study the practical usefulness of the proposed composite estimator, a Monte Carlo simulation is performed comparing the bias and MSE of composite estimators (with estimated optimal weight and with known optimal weight) with those of the simple expansion and the ratio estimators. Two examples, one regarding the estimation of dead fir trees via an aerial photo and the other regarding the estimation of the average sugarcane acres per county, are included to illustrate the use of the composite estimator developed here.

Key words: Composite estimator; ratio estimator; simple expansion estimator; odds ratio; phi correlation; regression estimator.

1. Introduction

The ratio estimator, which incorporates the information on the auxiliary measurements into estimation, is one of the most commonly-used estimators in surveys. For example, the Current Population Survey employed the ratio estimator accounting for the number of African Americans to estimate the number of unemployed African Americans (Scheaffer et al. 2012). When the correlation between the auxiliary measurement and the studied measurement is high, the ratio estimator can outperform the simple expansion estimator with respect to precision (or variance) (Cochran 1977, 157–158). However, if this correlation is low, the ratio estimator can be less precise than the simple expansion estimator (Cochran 1977). The ratio estimator is biased, but its bias decreases to 0 as the sample size increases to ∞ (Cochran 1977). There are publications on adjusting or reducing the bias of the ratio estimator (Pascual 1961; Sahoo 1987; Cochran 1977). Because the bias of the ratio estimator is of order $1/n$, we commonly focus our attention on variance when comparing the ratio estimator with the simple expansion estimator (Cochran 1977; Scheaffer et al. 2012). Some discussions on use of the composite estimation technique to improve the precision of existing estimators in small area estimation (Lui and Cumberland 1991; Schaible 1978, Royall 1970) or to extend the ratio estimator to multivariate ratio-typo estimators have been presented elsewhere (Sukhatme et al. 1984, 217–223; Cochran 1977, 184–185). A recent discussion on use of the

[1] Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182-7720. USA. Email: klui@sdsu.edu

Acknowledgments: The author wishes to thank the Associate Editor and four Referees for many useful and valuable comments to improve the contents and clarity of this article.

composite estimator with weights proportional to the reciprocal of the variance of its component estimator for two-stage cluster sampling has also been presented by Lee et al. (2016). However, the weight suggested by Lee et al. (2016) is optimal only when its two component estimators are independent.

This article suggests that the idea of the composite estimator to reduce the bias and variance (or the mean-squared-error (MSE) of the ratio estimator may be applied with no additional efforts of collecting extra data. The optimal weight minimizing variance (or MSE) for the composite estimator under the simple random sampling (SRS) is derived. A Monte Carlo simulation comparing the performance of the simple expansion estimator, the ratio estimator, the composite estimator with an estimated optimal weight and the composite estimator with known optimal weight is carried out in a variety of situations. A brief discussion is given on deriving the optimal weight minimizing variance (or MSE) when measurements are dichotomous. Two examples, one regarding the estimation of dead fir trees via an aerial photo and the other regarding the estimation of the average sugarcane acres per county, are included to illustrate the use of the composite estimator developed here.

2. Notation and Methods

Suppose that a population consists of $\{(X_i, Y_i)|X_i > 0,\ Y_i > 0,\ i = 1, 2, 3, ..., N\}$, where X_i and Y_i are measurements on subject i and N is the population size. For clarity, we summarize the definitions of notation in the following:

Let $X = \sum_{i=1}^{N} X_i$ and $Y = \sum_{i=1}^{N} Y_i$ denote the population totals of measurements X_i and Y_i ;

let $\bar{X} = \sum_{i=1}^{N} X_i/N$ and $\bar{Y} = \sum_{i=1}^{N} Y_i/N$ denote the population means of measurements X_i and Y_i;

let $R = Y/X = \bar{Y}/\bar{X}$ denote the population total or mean ratio;

let $S_x^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2/(N-1)$ and $S_y^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2/(N-1)$ denote the population variances;

let $S_{xy} = \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})/(N-1)$ denote the population covariance between X_i and Y_i;

let $\rho = S_{xy}/(S_x S_y)$ denote the simple correlation coefficient between X_i and Y_i, as well as let $CV_x = S_x/\bar{X}$ and $CV_y = S_y/\bar{Y}$ denote the coefficients of variation for X_i and Y_i.

Note that since we will consider using the ratio estimator only when X_i and Y_i are positively correlated, we assume $\rho > 0$ in the following discussion.

Suppose that we employ the SRS scheme and obtain n subjects. Let notation c denote the collection of labels in the sample. Furthermore,

let $\bar{x}_s = \sum_{i \in c} X_i/n$ and $\bar{y}_s = \sum_{i \in c} Y_i/n$ denote the sample means of X_i and Y_i;

let $s_x^2 = \sum_{i \in c} (X_i - \bar{x}_s)^2/(n-1)$ and $s_y^2 = \sum_{i \in c} (Y_i - \bar{y}_s)^2/(n-1)$ denote the sample variances;

let $\hat{R} = \bar{y}_s/\bar{x}_s$ denote the ratio of the two sample means;

let $s_{xy} = \sum_{i \in c} (X_i - \bar{x}_s)(Y_i - \bar{y}_s)/(n-1)$ denote the sample covariance between X_i and Y_i;

let $\hat{\rho} = s_{xy}/(s_x s_y)$ denote the estimated correlation coefficient between X_i and Y_i ; as well as let $cv_x = s_x/\bar{x}_s$ and $cv_y = s_y/\bar{y}_s$ denote the sample coefficients of variation for X_i and Y_i.

Here, we focus our attention on estimation of the population total Y. Note that the simplest unbiased estimator for Y under the SRS is the simple expansion estimator given by (Cochran 1977)

$$\hat{Y} = N\bar{y}_s. \tag{1}$$

The variance for $\hat{Y}$ is (Cochran 1977, 23)

$$\begin{aligned} Var(\hat{Y}) &= N^2(1-f)S_y^2/n \\ &= N^2(1-f)\bar{Y}^2CV_y^2/n, \end{aligned} \tag{2}$$

where $f = n/N$ is the sampling fraction. To estimate the population total Y, we can also employ the most commonly-used ratio estimator defined by (Cochran 1977)

$$\hat{Y}_R = \hat{R}X, \tag{3}$$

where $\hat{R} = \bar{y}_s/\bar{x}_s$. The variance for $\hat{Y}_R$ is approximately given by (Cochran 1977, 154; see also Appendix 1, Subsection 7.1.)

$$\begin{aligned} Var(\hat{Y}_R) &= N^2(1-f)[S_y^2 + R^2S_x^2 - 2RS_{XY}]/n, \\ &= N^2(1-f)\bar{Y}^2[CV_y^2 + CV_x^2 - 2\rho CV_xCV_y]/n, \end{aligned} \tag{4}$$

where $R = \bar{Y}/\bar{X}$. Furthermore, we can show that the covariance between $\hat{Y}$ and $\hat{Y}_R$ can be approximated by (Appendix 1)

$$Cov(\hat{Y}, \hat{Y}_R) = N^2(1-f)\bar{Y}^2[CV_y^2 - \rho CV_xCV_y]/n. \tag{5}$$

Consider use of the composite estimator $w\hat{Y} + (1-w)\hat{Y}_R$ (where $0 \le w \le 1$) to estimate the population total Y. Note that the simple expansion estimator $\hat{Y}$ (1), the ratio estimator $\hat{Y}_R$ (3) and the composite estimator $w\hat{Y} + (1-w)\hat{Y}_R$ can all be expressed as $\sum_{i\in c} a_iy_i$, a linear combination of the sampled units. The weights a_i for $\hat{Y}_R$ are calibrated towards the population total X (*i.e.*, $\sum_{i\in c} a_ix_i = X$), while the weights a_i for $\hat{Y}$ and $w\hat{Y} + (1-w)\hat{Y}_R$ are not. Note also that since the weight w in $w\hat{Y} + (1-w)\hat{Y}_R$ is between 0 and 1, the absolute magnitude of the bias for the composite estimator $|E(w\hat{Y} + (1-w)\hat{Y}_R - Y)| = |(1-w)E(\hat{Y}_R - Y)|$, is always smaller than $|E(\hat{Y}_R - Y)|$, the absolute magnitude of the bias for the ratio estimator. Note that the ratio estimator is an asymptotically unbiased estimator for Y (Cochran 1977, 161) and so is the composite estimator. The optimal weight w minimizing the variance $Var(w\hat{Y} + (1-w)\hat{Y}_R)$ can be shown to equal

$$\begin{aligned} w^* &= (Var(\hat{Y}_R) - Cov(\hat{Y}, \hat{Y}_R))/(Var(\hat{Y}_R) + Var(\hat{Y}) - 2Cov(\hat{Y}, \hat{Y}_R)), \\ &= 1 - \rho CV_y/CV_x. \end{aligned} \tag{6}$$

Define w^* to be 0 if $1 - \rho CV_y/CV_x < 0$. Because $E(\hat{Y}_R - Y)^2 \approx Var(\hat{Y}_R)$ and $Cov(\hat{Y}, \hat{Y}_R) = E(\hat{Y} - Y)(\hat{Y}_R - Y)$, the optimal weight w^* (6) minimizing the variance of the composite estimator is approximately equivalent to minimizing its MSE. We denote the composite estimator $w^*\hat{Y} + (1-w^*)\hat{Y}_R$ with the optimal weight w^* by $\hat{Y}_C(w^*)$. When $0 < w^* < 1$, the variance for $\hat{Y}_C(w^*)$ is obtained by simply substituting w^* for w in

$Var(\hat{Y}_C(w))$ and is given by (Cochran 1977, 185)

$$\begin{aligned}Var(\hat{Y}_C(w^*)) &= (Var(\hat{Y})Var(\hat{Y}_R) - (Cov(\hat{Y},\hat{Y}_R))^2)/(Var(\hat{Y}) + Var(\hat{Y}_R) - 2Cov(\hat{Y},\hat{Y}_R)) \\ &= N^2(1-f)\bar{Y}^2 CV_y^2(1-\rho^2)/n. \qquad (7)\end{aligned}$$

On the basis of (2) and (7), we can show that the proportional reduction of variance (PRV) by use of $\hat{Y}_C(w^*)$ instead of $\hat{Y}$ is given by

$$PRV_S = [Var(\hat{Y}) - Var(\hat{Y}_C(w^*))]/Var(\hat{Y}) = \rho^2. \qquad (8)$$

Furthermore, the PRV by use of $\hat{Y}_C(w^*)$ instead of $\hat{Y}_R$ is given by

$$\begin{aligned}PRV_R &= [Var(\hat{Y}_R) - Var(\hat{Y}_C(w^*))]/Var(\hat{Y}_R) \\ &= 1/[1 + (1-\rho^2)/(CV_x/CV_y - \rho)^2], \qquad (9)\end{aligned}$$

which depends only on the ratio CV_x/CV_y and ρ.

On the basis of $Var(\hat{Y})$ and $Var(\hat{Y}_R)$, we can see that $Var(\hat{Y}_R) < Var(\hat{Y})$ if and only if $\rho > CV_x/(2CV_y)$. As compared with the definitions of (8) and (9), we can further show that $PRV_R < PRV_s$ if and only if $Var(\hat{Y}_R) < Var(\hat{Y})$ and hence $\rho > CV_x/(2CV_y)$. In fact, we can also prove this result by directly comparing Equation (8) with Equation (9). When $\rho = CV_x/(2CV_y)$, the two variances $Var(\hat{Y}_R) = Var(\hat{Y})$. In this case, the optimal weight w^* becomes 0.5 and both PRV_S and PRV_R reduce to $CV_x^2/(4CV_y^2)$. When $\rho = CV_x/(2CV_y)$ and $CV_x = 2CV_y$, PRV_S and PRV_R will equal 1 in this extreme case. On the other hand, if $\rho = CV_x/CV_y$, PRV_R will reduce to 0. This is because the composite estimator $\hat{Y}_C(w^*)$ with $w^* = 0$ will be the same as the ratio estimator $\hat{Y}_R$. This accounts for the reason why PRV_R is 0.

To estimate $Var(\hat{Y})$, $Var(\hat{Y}_R)$ and $Cov(\hat{Y}, \hat{Y}_R)$, we can substitute $\hat{\rho}$, cv_x, cv_y and $\bar{y}_s$ for ρ, CV_x, CV_y and $\bar{Y}$, respectively. Furthermore, to estimate the optimal weight w^*, we can similarly substitute estimators $\hat{\rho}$, cv_x and cv_y for ρ, CV_x and CV_y, and obtain $\hat{w}^*$.

Note that if we want to employ the composite estimator $w\hat{\bar{Y}} + (1-w)\hat{\bar{Y}}_R$ (where $\hat{\bar{Y}} = \bar{y}_s$ and $\hat{\bar{Y}}_R = \hat{R}\bar{X}$) to estimate the population mean $\bar{Y}$, the optimal weight w^* for the composite estimator minimizing variance $Var(w\hat{\bar{Y}} + (1-w)\hat{\bar{Y}}_R)$ will remain the same as given in (6). We denote the composite estimator $w^*\hat{\bar{Y}} + (1-w^*)\hat{\bar{Y}}_R$ with w^* by $\hat{\bar{Y}}_C(w^*)$. The variances $Var(\hat{\bar{Y}})$, $Var(\hat{\bar{Y}}_R)$ and $Var(\hat{\bar{Y}}_C(w^*))$ are simply equal to $Var(\hat{Y})/N^2$, $Var(\hat{Y}_R)/N^2$ and $Var(\hat{Y}_C(w^*))/N^2$, respectively.

3. Monte Carlo Simulations

Because the true optimal weight $w^* = 1 - \rho CV_y/CV_x$ depends on unknown parameters, we need to employ the estimator $\hat{w}^* = 1 - \hat{\rho}cv_y/cv_x$ calculated from data in use of the composite estimator in practice. This may inflate variance (or MSE) of the composite estimator $\hat{Y}_C(w^*)$ with known optimal weight. We employ Monte Carlo simulation to compare the performance of $\hat{Y}$, $\hat{Y}_R$, $\hat{Y}_C(\hat{w}^*)$ and $\hat{Y}_C(w^*)$ with respect to the bias and MSE. For specified values of S_x^2, S_y^2 and $\rho S_x S_y$, we first generate $N = 10000$ pairs of units (X_i, Y_i): $X_i = S_x Z_{1i} + \bar{X}$ and $Y_i = \rho S_y Z_{1i} + S_y\sqrt{1-\rho^2}Z_{2i} + \bar{Y}$, where Z_{1i} and Z_{2i} are all mutually independent and follow the standard normal distribution with mean 0 and variance 1.

We arbitrarily choose $\bar{X} = \bar{Y} = 10$ in the simulations. To assure that the occurrence of obtaining a simulated sample for which $X_i < 0$ or $Y_i < 0$ is rare, we focus our attention on CV_x and $CV_y \leq 1/2$. We consider the situations in which the coefficients of variation $CV_x = 0.33$, 0.5 and $CV_y = 0.25$, 0.33 the correlation between X_i and Y_i, $\rho = 0.60$, 0.80; and the sample size $n = 10$, 25, 50. These cover the situations in which the optimal weight w^* ranges from 0.20 to 0.70. For each configuration determined by a combination of the above parameters, we generate a population of $N = 10000$ pairs of units (X_i, Y_i) via the standard normal random number generator in SAS (2009) according to the procedure as described in the above. We then take 10,000 simple random samples, each being of size n, from the population in calculation of the bias and MSE for $\hat{Y}$, $\hat{Y}_R$, $\hat{Y}_C(\hat{w}^*)$ and $\hat{Y}_C(w^*)$. We calculate the simulated bias (or MSE) for an estimator $e(Y)$ of Y as the average of $(e(Y) - Y)$ (or the average of $(e(Y) - Y)^2$) over 10,000 simulated samples of size n. If we obtained $\hat{w}^* > 1$ in a simulated sample, we would set $\hat{w}^* = 1$. Similarly, if we obtained $\hat{w}^* < 0$, we would set $\hat{w}^* = 0$. To help readers easily see the relative performance of different estimators, we present the simulated MSE for $\hat{Y}_R$, $\hat{Y}_C(\hat{w}^*)$ and $\hat{Y}_C(w^*)$ relative to the MSE of the simple expansion estimator $\hat{Y}$. Note that because the bias for $\hat{Y}$ is theoretically zero, the simulated bias of an estimator $e(Y)$ relative to the simulated bias of $\hat{Y}$ can be potentially misleading. Thus, we present the bias itself for an estimator $e(Y)$. To help readers easily see where the bias of an estimator is non-negligible, we indicate the entry by a superscript "†" whenever its value is larger than 1,000 (or 1% of the expected population total 100,000 in simulations).

To study the usefulness of PRV_S (8) and PRV_R (9) in practice, we compare the underlying values for these parameters with those obtained by use of $\hat{Y}_C(\hat{w}^*)$ in the same set of configurations considered previously. We calculate the simulated PRV_S and PRV_R as

$$SPRV_S = [SMSE(\hat{Y}) - \{SMSE(\hat{Y}_C(\hat{w}^*)) - (SBIAS(\hat{Y}_C(\hat{w}^*)))^2\}]/SMSE(\hat{Y}), \qquad (10)$$

and

$$SPRV_R = [\{SMSE(\hat{Y}_R) - (SBIAS(\hat{Y}_R))^2\} - \{SMSE(\hat{Y}_C(\hat{w}^*)) - (SBIAS(\hat{Y}_C(\hat{w}^*)))^2\}] \\ /[SMSE(\hat{Y}_R) - (SBIAS(\hat{Y}_R))^2], \qquad (11)$$

where $SMSE(\hat{Y})$, $SMSE(\hat{Y}_R)$ and $SMSE(\hat{Y}_C(\hat{w}^*))$ are the simulated MSEs for $\hat{Y}$, $\hat{Y}_R$ and $\hat{Y}_C(\hat{w}^*)$, as well as $SBIAS(\hat{Y}_R)$ and $SBIAS(\hat{Y}_C(\hat{w}^*))$ are the simulated biases for $\hat{Y}_R$ and $\hat{Y}_C(\hat{w}^*)$, respectively.

4. Results

Table 1 summarizes the simulated bias and the simulated MSE relative to the MSE of $\hat{Y}$ for estimators $\hat{Y}$ $\hat{Y}_R$, $\hat{Y}_C(\hat{w}^*)$ and $\hat{Y}_C(w^*)$ in situations in which the coefficients of variation $CV_x = 0.33$, 0.5 and $CV_y = 0.25$, 0.33 the correlation between X_i and Y_i, $\rho = 0.60$, 0.80 and the sample size $n = 10$, 25, 50. As what we expect, the bias of the ratio estimator can be non-negligible, especially when the sample size n is not large ($= 10$). The biases for the other estimators are generally smaller than that of the ratio estimator (Table 1). When we examine the relative simulated MSE, the composite estimator $\hat{Y}_C(w^*)$ with known optimal weight has the smallest MSE (in boldface). Table 1 also suggests that the composite estimator $\hat{Y}_C(\hat{w}^*)$ can consistently outperform the ratio estimator $\hat{Y}_R$ with

Table 1. The simulated bias and the simulated MSE (in parenthesis) relative to the MSE of the simple expansion estimator $\hat{Y}$, for the ratio estimator $\hat{Y}_R$, the composite estimator $\hat{Y}_C(\hat{w}^*)$ with the estimated optimal weight and the composite estimator $\hat{Y}_C(w^*)$ with the known optimal weight in situations in which the coefficient of variation $CV_x = 1/3, 1/2$; $CV_y = 1/4, 1/3$; the correlation between X_i and Y_i, $\rho = 0.60, 0.80$; and the sample size $n = 10, 25, 50$. The entry with the bias larger than 1% of the expected population total is indicated by a superscript "†" and the entry with the smallest relative simulated MSE is printed in boldface. Each entry is calculated on the basis of 10,000 repeated samples.

CV_x	CV_y	ρ	n	$\hat{Y}$	$\hat{Y}_R$	$\hat{Y}_C(\hat{w}^*)$	$\hat{Y}_C(w^*)$
0.33	0.25	0.60	10	55	770	66	372
				(1.000)	(1.289)	(0.731)	(**0.662**)
			25	−66	279	−23	87
				(1.000)	(1.202)	(0.667)	(**0.637**)
			50	−34	149	−11	47
				(1.000)	(1.221)	(0.661)	(**0.651**)
		0.80	10	1	468	32	281
				(1.000)	(0.685)	(0.403)	(**0.363**)
			25	−57	150	−32	67
				(1.000)	(0.640)	(0.354)	(**0.342**)
			50	−64	105	−15	37
				(1.000)	(0.647)	(0.358)	(**0.353**)
	0.33	0.60	10	67	491	23	321
				(1.000)	(0.836)	(0.707)	(**0.653**)
			25	23	173	15	113
				(1.000)	(0.804)	(0.661)	(**0.640**)
			50	−42	66	−41	23
				(1.000)	(0.784)	(0.640)	(**0.631**)
		0.80	10	−107	115	−43	71
				(1.000)	(0.404)	(0.391)	(**0.360**)
			25	84	108	47	103
				(1.000)	(0.402)	(0.371)	(**0.361**)
			50	59	24	1	31
				(1.000)	(0.393)	(0.359)	(**0.352**)
0.50	0.25	0.60	10	87	2222†	165	728
				(1.000)	(3.172)	(0.743)	(**0.696**)
			25	45	655	−5	228
				(1.000)	(2.826)	(0.686)	(**0.667**)
			50	70	364	42	158
				(1.000)	(2.647)	(0.660)	(**0.650**)
		0.80	10	−155	1826†	−53	637
				(1.000)	(2.409)	(0.405)	(**0.400**)
			25	−20	552	−37	208
				(1.000)	(1.982)	(0.379)	(**0.374**)
			50	−49	412	18	135
				(1.000)	(1.875)	(0.365)	(**0.363**)
	0.33	0.60	10	16	1613†	−23	671
				(1.000)	(1.675)	(0.714)	(**0.655**)
			25	49	597	24	274
				(1.000)	(1.474)	(0.654)	(**0.638**)
			50	14	377	33	163
				(1.000)	(1.411)	(0.626)	(**0.622**)
		0.80	10	151	1241†	−2	730
				(1.000)	(1.029)	(0.405)	(0.**385**)
			25	36	461	−13	262
				(1.000)	(0.901)	(0.372)	(**0.366**)
			50	−26	214	−33	101
				(1.000)	(0.902)	(0.362)	(**0.358**)

Table 2. The underlying values for PRV_S (8) and PRV_R (9) and their corresponding simulated values $SPRV_S$ (10) and $SPRV_R$ (11) for using the composite estimator $\hat{Y}_C(\hat{w}^*)$ with the estimated optimal weight in situations in which the coefficient of variation $CV_x = 1/3, 1/2$; $CV_y = 1/4, 1/3$; the correlation between X_i and Y_i, $\rho = 0.60, 0.80$; and the sample size $n = 10, 25, 50$. Each entry is calculated on the basis of 10,000 repeated samples.

CV_x	CV_y	ρ	n	PRV_S	$SPRV_S$	PRV_R	$SPRV_R$
0.33	0.25	0.60	10	0.36	0.27	0.46	0.43
			25	0.36	0.33	0.46	0.44
			50	0.36	0.34	0.46	0.46
		0.80	10	0.65	0.60	0.45	0.41
			25	0.65	0.65	0.45	0.45
			50	0.65	0.64	0.45	0.45
	0.33	0.60	10	0.36	0.29	0.20	0.15
			25	0.36	0.34	0.20	0.18
			50	0.36	0.36	0.20	0.18
		0.80	10	0.64	0.61	0.10	0.03
			25	0.64	0.63	0.10	0.08
			50	0.64	0.64	0.10	0.09
0.50	0.25	0.60	10	0.35	0.26	0.75	0.76
			25	0.35	0.31	0.75	0.76
			50	0.35	0.34	0.75	0.75
		0.80	10	0.65	0.59	0.80	0.83
			25	0.65	0.62	0.80	0.81
			50	0.65	0.64	0.80	0.80
	0.33	0.60	10	0.37	0.29	0.55	0.57
			25	0.37	0.35	0.55	0.55
			50	0.37	0.37	0.55	0.55
		0.80	10	0.64	0.59	0.58	0.60
			25	0.64	0.63	0.58	0.59
			50	0.64	0.64	0.58	0.60

respect to the MSE, although one needs to estimate w^* from the sample data (Table 1). Comparing the relative MSE of $\hat{Y}_C(\hat{w}^*)$ with that of $\hat{Y}_C(w^*)$ reveals that the percentage of inflation in the MSE because of estimation for w^* is generally small for $n \geq 25$. For example, consider the case $CV_x = 1/3$, $CV_y = 1/3$, $\rho = 0.60$, and $n = 50$. The ratio of the MSE of $\hat{Y}_C(\hat{w}^*)$ relative to the MSE of $\hat{Y}_C(w^*)$ is 1.014.

To study the influence due to CV_x, CV_y and ρ, as well as the effect due to estimation of w^* on the PRV, we summarize in Table 2 PRV_S (8), $SPRV_S$ (10), PRV_R (9), and $SPRV_R$ (11) in the same set of configurations as those considered in Table 1. When $n \geq 25$, the values of $SPRV_S$ agree well with those of PRV_S, and so do the values of $SPRV_R$ compared to PRV_R. This is consistent with the findings that the loss of precision due to estimation of the optimal weight is relatively small for $n \geq 25$ (Table 1). When the ratio CV_y/CV_x is small ($= 1/2$), and ρ is large ($= 0.80$), the gain in efficiency (according to PRV_R) for using the composite estimator as compared with the ratio estimator seems to be the largest in situations considered in Table 2.

5. Examples

We consider two examples, one regarding the estimation of dead fir trees via an aerial photo and the other regarding the estimation of the average sugarcane acres per county. Because the number of sampled units is small in both two examples, the bias of the ratio

Table 3. The photo count and ground actual count of dead fir trees on ten sampled plots.

$i =$	Photo Count X_i	Ground Count Y_i	$i =$	Photo Count X_i	Ground Count Y_i
1	12	18	6	30	36
2	30	42	7	12	14
3	24	24	8	6	10
4	24	36	9	36	48
5	18	24	10	42	54

Scheaffer et al. 2012, 206.

estimator can be non-negligible. However, for the purpose of illustration only, we will assume that the bias of the ratio estimator is not of our concern, as done elsewhere (Cochran 1977; Scheaffer et al. 2012).

Consider the first example regarding a forest resource manager who wishes to estimate the total number of dead fir trees in a 300 acre area of heavy infestation by use of an aerial photo. The area was divided into ($N =$) 200 1.5-acre plots. We summarize in Table 3 the photo count (X_i) and the actual ground count (Y_i) of dead fir trees for a simple random sample of $n = 10$ plots (Scheaffer et al. 2012, 206). The total number of dead fir trees obtained from the photo account is ($X =$) 4200. Given these data (Table 3), we obtain the estimated coefficient of variation $cv_x = 0.490$, $cv_y = 0.485$ and $\hat{\rho} = 0.973$. These lead the optimal weight estimate to be $\hat{w}^* = 0.037$. Because $\widehat{Var}(\hat{Y})$ is much larger than $\widehat{Var}(\hat{Y}_R)$, the estimated optimal weight $\hat{w}^*$ associated with the component $\hat{Y}$ of the composite estimator $\hat{Y}_C(\hat{w}^*)$ is close to 0 (or the weight 1-$\hat{w}^*$ associated with the component $\hat{Y}_R$ is close to 1). We get the estimates: $\hat{Y} = 6120$, $\hat{Y}_R = 5492$, and $\hat{Y}_C(\hat{w}^*) = 5516$ with $\widehat{Var}(\hat{Y}) = 837858$, $\widehat{Var}(\hat{Y}_R) = 45890$, and $\widehat{Var}(\hat{Y}_C(\hat{w}^*)) = 44722$. The variance estimate for $\hat{Y}_C(\hat{w}^*)$ is smaller than those of $\hat{Y}$ and $\hat{Y}_R$; the corresponding estimates of PRVs are $\widehat{PRV}_S = 0.947$ and $\widehat{PRV}_R = 0.025$.

Consider the second example regarding the estimate of the average sugarcane acres per county over ($N =$) 32 counties in Florida, Hawaii, Louisiana and Texas (United States) in 1999. We summarize in Table 4 the data on ($n =$) the six sampled counties for years 1997 and 1999 (Scheaffer et al. 2012, 178). Given these data (Table 4), we obtain $cv_x = 0.995$, $cv_y = 0.930$ and $\hat{\rho} = 0.993$. The optimal weight estimate is $\hat{w}^* = 0.072$. Again, since $\widehat{Var}(\hat{Y})$ is much larger than $\widehat{Var}(\hat{Y}_R)$, the optimal weight $\hat{w}^*$ obtained here is also small. The average ($\bar{X} =$) is known to be 27,752 sugarcane acres over 32 counties in 1997. Given

Table 4. The number of sugarcane acres of the six sampled counties for year 1997 and year 1999 in the four states: Florida, Hawaii, Louisiana, and Texas.

County State	Hendry, FL	Kauai, HI	Saint Landry, LA	Calcasieu, LA	Iberia, LA	Cameron, TX
1997 Acreage X_i	54000	12300	9100	1700	57200	12900
1999 Acreage Y_i	57000	13900	15500	3900	59900	10400

Scheaffer et al. 2012, 178.

these data, we obtain the estimates: $\hat{\bar{Y}} = 26767$, $\hat{\bar{Y}}_R = 30278$, and $\hat{\bar{Y}}_C(\hat{w}^*) = 30025$ with variance estimates $\widehat{Var}(\hat{\bar{Y}}) = 83825986$, $\widehat{Var}(\hat{\bar{Y}}_R) = 1595550$ and $\widehat{Var}(\hat{\bar{Y}}_C(\hat{w}^*)) = 1097183$. These give $\widehat{PRV}_S = 0.987$ and $\widehat{PRV}_R = 0.312$. Thus, we may gain a moderate amount of efficiency by use of $\hat{\bar{Y}}_C(\hat{w}^*)$ as compared with $\hat{\bar{Y}}(= \bar{y}_s)$ and $\hat{\bar{Y}}_R(= \hat{R}\bar{X})$, although the weight $1 - \hat{w}^* = 0.928$ associated with the ratio estimator in the composite estimator is large. This example may illustrate the situation in which $CV_x > CV_y$, using the composite estimator can be of use to improve the precision of the ratio estimator even for a large correlation ρ. We may wish to note that the estimated gain in efficiency presented in both examples does not account for the uncertainty due to estimation of the optimal weight. Thus, to further confirm this finding of gaining efficiency in the above cases with high correlation and small sample size, we have carried out additional similar simulations as described previously for correlation $\rho = 0.90, 0.95$. We have obtained the same results on the relative performance among different estimators with respect to the MSE as found for $\rho = 0.80$ in Table 1. For brevity, we do not present these simulation results, which are, however, available to readers upon request.

6. Discussion

The composite estimator $\hat{Y}(w^*)$ with known optimal weight can be shown to outperform the ratio estimator with respect to both bias and variance algebraically. Furthermore, the simulation results presented here suggest that both the bias and MSE of the composite estimator with the estimated optimal weight can be still smaller than those of the ratio estimator. Because there is no need for us to obtain any extra information besides the data required to apply the ratio estimator, the composite estimator proposed here should be of use to improve the efficiency of the ratio estimator in practice. We also wish to note that it is simply straightforward to extend the composite estimators discussed here to accommodate the stratified random sampling, in which the SRS is employed within each stratum.

When estimating the population mean $\bar{Y}$, Cochran (1977, 190–192) considered a class of regression estimators $\{\bar{y}_s + b_0(\bar{X} - \bar{x}_s)$, where b_0 is preassigned fixed constant$\}$. It is easy to see that the composite estimator $\hat{\bar{Y}}(w^*)$ can be rewritten as $\bar{y}_s + \rho(CV_y/CV_x)\hat{R}(\bar{X} - \bar{x}_s)$, which is not a member of the above class of regression estimators. This is because $\rho(CV_y/CV_x)\hat{R}$ is random rather than a fixed constant. However, the composite estimator $\hat{\bar{Y}}(w^*)$ with the estimated optimal weight $\hat{w}^*$ will lead to the estimator $\bar{y}_s + \hat{b}_0(\bar{X} - \bar{x}_s)$, where $\hat{b}_0^* = \hat{\rho}(s_y/s_x)$ is the estimator for the constant b_0^* $(= \rho(CV_y/CV_x))$ that minimizes the variance $Var(\bar{y}_s + b_0(\bar{X} - \bar{x}_s))$ under SRS (Cochran 1977, 191).

Note that we focus our attention on the classical finite-population sampling, in which the expectation is taken with respect to sampling scheme (or random permutations). As long as the auxiliary variate X_i is highly correlated with Y_i, we may consider use of the ratio estimator accounting for the auxiliary information to improve the precision of the simple expansion estimator without making assumptions of any model relationship between Y_i and X_i (Cochran 1977; Scheaffer et al. 2012; Thompson 2012). Because the composite estimator includes the ratio estimator as a special case, the composite estimator with known optimal weight will have, as noted previously, the MSE smaller than the ratio estimator for all bivariate models of (X_i, Y_i) under the SRS. Furthermore, we may find numerous real-life examples illustrating the use of ratio estimator under the

SRS in data likely fitting the common linear regression model $Y_i = \alpha + \beta X_i + e_i$, where random errors e_i have mean equal to 0 and constant variance $Var(e_i)$ (Scheaffer et al. 2012, 173–186). In our simulations, we assume that the bivariate normal model for (X_i, Y_i) in which all assumptions of the linear regression model with constant variance are implicitly satisfied (Casella and Berger 1990). Our simulated results (Table 1) suggest that use of composite estimators even with use of the estimated optimal weights may still substantially reduce the bias and MSE of the ratio estimator in these cases. On the other hand, the ratio estimator is often considered, provided that the population model: $Y_i = \beta X_i + e_i$, where random errors e_i have mean equal to 0 and variance $Var(e_i)$ proportional to X_i is assumed under the super-population model-based approach (SPMBA) (Royall 1970). This is because the ratio estimator $\hat{Y}_R$ is the best linear unbiased estimator for the population total Y with respect to random errors assumed in this linear model under the SPMBA (Cochran 1977, 158–159). However, for a given correlation ρ between X_i and Y_i, creating a bivariate model for $(X_i,\ Y_i)$ with the conditional mean $E(Y_i|X_i) = \beta X_i$ and conditional variance $Var(Y_i|X_i) = \sigma^2 X_i$ is extremely difficult. Thus, a topic for future research is to extend our simulation results under the SRS for a given fixed correlation ρ to cover cases with the conditional variance $Var(Y_i|X_i)$ proportional to X_i when one needs to estimate the optimal weight for the composite estimator from data.

Finally, when X_i and Y_i are dichotomous (i.e., X_i and $Y_i = 1$ if subject i possesses the characteristics of interest, and $= 0$, otherwise), we note that one may follow the same arguments as presented here and derive the optimal weight minimizing the variance (or MSE) of the composite estimator under the SRS. We can further derive the PRVs for using the composite estimator as compared with the ratio and simple expansion estimators. We leave some details in Appendix 2 (Subsection 7.2.) for readers' information.

7. Appendix

7.1. *Appendix 1*

Note that

$$\begin{aligned} E(\hat{Y} - Y)(\hat{Y}_R - Y) &= E(\hat{Y} - Y)(\hat{Y}_R - E(\hat{Y}_R) + E(\hat{Y}_R) - Y) \\ &= E(\hat{Y} - Y)(\hat{Y}_R - E(\hat{Y}_R)) + E(\hat{Y} - Y)(E(\hat{Y}_R) - Y). \end{aligned} \tag{A.1}$$

Because $(E(\hat{Y}_R) - Y)$ is constant and $E(\hat{Y}) = Y$, the second part in (A.1) is zero. Therefore, we have $Cov(\hat{Y}, \hat{Y}_R) = E(\hat{Y} - Y)(\hat{Y}_R - Y)$.

Furthermore, note that

$$\begin{aligned} E(\hat{Y} - Y)(\hat{Y}_R - Y) &= E(N\bar{y}_s - Y)(\hat{R}X - Y) = N^2\bar{X}E(\bar{y}_s - \bar{Y})((\bar{y}_s - R\bar{x}_s)/\bar{x}_s) \\ &\approx N^2E(\bar{y}_s - \bar{Y})((\bar{y}_s - \bar{Y} - R(\bar{x}_s - \bar{X})) \\ &= N^2[E(\bar{y}_s - \bar{Y})^2 - RE(\bar{y}_s - \bar{Y})(\bar{x}_s - \bar{X})] \\ &= N^2(1 - f)[S_y^2 - RS_{xy}]/n \\ &= N^2(1 - f)\bar{Y}^2[CV_y^2 - \rho CV_x CV_y]/n. \end{aligned} \tag{A.2}$$

Note also that

$$\begin{aligned} E(\hat{Y}_R - Y)^2 &= N^2\bar{X}^2E(\hat{R} - R)^2 = N^2\bar{X}^2E((\bar{y}_s - R\bar{x}_s)/\bar{x}_s)^2 \\ &\approx N^2E(\bar{y}_s - R\bar{x}_s)^2 = N^2E(\bar{y}_s - \bar{Y} - R(\bar{x}_s - \bar{X}))^2 \\ &= N^2[E(\bar{y}_s - \bar{Y})^2 + R^2(\bar{x}_s - \bar{X})^2 - 2RE(\bar{y}_s - \bar{Y})(\bar{x}_s - \bar{X})] \\ &= N^2(1-f)[S_y^2 + R^2S_x^2 - 2RS_{xy}]/n \\ &= N^2(1-f)\bar{Y}^2[CV_y^2 + CV_x^2 - 2\rho CV_xCV_y]/n. \end{aligned} \tag{A.3}$$

Because $E(\hat{Y}_R)$ is actually not equal to Y, it can be more appropriate to call $E(\hat{Y}_R - Y)^2$ (A.3) the MSE of $\hat{Y}_R$ as what was done by Sukhatme et al. (1984, 192). However, the relative bias $E(\hat{Y}_R - Y)/Y \approx (1-f)CV_x^2(1 - \rho CV_y/CV_x)/n$ converges to 0 as n increases. Therefore, despite of $E(\hat{Y}_R) \neq Y$, Formula (A.3) is also the traditional variance formula used for $Var(\hat{Y}_R)$ in many textbooks (Cochran 1977; Scheaffer et al. 2012; Govindarajulu 1999; Thompson 2012).

Because $E(\hat{Y}_R - Y)^2 \approx Var(\hat{Y}_R)$ and $Cov(\hat{Y}, \hat{Y}_R) = E(\hat{Y} - Y)(\hat{Y}_R - Y)$, the optimal weight w^* minimizing the variance $Var(\hat{Y}_C(w))$ of the composite estimator $\hat{Y}_C(w) = w\hat{Y} + (1-w)\hat{Y}_R$ is equivalent to minimizing the MSE $E(\hat{Y}_C(w) - Y)^2$. To derive the optimal weight w^* minimizing $Var(\hat{Y}_C(w))$, we set the first derivative of $Var(\hat{Y}_C(w))$ taken with respect to w equal 0 and solve the equation.

7.2. *Appendix 2*

When X_i (and Y_i) = 1 if subject i possesses the characteristics of interest, and = 0, otherwise. One can show that $CV_x = \sqrt{N(1-P_x)/[(N-1)P_x]}$, where $P_x = \sum_{i=1}^{N} X_i/N$ represents the proportion of subjects possessing the characteristics of interest in the population of X_i. Similarly, $CV_y = \sqrt{N(1-P_y)/[(N-1)P_y]}$, where $P_y = \sum_{i=1}^{N} Y_i/N$ represents the proportion of subjects possessing the characteristics of interest in the population of Y_i. Note that ρ is simply identical to the phi φ correlation for dichotomous data defined elsewhere (Fleiss et al. 99). Thus, the optimal weight w for the composite estimator $\hat{Y}_C(w) = w(N\hat{p}_y) + (1-w)(\hat{p}_y/\hat{p}_x)(NP_x)$ (where $\hat{p}_y = \sum_{i \in c} y_i/n$ and $\hat{p}_x = \sum_{i \in c} x_i/n$) minimizing variance $Var(\hat{Y}_C(w))$ is

$$w^* = 1 - \varphi\sqrt{OR_{xy}} \tag{A.4}$$

where $OR_{xy} = P_x(1-P_y)/[(1-P_x)P_y]$ is the odds ratio (OR) of the marginal proportions of possessing the characteristics between populations of X_i and of Y_i. Note that this OR_{xy} is different from the OR, $(p_{11}p_{00})/(p_{10}p_{01})$, where $p_{rs} = P(X_i = r, Y_i = s)$, $r = 1, 0$ and $s = 1, 0$. The latter OR is related to the φ correlation, measuring the strength of association between X_i and Y_i. Note that the OR has many mathematically inherent and desirable properties and is one of the most common-used indices measuring the extent of association between variables in Epidemiology (Lui 2004, 89–90). On the basis of Formula (A.4), we see that the higher the φ or the larger the OR_{xy}, the smaller is the optimal weight w^*

(subject to $w^* \geq 0$). The PRV based on (8) by use of $\hat{Y}_C(w^*)$ instead of $\hat{Y}$ becomes

$$PRV_S = \varphi^2. \tag{A.5}$$

Furthermore, the PRV based on (9) by use of $\hat{Y}_C(w^*)$ instead of $\hat{Y}_R$ is

$$PRV_R = (1 - \varphi\sqrt{OR_{xy}})^2/[(1 - \varphi\sqrt{OR_{xy}})^2 + OR_{xy}(1 - \varphi^2)]. \tag{A.6}$$

When the prevalence rate of possessing the characteristics of interest between populations X_i and Y_i are equal (*i.e.,* $OR_{xy} = 1$), PRV_R Formula (A.6) simplifies to $(1 - \varphi)/2$. A systematic comparison of the performance for the composite estimator with the estimated optimal weight minimizing the variance (or the MSE) versus its two component estimators in dichotomous data can be a small research project.

8. References

Casella, G. and R.L. Berger. 1990. *Statistical Inference*. Belmont, CA: Duxbury.

Cochran, W.G. 1977. *Sampling Techniques* (3rd ed.). New York: Wiley.

Fleiss, J.L., B. Levin, and M.C. Paik. 2003. *Statistical Methods for Rates and Proportions* (3rd ed.). New York: Wiley. DOI: https://doi.org/10.1002/0471445428.

Govindarajulu, Z. 1999. *Elements of Sampling Theory and Methods*. Upper Saddle River, NJ: Prentice Hall.

Lee, S.E., P.R. Lee, and K-II. Shin. 2016. "A Composite Estimator for Stratified Two-Stage Cluster Sampling." *Communications for Statistical Applications and Methods* 23: 47–55. DOI: https://doi.org/10.5351/CSAM.2016.23.1.047.

Lui, K.-J. and W.G. Cumberland. 1991. "A Model-Based Approach: Composite Estimators for Small Area Estimation." *Journal of Official Statistics* 7: 69–76.

Lui, K.-J. 2004. *Statistical Estimation of Epidemiological Risk*. New York: Wiley. DOI: https://doi.org/10.1002/0470094087.

Pascual, J.N. 1961. "Unbiased Ratio Estimators in Stratified Sampling." *Journal of American Statistical Association* 56: 70–87. DOI: https://doi.org/10.2307/2282332.

Royall, R.M. 1970. "On Finite Population Sampling Theory under Certain Linear Regression Models." *Biometrika* 57: 377–387. DOI: https://doi.org/10.1093/biomet/57.2.377.

Sahoo, L.N. 1987. "On a Class of Almost Unbiased Estimators for Population Ratio." *Statistics* 18: 119–121. DOI: https://doi.org/10.1080/02331888708801998.

SAS Institute Inc. 2009. *SAS/STAT 9.2 User's Guide* (2nd ed.). Cary, NC: SAS Institute.

Schaible, W.L. 1978. "Choosing Weight for Composite Estimators for Small Area Statistics." In *Proceedings of the Section on Survey Research Methods*, 741–746. Washington: American Statistical Association.

Schaible, W.L. 1979. "A Composite Estimator for Small Area Statistics." In *Proceedings of the Section on Social Statistics*, 1017–1021. Washington: American Statistical Association

Scheaffer, R.L., W. Mendenhall III, R.L. Ott, and K.G. Gerow. 2012. *Elementary Survey Sampling* (7th ed.). Boston, MA: Brooks/Cole.

Sukhatme, P.V., B.V. Sukhatme, S. Sukhatme, and C. Asok. 1984. *Sampling Theory of Surveys Applications* (3rd ed.). Ames, Iowa: Iowa State University Press.

Thompson, S.K. 2012. *Sampling* (3rd ed.). New York: Wiley. DOI: https://doi.org/10.1002/9781118162934.

Received November 2018
Revised April 2019
Accepted November 2019

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 151–172, http://dx.doi.org/10.2478/JOS-2020-0008

An Appraisal of Common Reweighting Methods for Nonresponse in Household Surveys Based on the Norwegian Labour Force Survey and the Statistics on Income and Living Conditions Survey

Nancy Duong Nguyen[1] *and Li-Chun Zhang*[2]

Despite increasing efforts during data collection, nonresponse remains sizeable in many household surveys. Statistical adjustment is hence unavoidable. By reweighting the design, weights of the respondents are adjusted to compensate for nonresponse. However, there is no consensus on how this should be carried out in general. Theoretical comparisons are inconclusive in the literature, and the associated simulation studies involve hypothetical situations not all equally relevant to reality. In this article we evaluate the three most common reweighting approaches in practice, based on real data in Norway from the two largest household surveys in the European Statistical System. We demonstrate how cross-examination of various reweighting estimators can help inform the effectiveness of the available auxiliary variables and the choice of the weight adjustment method.

Key words: Unit nonresponse; auxiliary variable selection: inverse propensity weighting; generalised regression estimation; doubly robust estimation.

1. Introduction

Response rates in household surveys have declined steadily in many Western countries (De Leeuw and De Heer 2002; Stoop et al. 2010; Meyer et al. 2015). Post data collection, statistical adjustment is needed due to a sizeable amount of nonresponse. A standard process to compensate for unit nonresponse is reweightings (Little 1986; Kalton and Flores-Cervantes 2003; Särndal and Lundström 2005; Brick 2013). Generally speaking, this requires making two interrelated decisions on *auxiliary variable selection* and *weight adjustment method*. However, there is no consensus on a general approach.

We distinguish between the three most common reweighting approaches in practice. Firstly, the *two-step* approach combines response propensity weighting (from respondents to sample) and calibration (from sample to population); see, for example Kalton and Kasprzyk (1986). In general two *different* sets of auxiliary variables are used at the two steps. The first step weight may either be directly given by the inverse of the estimated response propensities (Cassel et al. 1983; Little and Rubin 1987), or indirectly based on adjustment cells formed using these propensities (Little 1986; Eltinge and Yansaneh

[1] School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Ireland. Email: duong.nguyen@ucdconnect.ie
[2] Department of Social Statistics and Demography, University of Southampton, Southampton, UK. Email: L.Zhang@soton.ac.uk

1997). Secondly, applying calibration of the sampling weights (from respondents to population) directly yields the *one-step* approach (Lundström and Särndal 1999), for which a set of auxiliary variables should ideally have high association with both the response indicator and the target outcome variable. Adopting the linear calibration function yields the modified generalised regression (MGR) estimator (Bethlehem 1988). Thirdly, using the *same* covariates for both response propensity modelling and calibration, the two-step approach could yield the *doubly robust (DR)* estimators (see e.g., Robins et al. 1994; Robins and Wang 2000; Bang and Robins 2005; Carpenter et al. 2006; Kang and Schafer 2007).

Despite their long tradition, the choice between the two- and one-step approaches is still not conclusive in the literature. For instance, one may easily find motivations for the one-step approach (Little and Vartivarian 2005; Särndal and Lundström 2008, 2010), but there exist also several warnings against its potential pitfalls (Brick 2013; Kott and Liao 2015; Haziza and Lesage 2016). Although the DR estimators have attracted much attention outside the field of survey sampling, we did not come across any reports on their performance in real household (or business) surveys.

We believe theoretical comparisons are unable to reach a clear-cut choice because the ‘true’ nonresponse mechanism cannot be identified based on the observed data alone. Moreover, while simulation studies are useful for illustrating certain properties of one approach or another, not all the hypothetical set-ups are relevant to reality. It is therefore essential to examine situations in actual household surveys, which are limited in number. For instance, in the context of European Statistical System (ESS), there are currently only about ten major household surveys. Moreover, relevant auxiliary variables consist mostly (or entirely) of categorical variables, unlike what is common in simulation studies.

In this article, we assess *empirically* the three reweighting approaches outlined above, based on the Norwegian Labour Force Survey (LFS) and Survey of Income and Living Conditions (SILC), which are the two largest household surveys in the ESS. The protocol of the appraisal is generally applicable to other surveys or countries.

We begin with a description of the sampling designs of the Norwegian LFS and the SILC in Section 2. In Section 3, we describe a set of reweighting estimators to be investigated and some common variations. Then, we introduce simple Analysis of Variance (ANOVA)-type measures to understand the potential effects of an auxiliary variable based on its association with the outcome variable and the response indicator in Section 4, and use real data to illustrate how these may be related to the resulting change in the point estimate and the associated variance. Our discussion brings forward greater nuances of the reweighting effects than those that have been delineated previously by Thomsen (1973, 1978), Oh and Scheuren (1983) and Little and Vartivarian (2005). In Section 5 we present an empirical study of the Norwegian LFS and SILC data. As will be demonstrated, cross-examination of the different point estimates and their variances can inform the effectiveness of the available auxiliary variables and the choice of the weight adjustment method. Some general conclusions that emerge from the empirical appraisal will be summarised in Section 6.

In summary, regarding auxiliary variable selection, we find that it is always useful to increase the association with the outcome variable, but seeking the highest possible association with nonresponse is not necessarily helpful. Moreover, we find that the choice of weight adjustment method matters, especially when there exist strong auxiliary variables for the outcome available; whereas provided only weak auxiliary variables for the outcome variable, limiting the loss of efficiency and avoiding spurious adjustment may be a relevant priority. Overall, we found no evidence in the situations examined to support an uncritical adoption of the two-step approach. Since the 'true' nonresponse model envisaged for a two-step approach cannot be identified based on the observed data, regardless of whether the available auxiliary variables have low or high association with nonresponse, it makes sense to choose based on cross-examination of the alternatives in a given situation.

2. Sampling Designs

In this article we use or relate the discussions to the LFS data in Subsections 4.2, 4.3. and 5.1., and the SILC data in Subsection 5.2. We now briefly describe the sampling designs of these two surveys and a relevant variable called *Panel Response Status*.

2.1. The LFS

The Norwegian LFS has a stratified cluster sampling design, where the 19 counties make up the strata and family units form the clusters. The population register provides the sampling frame. The target population consists of residents aged 15–74 years old in Norway. Every in-scope person stays in the LFS for eight quarters, and there is approximately an 7/8 overlap between two consecutive quarters. The quarterly sample contains approximately 24,000 individuals, and the current response rate is around 80%. All interviews are conducted by telephone.

The overlap between two consecutive quarters means that approximately one in eight persons is new in each quarterly sample. It is possible to create a variable called Panel Response Status that identifies every person as new in sample, or previous quarter respondent, or previous quarter nonrespondent. This variable has very high association with the current quarter response indicator, in that previous quarter respondents (or nonrespondents) are more likely to respond (or not respond) again. Later on we will use this Panel Response Status to demonstrate the effects of a variable that has high association with the response indicator on the point estimate and the variance of an estimator.

2.2. The SILC

The annual SILC collects data on housing, finance, health, and work, and so on. The target population consists of residents aged 16 years and over and not living in institutions. It has a four-year rotating panel design. Individuals are selected from the population register using the SRS design. The interviews are largely conducted over telephone, although face-to-face interviews can take place by way of exception. Just like with the LFS, the panel design of the SILC allows one to create the Panel Response Status variable, distinguishing new persons in the sample, previous year respondents and previous year nonrespondents.

3. Reweighting Estimators to be Investigated

Consider a finite population U of size N. Let Y be an outcome variable of interest which takes the value y_i for unit $i \in U$. Assume that a sample s of size n is selected from U by probability sampling, where π_i is the inclusion probability and $d_i = 1/\pi_i$ is the design weight of unit $i \in s$. Let R be the response indicator defined as $r_i = 1$ if unit i responds and $r_i = 0$ otherwise, for $i \in s$. Let r denote the respondent sample of n_r units such that $r \subset s$ and $n_r < n$. We describe the various methods to be included in a schematic investigation in terms of the estimator of the population total $t = \sum_{i \in U} y_i$.

As a baseline for comparison, consider the design weighted estimator

$$\hat{t}_d = \frac{n}{n_r} \sum_{i \in r} d_i y_i. \tag{1}$$

This estimator takes the sampling design into account, and is approximately unbiased for t provided nonresponse *missing completely at random* (MCAR, Little and Rubin 1987). An alternative baseline estimator is the sample respondent expansion estimator

$$\hat{t} = \frac{N}{n_r} \sum_{i \in r} y_i. \tag{2}$$

It is unbiased for t provided MCAR and equal probability selection method (epsem), and allows one to gauge both the effects of sampling design and nonresponse on reweighting. In many household surveys, epsem holds either exactly or approximately, such that the difference between $\hat{t}_d$ and $\hat{t}$ may be small, when compared to the various reweighting estimators described below, which aim to adjust for the potential bias caused by nonresponse.

To begin with, when it comes to auxiliary variable selection, it is often recommended to select variables that have high association with both the survey variable (Y) and the response indicator (R); see, for example Little and Vartivarian (2005), Schouten (2007), Särndal and Lundström (2008) or Bethlehem et al. (2011). In practice, instead of building a bivariate model of (Y, R), it is common to model R and Y *separately*. Denote by Z the selected predictors of the R-model and by X those of the Y-model. The two generally do not coincide. Not all the variables in Z (or X) are equally important to R (or Y). In a sense, one may consider the variables in the joint subset, denoted by $A = Z \wedge X$, to be explanatory of both R and Y, but we are unaware of any recommended reweighting approach that *only* makes use of A. There exist also other variable selection approaches that are not based on explicit R- and Y-modelling; see, for example Schouten (2007) and Särndal and Lundström (2010). However, we shall focus on the modelling approach to auxiliary variable selection in this article, because it is more generally applicable and has a more direct connection to the weight adjustment methods, as will be explained shortly. Notice that in this article, we consider the y-values in the population to be fixed when calculating the expectation and variance of an estimator, even when Y-modelling is used to 'assist' its construction.

Denote the response propensity of unit i, for $i \in s$, by

$$p_i = p(z_i; \alpha) = \Pr(r_i = 1 | z_i)$$

for example, defined via a logistic regression model. Let

$$\mu_i = E(Y_i | x_i) = m(x_i; \beta)$$

be the conditional expectation of Y_i given x_i. For illustration, we shall assume the most common linear regression, that is, $\mu_i = x_i^T \beta$; but other types of regression models of μ_i are equally feasible. The two-step weight adjustment that uses Z and X separately can now be given as

$$\hat{t}_{2sts} = \sum_{i \in U} m(x_i; \hat{\beta}) + \sum_{i \in r} \frac{d_i}{p(z_i; \hat{\alpha})} \{y_i - m(x_i; \hat{\beta})\}, \tag{3}$$

where $\hat{\beta} = \left[\sum_{i \in r} d_i x_i x_i^T / p(z_i; \hat{\alpha})\right]^{-1} \sum_{i \in r} d_i x_i y_i / p(z_i; \hat{\alpha})$, and $\hat{\alpha}$ is the estimator of α, which is typically obtained from fitting an appropriate logistic regression model to the sample by solving for $\sum_{i \in s} z_i[r_i - p(z_i; \alpha)] = 0$. The estimator (3) is approximately unbiased for t provided nonresponse is missing-at-random (MAR, Little and Rubin 1987) given Z, and the model of p_i is correctly specified.

By itself, the first step of (3) yields the Inverse Propensity Weighting (IPW) estimator

$$\hat{t}_{IPW} = \sum_{i \in r} \frac{d_i}{p(z_i; \hat{\alpha})} y_i. \tag{4}$$

It is approximately unbiased under the same condition as (3), but may be less efficient if X can help reduce the variance. Extreme weights can arise by IPW, when large weights are assigned to relatively few respondents with similar characteristics to nonrespondents. Some authors propose to stratify the sample into several groups (or adjustment cells) based on similar $p(z_i; \hat{\alpha})$, that is, Response Propensity Stratification (RPS), and use the inverse within-group response rate as the 1st-step weight. RPS is reported to be more efficient than IPW in some studies (Little 1986; Kang and Schafer 2007), although Lunceford and Davidian (2004) warn against their routine use based on their theoretical and empirical results. In general, while potential modification of the IPW weight $p(z_i; \hat{\alpha})^{-1}$ is always a relevant practical issue, the IPW weight is more easily interpretable when comparisons are made to other weight adjustment methods. We recommend $\hat{t}_{IPW}$ to be computed and included in a schematic investigation of reweighting methods.

Next, applying the second weight adjustment of (3) directly to the respondents yields the one-step MGR estimator

$$\hat{t}_{MGR} = \sum_{i \in U} m(x_i; \hat{B}) + \sum_{i \in r} d_i \{y_i - m(x_i; \hat{B})\}, \tag{5}$$

where $\hat{B} = \left[\sum_{i \in r} d_i x_i x_i^T\right]^{-1} \sum_{i \in r} d_i x_i y_i$. As mentioned before, other one-step calibration estimators are possible by other calibration functions. But the linear calibration (5) is the most routine choice, and we shall focus on it to compare the one-step approach to other adjustment methods. The MGR estimator is approximately unbiased, if nonresponse is MAR given X, and if the linear model of μ_i is correctly specified or if the response propensity p_i is the inverse of a linear combination of x_i (Lundström and Särndal 1999). An extra feature sometimes included in the discussion of the one-step approach is when some variables in X are observed in the whole sample but have unknown population totals (Särndal and Lundström 2005; Andersson and Särndal 2016). However, this is not an essential difference to the two-step approach, because the same possibility can as well be accommodated by the two-step approach.

Now, the variables Z selected by R-modelling generally differ from X by Y-modelling. Moreover, none of the associated MAR assumptions can be entirely true. Under the DR

Table 1. A minimal set of reweighting estimators.

Selection and use of auxiliary variable	Weight adjustment method		
	One-step IPW	One-step MGR	Two-step
Separate R- and Y-modelling	$\hat{t}_{IPW}(Z,-)$	$\hat{t}_{MGR}(-,X)$	$\hat{t}_{2sts}(Z,X)$
Refitting *after* R- and Y-modelling	$\hat{t}_{IPW}(V,-)$	$\hat{t}_{MGR}(-,V)$	$\hat{t}_{DR}(V,V)$

approach, one uses the same variables to build an R-model and a Y-model; see, for example Kim and Haziza (2014). The resulting estimator is approximately unbiased if either one of the two models is correctly specified. In practice, without actually building a bivariate (R, Y)-model, taking the auxiliary variables $V = Z \vee X$, as the union of Z and X following separate R- and Y-modelling, appears a likely course of variable selection. The DR estimator for t that we adopt for this study is thus given by applying the two-step approach (3) to (V, V) instead of (Z, X), that is,

$$\hat{t}_{DR} = \sum_{i\in U} m(v_i;\hat{\xi}) + \sum_{i\in r}\frac{d_i}{p(v_i;\hat{\eta})}\{y_i - m(v_i,\hat{\xi})\}, \tag{6}$$

where $\hat{\xi} = \left[\sum_{i\in r} d_i v_i v_i^T/p(v_i;\hat{\eta})\right]^{-1}\sum_{i\in r} d_i v_i y_i/p(v_i;\hat{\eta})$ under the linear Y-model $\mu_i = v_i^T\xi$, and $\hat{\eta}$ is the estimator of the R-model parameter η in $p_i = p(v_i;\eta)$. Note that this requires known population total of z_i, unlike the IPW estimator for which one only needs the z_i's in the sample. Provided nonresponse is MAR given V, the estimator (6) is approximately unbiased when either the R- or Y-model is correctly specified. Notice that unless separate modelling happens to result in $Z = X$, adopting $V = Z \vee X$ would imply over-fitting for p_i or μ_i. However, in the situation of $v_i = x_i$, Lunceford and Davidian (2004) demonstrate the potential gains of the DR approach, that is, to “over-model” $p(z_i;\alpha)$ by $p(v_i;\eta)$. So it is of interest to investigate the performance of $\hat{t}_{DR}$, despite the heuristic construction of V.

Thus we arrive at a *minimal* set of estimators for a schematic investigation in any given situation (Table 1). Also specified are the respective auxiliary variables to be used for each reweighting estimator. For the estimators using $V = Z \vee X$, refitting of $p_i(v_i;\eta)$ and $\mu_i(v_i;\xi)$ is needed in practice. Cross-examination of the different point estimates and their associated variances in a given survey will be illustrated in Section 5.

4. Effects of Auxiliary Variable

4.1. Subclass Reweighting and Association Measures

Not all the selected variables in Z or X are equally effective. To gauge the potential effects of a categorical auxiliary variable, $c = 1, 2, \ldots, C$, let the population be partitioned accordingly into C subclasses with known population sizes $N_1, \ldots, N_C$, and $N = \sum_{c=1}^{C} N_c$. Let each subclass consist of a respondent stratum and a nonrespondent stratum (Cochran 1953), respectively, of the population sizes N_c' and N_c'' and means $\bar{Y}_c'$ and $\bar{Y}_c''$. Let $\bar{Y}' = \sum_c N_c'\bar{Y}_c'/N'$ be the population respondent mean, where $N' = \sum_c N_c'$, and $\bar{Y}'' = \sum_c N_c''\bar{Y}_c''/N''$ the population nonrespondent mean, where $N'' = \sum_c N_c''$.

Let $\bar{Y} = \bar{Y}'N'/N + \bar{Y}''N''/N$ be the population mean. Consider the unweighted sample respondent mean

$$\bar{y} = \sum_{i \in r} y_i / n_r = \hat{t}/N,$$

as an estimator of $\bar{Y} = t/N$, against the reweighted respondent mean

$$\bar{y}_W = \sum_{c=1}^{C} W_c \bar{y}_c,$$

where $W_c = N_c/N$ and $\bar{y}_c$ is the respondent mean in sample subclass c.

The set-up is convenient for several reasons. Previously, Thomsen (1973 1978), Oh and Scheuren (1983) and Little and Vartivarian (2005) all used it to study the effects of reweighting, which is natural for household surveys where the auxiliary variables are either categorical or can be categorised, and the subclasses may arise from cross-classifying several variables. Based on subclasses $1, \ldots, C$, all the reweighting estimators described in Section 3 reduce to $\bar{y}_W$, provided simple random sampling (SRS), which allows us to isolate away the choice of adjustment method. Moreover, one can estimate the randomisation variances of $\bar{y}$ and $\bar{y}_W$ based on the observed sample (Thomsen 1978), where the population y- and r-values are treated as fixed. As pointed out by Little and Vartivarian (2005), the SRS assumption allows one to gain an appreciation of the relative efficiency, that is, RE $= \mathrm{Var}(\bar{y}_W)/\mathrm{Var}(\bar{y})$, without complicating the technical details due to complex designs. Notice that, even when the sampling design is complex, or if one prefers the model-based or quasirandomisation-based inference in the end, it is still possible to make use of the randomisation-based results below, obtained under the SRS assumption, in order to easily gauge the potential effects of an auxiliary variable.

Now, to examine the change of the point estimate due to subclass reweighting, let

$$B = E(\bar{y} - \bar{y}_W) = \frac{1}{\bar{h}} \sum_{c=1}^{C} W_c \bar{Y}_c'(h_c - \bar{h}) = \frac{1}{\bar{h}} \sum_{c=1}^{C} W_c(\bar{Y}_c' - \bar{Y}')(h_c - \bar{h}), \qquad (7)$$

where $h_c = N_c'/N_c$, for $h_c > 0$, is the population subclass respondent proportion, and $\bar{h} = \sum_c W_c h_c$ is the population respondent proportion. The second last expression in (7) is given by Thomsen (1973), and the last one follows since $\sum_c W_c(h_c - \bar{h}) = 0$. Considering $\{W_1, \ldots, W_C\}$ as a probability mass function, one may interpret B as the covariance between $\bar{Y}_c'$ and h_c as c varies, denoted by $\mathrm{Cov}_W(\bar{Y}_c', h_c)$. Since $\bar{h}$ is fixed at the estimation stage, different subclass formations can only affect $\mathrm{Cov}_W(\bar{Y}_c', h_c)$. Thus, B would be large if either $\bar{Y}_c'$ or h_c varies much across the subclasses, that is, if the subclasses are heterogeneous either with respect to the outcome variable or the response indicator, or both.

Next, regarding the RE of subclass reweighting, Thomsen (1978) shows that

$$\mathrm{Var}(\bar{y}) = \frac{1}{n\bar{h}^2} \left\{ \sum_{c=1}^{C} W_c h_c S_c^2 + \sum_{c=1}^{C} W_c h_c (\bar{Y}_c' - \bar{Y}')^2 \right\} = \frac{1}{n\bar{h}^2}(\tau_1 + \tau_2),$$

$$\mathrm{Var}(\bar{y}_W) \approx \frac{1}{n}\sum_{c=1}^{C} W_c S_c^2 / h_c,$$

where $S_c^2 = \sum_{i=1}^{N_c'}(Y_{ci} - \bar{Y}_c')^2/(N_c' - 1)$ is the population subclass respondent variance. Notice that $\mathrm{Var}(\bar{y})$ can be decomposed into two terms of within- and between-subclass respondent variances, denoted by τ_1 and τ_2, respectively, with fixed sum $\tau_1 + \tau_2$. A corresponding ANOVA-type measure of the association between c and Y can be given by

$$\lambda_{cY} = \tau_2/(\tau_1 + \tau_2).$$

The association measure λ_{cY} provides an easy appreciation of the potential effects of the auxiliary variable (or variables) underlying the subclasses $c = 1, \ldots, C$. In the extreme case of $\lambda_{cY} = 1$ and $\tau_1 = 0$, we would have $B = \mathrm{Bias}(\bar{y}) = E(\bar{y}) - \bar{Y}$ and $\mathrm{Var}(\bar{y}_W) = 0 < \mathrm{Var}(\bar{y})$. At the other end, where $\lambda_{cY} = 0$, $\tau_2 = 0$ and $S_c^2 \equiv S^2$, we would have $B = 0$ and

$$\mathrm{Var}(\bar{y}) = \frac{S^2}{n} \cdot \frac{1}{\bar{h}} \leq \frac{S^2}{n} \cdot \prod_{c=1}^{C} \left(\frac{1}{h_c}\right)^{W_c} \leq \frac{S^2}{n} \cdot \sum_{c=1}^{C} \frac{W_c}{h_c} \approx \mathrm{Var}(\bar{y}_W),$$

by applying twice the inequality of weighted arithmetic and geometric means, or directly the Titu's lemma as a special case of Cauchy-Schwarz inequality. Between the two extreme cases, increasing λ_{cY} makes the subclasses more heterogeneous with respect to Y, which tends to decrease the within-subclass variances S_c^2 and $\mathrm{Var}(\bar{y}_W)$, as well as increasing the change of point estimate, that is, provided fixed $\mathrm{h}_1, \ldots, h_C$.

Similarly, an ANOVA-type measure of the association between c and R is given as

$$\lambda_{cR} = \sum_{c=1}^{C} W_c(h_c - \bar{h})^2 \bigg/ \left\{ \sum_{c=1}^{C} W_c h_c (1 - h_c) + \sum_{c=1}^{C} W_c (h_c - \bar{h})^2 \right\} = \nu_2/(\nu_1 + \nu_2)$$

where ν_1 and ν_2 are the within- and between-subclass variances of R, respectively, with fixed sum $\nu_1 + \nu_2$. In the extreme case of $\lambda_{cR} = 1$ and $\nu_1 = 0$, h_c would be either 0 or 1, such that the subclasses are nested in the respondent and nonrespondent strata. We would have $B = 0$, despite perfect association between c and R, so that subclass reweighting affects only the variance depending on λ_{cY}. At the other end, where $\lambda_{cR} = 0$, $\nu_2 = 0$ and $h_c \equiv \bar{h}$, we would again have $B = 0$, where subclass reweighting affects only the variance. Between the two extreme cases, both B and $\mathrm{Cov}_W(\bar{Y}_c', h_c)$ are likely to increase with $\nu_2 = \mathrm{Var}_W(h_c)$ and λ_{cR}. To appreciate what might happen to the variance at the same time, rewrite

$$\mathrm{Var}(\bar{y}_W) \approx \frac{1}{n}\left\{ \sum_{c=1}^{C} W_c S_c^2/\bar{h} - \sum_{c=1}^{C} W_c S_c^2 (h_c - \bar{h})/\bar{h}^2 + \sum_{c=1}^{C} W_c S_c^2 (h_c - \bar{h})^2/\bar{h}^3 \right\},$$

based on Taylor expansion of h_c around $\bar{h}$. As ν_2 increases, the term involving $(h_c - \bar{h})^2$ may increase accordingly, while that involving $(h_c - \bar{h})$ remains small since $\sum_c W_c(h_c - \bar{h}) = 0$. In particular, even if λ_{cY} is high and S_c^2's are relatively small, it is possible for the term involving $(h_c - \bar{h})^2$ to increase to such an extent that we would have

$\mathrm{Var}(\bar{y}_W) > \mathrm{Var}(\bar{y})$. Thus, as λ_{cR} increases, subclass reweighting is likely to achieve greater change of the point estimate while increasing the variance at the same time.

Remark Särndal and Lundström (2010) consider three indicators, $H_1 - H_3$, for the usefulness of auxiliary information. They consider H_2 to be ad hoc, which is only included for exploration. According to their conclusion, they prefer H_1 for a given y-variable, and they argue for H_3 as a tentative choice for the "many y-variables situation", but call for more research to develop other indicators (than H_3).

The indicator H_1 is given by $H_1 = |H_0|$ and $H_0 = \Delta_A/S_y$. Combining Equations (2.1), (5.2), (5.7), (5.8) and (5.11) in Särndal and Lundström (2010), we have

$$H_0 = \frac{\Delta_A}{S_y} = -R_{y,m} \times cv_m = -\frac{Cov(y,m)}{S_y S_m} \times \frac{S_m}{\bar{m}_{r;d}} = -\frac{P}{S_y} Cov(y,m)$$

where P is the weighted response rate, that is, an estimate of $\bar{h}$ in our set-up, and $\Delta_A = (\tilde{Y}_{EXP} - \tilde{Y}_{CAL})/\hat{N}$, with the "expansion" estimator $\tilde{Y}_{EXP}$ and the "calibration" estimator $\tilde{Y}_{CAL}$. Thus, Δ_A is similar to the B-term by Equation (7) in this article, defined as the expectation of $\bar{y} - \bar{y}_W$ under SRS, where $\bar{y} = \hat{t}_d/N = \tilde{Y}_{EXP}/\hat{N}$ and $\bar{y}_W = \tilde{Y}_{CAL}/\hat{N}$ by subclass reweighting. Notice that by Equation (7) in this article, B is a function of $\bar{h}$ and $Cov_W(\bar{Y}_c^{'}, h_c)$. The key difference between Δ_A and B is that the latter is based on the response propensity p_i's, whereas the former is based on m_i's which are on the scale of $1/p_i$.

Next, $H_3 = cv_m$, which is based on the auxiliary variables and the response indicator but not the y-values. In this sense, it is similar to λ_{cR} in our article, which measures the association between the auxiliary variables and the response indicator. While H_3 is related to the variance of m_i, λ_{cR} is related to the variance of p_i; while H_3 depends in addition on $\bar{h}$, λ_{cR} depends in addition on the decomposition of the variance of p_i.

Thus, by introducing λ_{cY} and λ_{cR}, we move into areas not covered by Särndal and Lundström (2010). In particular, we find that as λ_{cY} increases, reweighting tends to increase both bias adjustment (B or Δ_A) and efficiency gains; whereas as λ_{cR} increases, reweighting is likely to increase bias adjustment, but inflate the variance at the same time.

4.2. A Simulation Study

In Subsection 4.1, we presented the formula for B, the change in point estimate due to subclass reweighting, as well as the fomulas for the variance of the unweighted respondent mean $\bar{y}$ and weighted mean $\bar{y}_W$. These formulas hold exactly under SRS. In practice, strict SRS is not the most common design, despite the household survey, inclusion probabilities tend not to vary greatly across the population. They can still provide useful indications for the relative importance and potential effects of the different auxiliary variables in reweighting, as we will discuss in more detail in Section 5, even though they do not suffice as the final uncertainty measures to be reported together with the survey estimates. We feel that such uses are warranted based on our past experience of in-house empirical evaluations. Below, we carry out a simple simulation study to illustrate this point.

First, we generate a Norwegian Labour Force population that resembles the LFS in the first quarter of 2015, including the response indicator. This proceeds as follows.

- The population of approximately 3.8 million Norwegians aged 15–74 are distributed in the 19 counties, according to the situation in the first quarter of 2015. The county population size varies from approximately 58,000 to 506,000. For more details on the population, we refer to https://www.ssb.no/en/befolkning/statistikker/folkemengde.
- Within each county, assign each person a binary registered employment status, such that the total number of registered employed people is as given in the first quarter of 2015.
- Within each county, simulate independently the LFS classification (employed, unemployed, inactive) for each person, by the multinomial distribution with the corresponding proportions observed among the LFS respondents in that county.
- Within each county h, simulate independently the response indicator (yes, no) for each person, using the Bernoulli distribution with a probability $0.81 + \mathrm{d}_{1h}$ if the person is registered employed and $0.76 + d_{0h}$ if the person is not registered employed. The Figures 0.81 and 0.76 are respectively the average response rates for the registered employed and not registered employed in the first quarter of 2015. Within each stratum, the response rates for these two groups vary slightly, about 2% above or below the averages. Hence, d_{0h} and d_{1h} are simulated to have a normal distribution with mean 0 and standard deviation 0.01 to reflect the range of the corresponding stratum response rates observed in the LFS sample.

We then repeatedly draw samples (of the same size as the LFS) from this population using SRS or Stratified SRS (StrSRS), where the strata are the 19 counties and the stratum sample sizes are the same as in the Norwegian LFS. The county sample size varies from 610 to 2,745. Based on m simulated samples, with sufficiently large m, we may compare the true values of B, $Var(\bar{y})$ and $Var(\bar{y}_W)$ under each sampling design, with the expected sample estimates of them using the formulas in Subsection 4.1 under the assumption of SRS. The results for the proportions of unemployed and employed persons are given in Table 2. The table shows that the formulas under the SRS assumption ("Estimated") hold as approximately well under the Stratified SRS sampling design.

4.3. Examples from the Norwegian LFS Data

In practice, λ_{cY} and λ_{cR} are neither 0 nor 1, and they vary simultaneously with the auxiliary variables. In the literature, such as those cited in Section 3, it is often suggested that one should select variables that have high associations with *both* Y and R. Little and Vartivarian (2005) summarise in their "Table 1" the effects of reweighting, depending on the association of the auxiliary variables to Y and R, which is reproduced here as Table 3. However, our own experiences (Zhang et al. 2013) suggest that there exist greater nuances

Table 2. Simulation results (× 10^{-3}), $m = 1{,}000$.

	Unemployment				Employment		
	Estimated	SRS	StrSRS		Estimated	SRS	StrSRS
B	− 1.00	− 1.00	− 1.00	B	12.48	12.44	12.62
$s.e.(\bar{y})$	1.14	1.14	1.20	$s.e.(\bar{y})$	3.34	3.33	3.45
$s.e.(\bar{y}_W)$	1.16	1.16	1.22	$s.e.(\bar{y}_W)$	1.90	2.01	1.93

Table 3. Effects of nonresponse reweighting, from Little and Vartivarian (2005).

	Association with outcome variable	
Association with nonresponse	Low	High
Low	Effect on bias: — Effect on variance: —	Effect on bias: — Effect on variance: ↓
High	Effect on bias: — Effect on variance: ↑	Effect on bias: ↓ Effect on variance: ↓

in reality, which we demonstrate below using four examples based on the Norwegian LFS data. The examples also illustrate how (λ_{cY}, λ_{cR}) may be related to the changes of the point estimate and the associated variance.

We use the Norwegian LFS in the first quarter of 2015. The sample size is $n = 24{,}353$ and the response rate is $\bar{h} = 0.79$. We consider two binary Y-variables: employment and unemployment status. All the terms B, Var($\bar{y}$), Var($\bar{y}_W$), and so on are estimated based on the observed sample. However, for simplicity we do not introduce extra notations to emphasise that the values presented are estimates instead of population quantities.

Example 1 Let Y be the LFS Unemployment Status. Let two subclasses be formed based on the Registered Employment Status, where $c = 1$ for not registered employed and $c = 2$ for registered employed. We have $W_c = (0.35, 0.65)$ and $h_c = (0.74, 0.81)$, for $c = (1, 2)$, with the corresponding subclass respondent means $\bar{y}_c = (0.07, 0.00)$ and respondent variances $S_c^2 = (0.06, 0.00)$. We obtain

$$\lambda_{cY} = 0.04,\ \lambda_{cR} = 0.01,\ B = -1.41 \times 10^{-3},\ \text{s.e.}(\bar{y}) = 1.13 \times 10^{-3},\ \text{RE} = 1.07.$$

Both λ_{cY} and λ_{cR} are close to zero. This provides an example of the top-left scenario in Table 3, according to which reweighting has little effect. However, the point estimate is actually changed by about 120% of the standard error (s.e.) of $\bar{y}$, while it increases the variance only slightly. Previous studies of the Norwegian data (Zhang 1999; Thomsen and Zhang 2001; Zhang 2005) all conclude that employment is overestimated and unemployment underestimated, based on the unadjusted respondent sample. Therefore, the adjustment B is in the direction one would expect, and it is by no means 'negligible' in size, despite the low association of the auxiliary variable with both Y and R.

Example 2 Let Y be the LFS Employment Status, and keep the same subclasses as in Example 1. We have $\bar{y}_c = (0.14, 0.96)$ and $S_c^2 = (0.12, 0.04)$, and

$$\lambda_{cY} = 0.69,\ \lambda_{cR} = 0.01,\ B = 1.68 \times 10^{-2},\ s.e.(\bar{y}) = 3.31 \times 10^{-3},\ \text{RE} = 0.34.$$

It can be seen that λ_{cR} stays the same but λ_{cY} is greatly increased, compared to when the outcome variable is Unemployment Status. This provides an example of the top-right scenario in Table 3, according to which reweighting leads to little bias adjustment, although it may reduce the variance. However, it can be seen that in addition to the huge variance reduction, the change in the point estimate is also several times the standard error.

Example 3 Let Y be the LFS Employment Status. Let the subclasses be formed using the Panel Response Status, where $c = 1$ if previous nonrespondent, $c = 2$ if previous respondent, and $c = 3$ if new sample unit. For $c = (1, 2, 3)$, we obtain $W_c = (0.20, 0.67, 0.13)$, $h_c = (0.29, 0.94, 0.77)$, $\bar{y}_c = (0.66, 0.71, 0.66)$, and $S_c^2 = (0.22, 0.21, 0.22)$, so that

$$\lambda_{cY} = 0.00,\ \lambda_{cR} = 0.39,\ B = 5.50 \times 10^{-3},\ s.e.(\bar{y}) = 3.31 \times 10^{-3},\ \text{RE} = 1.28.$$

Compared to Example 1, λ_{cR} is considerably increased, but λ_{cY} remains almost zero. This provides an example of the low-left scenario in Table 3, according to which reweighting leads to little bias adjustment, although it may increase the variance. Actually, however, in addition to the increasing variance, the change in the point estimate is again by no means 'negligible' in size, despite the low association between the auxiliary variable and Y.

Example 4 Let Y be the LFS Employment Status. Crossing the Panel Response Status and the Registered Employment Status yields the subclasses, where $c = 1$ if previous nonrespondent and not registered employed, $c = 2$ if previous nonrespondent and registered employed, $c = 3$ if previous respondent and not registered employed, $c = 4$ if previous respondent and registered employed, $c = 5$ if new sample unit and not registered employed, and $c = 6$ if new sample unit and registered employed. Then, for $c = (1, 2, 3, 4, 5, 6)$, we obtain $W_c = (0.08, 0.12, 0.21, 0.47, 0.05, 0.08)$, $h_c = (0.25, 0.31, 0.93, 0.94, 0.72, 0.79)$, $\bar{y}_c = (0.14, 0.95, 0.14, 0.96, 0.10, 0.95)$, $S_c^2 = (0.12, 0.05, 0.12, 0.04, 0.09, 0.05)$, and

$$\lambda_{cY} = 0.69,\ \lambda_{cR} = 0.39,\ B = 1.78 \times 10^{-2},\ s.e.(\bar{y}) = 3.31 \times 10^{-3},\ \text{RE} = 0.43.$$

Compared to Example 2, λ_{cR} is considerably increased in addition to high λ_{cY}. This provides an example of the low-right scenario in Table 3, which is 'ideal' according to the prevailing recommendation in the literature. However, while the adjustment B is increased by about 6% compared to the reweighting in Example 2, there is also a loss of efficiency by about 26%. In other words, it is not unreservedly beneficial to increase the association with R, while the association with Y remains the same. In fact, we now demonstrate the caveat of doing so with the following thought experiment.

Example 4* The first two h_c's in Example 4 are the response rates of the previous nonrespondents, the next two of the previous respondents, and the last two of the new sample members. To vary the response rates more extremely, suppose we have full response among the previous respondents, so that $h_3 = h_4 = 1$; suppose the response rates among the new sample units stay the same, so that $h_5 = 0.72$ and $h_6 = 0.79$; suppose the response rates among the previous nonrespondents are reduced to $h_1 = 0.05$ and $h_2 = 0.10$. This yields $h_c = (0.05, 0.10, 1.00, 1.00, 0.72, 0.80)$, with the same overall response rate $\bar{h} = 0.79$. Keeping everything else the same as in Example 4, we obtain

$$\lambda_{cY} = 0.69,\ \lambda_{cR} = 0.78,\ B = 1.99 \times 10^{-2}, s.e.(\bar{y}) = 3.31 \times 10^{-3},\ \text{RE} = 1.13.$$

As we remarked earlier in Subsection 4.1, *without* increasing λ_{cY} at the same time, increasing λ_{cR} on its own can result in $\text{Var}(\bar{y}_W) > \text{Var}(\bar{y})$, despite high association λ_{cY}.

5. Empirical Study of Reweighting

For this study of reweighting, a number of auxiliary variables are extracted from the statistical register system at Statistics Norway and linked to the samples at the individual level. For the LFS, these include age (11), sex (2), county (19), education level (4), marital status (3), family type (3), immigration (3), birth country (2), income (5), household income (5), and registered employment (2), where the numbers in parentheses indicate the number of categories each variable has. The same variables are used for the SILC, except for registered employment due to data protection regulations. In addition, some of the variables are adjusted to have fewer categories due to the smaller SILC sample size, for example 4 age groups instead of 11, 7 regions instead 19 counties, and so on.

For both *R*- and *Y*-modelling, variable selection is carried out stepwise according to the Akaike Information Criterion. While this is somewhat simplistic, it suffices for the purpose of this study and reflects well the existing process at national statistical offices. All six estimators listed in Table 1 are applied to each of the outcome variables to be presented, in terms of the corresponding population mean estimators, denoted by $\bar{y}_{method} = \hat{t}_{method}/N$ where the subscript *method* identifies the weight adjustment method. The baseline estimator to be presented is $\bar{y} = \hat{t}/N$ for $\hat{t}$ given by (2). The difference to $\hat{t}_d/N$ is negligible compared to their differences to the various reweighting estimates. To save space, other estimators that have been calculated may be mentioned in comments but not presented in detail. This include, for example, using RPS instead IPW under the two-step approach. All the estimated variances are calculated in R using 500 bootstrap samples with the same sampling design as the LFS/SILC, except for one case to be specified later. The bootstrap follows the procedure of Canty and Davison (1999), where to mimic the effect of sampling without replacement, the bootstrap population is made by concatenating copies of the observed sample, from which the bootstrap replicate samples are taken without replacement according to the given sampling design. For each sample, we calculate the estimates for each of the estimators discussed in Section 3, and the standard deviation of these estimates is used to estimate the standard error of each estimator.

5.1. The LFS

We have carried out the same analysis for five quarterly samples. The results are very similar, so only those based on the first quarter of 2015 are presented here, where we focus on two binary outcome *Y*-variables, employment and unemployment, denoted by Y_{em} and Y_{un}, respectively.

The association measures of each covariate with R, Y_{em} and Y_{un} are given in Table 4, together with B and RE by the respective subclass reweighting, as described in Subsection 4.1. It can be seen that the available covariates have very different associations with the two outcome variables. While registered employment, age, income and education all have a high association with Y_{em}, the association with Y_{un} is much lower across the board, although registered employment and income remain the two with the highest associations there. The covariates selected for the *R*-model and the two *Y*-models are marked (by †) for the corresponding λ_{cR}, $\lambda_{cY_{em}}$ and $\lambda_{cY_{un}}$ (Table 4). No interaction terms are selected for any of the models based on these data. Largely the same variables are selected for both *Y*-models, denoted by X_{em} and X_{un}, respectively. Each model includes the covariates that have the

Table 4. Association with (R, Y_{em}, Y_{un}), selected†, B in 10^{-2}.

Auxiliary variable	λ_{cR}	$\lambda_{cY_{em}}$	$\lambda_{cY_{un}}$	B_{em} (RE)	B_{un} (RE)
Registered employment	$0.01^{\dagger}$	$0.69^{\dagger}$	$0.04^{\dagger}$	1.68 (0.33)	−0.14 (1.07)
Age	$0.02^{\dagger}$	$0.28^{\dagger}$	$0.01^{\dagger}$	−1.04 (0.71)	−0.10 (1.08)
Sex	$0.00^{\dagger}$	$0.00^{\dagger}$	$0.00^{\dagger}$	0.01 (1.00)	0.00 (1.00)
County	$0.00^{\dagger}$	$0.01^{\dagger}$	0.00	0.07 (0.99)	0.00 (1.00)
Family type	$0.02^{\dagger}$	0.02	0.00	0.19 (0.99)	−0.02 (1.02)
Birth country	0.02	0.00	0.01	−0.25 (1.01)	−0.15 (1.11)
Immigration status	$0.02^{\dagger}$	0.00	$0.01^{\dagger}$	−0.20 (1.02)	−0.16 (1.12)
Education	$0.01^{\dagger}$	$0.11^{\dagger}$	$0.01^{\dagger}$	0.59 (0.91)	−0.06 (1.04)
Marital status	$0.02^{\dagger}$	0.01	0.00	0.30 (1.00)	−0.06 (1.05)
Income	$0.02^{\dagger}$	$0.26^{\dagger}$	$0.02^{\dagger}$	1.47 (0.80)	−0.13 (1.08)
Household income	$0.04^{\dagger}$	$0.09^{\dagger}$	0.01	1.39 (0.96)	−0.14 (1.13)

highest association with either Y_{em} or Y_{un}. The R-model includes all the available covariates (Z), except for birth country that is similar to immigration status. In particular, both X_{em} and X_{un} are nested in Z, such that $V = Z$ for both Y_{em} and Y_{un}.

The different estimates and their associated s.e.'s (in parentheses) are given in Table 5. Compared to the baseline estimate, all the reweighting estimates adjust the employment rate downwards and the unemployment rate upwards, that is, in the expected direction. In the case of employment, all the one-step MGR and two-step estimators reduce the variance, while the one-step IPW estimator increases the variance. In the case of unemployment, all the reweighting estimators increase the variance, but have similar RE to each other. For both Y_{em} and Y_{un}, the point-estimate changes are very large compared to the s.e.'s. Bias exploration by the method described in Zhang (1999) suggests that, provided informative nonresponse, the reweighted employment estimators may still have a positive bias, so that the risk is low that the reweighted estimators are more biased than the baseline estimator. Likewise for the reweighted unemployment estimators, since the upward adjustments of unemployment resulted from reweighting appear plausible in magnitude compared to the downward adjustments of employment.

To a large extent, these results have confirmed the potential adjustment effects, which are suggested by simple subclass reweighting and association measures in Subsection 4.3.

Table 5. LFS estimates (s.e.) in 10^{-2}, the first quarter 2015.

Auxiliary for (IPW, MGR)	Mean employment, $\bar{y} = 69.84$ (0.35)		
	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step estimator
(Z, X_{em}) (Z, Z)	67.47 (0.44) ”	67.10 (0.19) 67.10 (0.19)	$\bar{y}_{2sts} = 67.08$ (0.19) $\bar{y}_{DR} = 67.09$ (0.19)
Auxiliary for (IPW, MGR)	Mean unemployment, $\bar{y} = 2.45$ (0.12)		
	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step estimator
(Z, X_{un}) (Z, Z)	2.99 (0.14) ”	3.06 (0.14) 3.05 (0.14)	$\bar{y}_{2sts} = 3.18$ (0.15) $\bar{y}_{DR} = 3.19$ (0.15)

As indicated in Example 2, it is possible to achieve large adjustment of the point estimate *and* variance reduction for Y_{em} without high association with R but provided high association with the outcome variable. Moreover, as indicated in Example 1, the reweighting estimators can yield appreciable adjustment of the point estimate of Y_{un} but also slightly increase the variance, despite the low association with both Y_{un} and R.

Cross-examination of the estimators gives rise to additional noteworthy observations. Firstly, a striking result in Table 5 is the large variances of the IPW estimators, for example, $\bar{y}_{IPW}$ is even less efficient than the baseline estimator $\bar{y}$ for Y_{em}. Using RPS with five groups does not result in a smaller variance compared to that of the IPW estimator either. Recall that in the case of $V = Z \vee X = X$, Lunceford and Davidian (2004) show that "over-modelling" $p(z_i;\ \alpha)$ by $p(v_i;\ \eta)$ can reduce the variance of the IPW estimator. However, since X_{em} is a subset of Z here, the predictive covariates are already included in Z and the strategy of "over-modelling" does not work. This shows that having predictive variables for Y in the R-model does not guarantee efficiency by itself, without an appropriate weight adjustment method. For instance, the MGR estimator based on "over-modelling" $\mu(v;\ \xi)$ with $v = z$ is basically as efficient as $\bar{y}_{MGR}$ that only uses X_{em}. Moreover, the two-step estimator $\bar{y}_{2sts}$ is able to recover almost all the lost efficiency of $\bar{y}_{IPW}$ by calibration of the IPW-adjusted weights $d_i/p(z_i;\hat{\alpha})$ with respect to X_{em}.

Secondly, the two-step approach $\bar{y}_{2sts}$ does not offer any noticeable advantage over the one-step MGR for the Norwegian LFS. In theory, correct modelling of the unit nonresponse could yield approximately unbiased estimation for any outcome variable. In reality, however, the true nonresponse model is unobtainable. This is certainly the case with the LFS data here, given the low association between the available covariates and R. Empirically, $\bar{y}_{2sts}$ does not yield any notable improvement over $\bar{y}_{MGR}$ here, but is more complicated due to an extra step of model-fitting and reweighting.

Thirdly, the DR approach does not offer any noticeable advantage compared to the traditional one- and two-step approaches for the Norwegian LFS. In the case of Y_{em}, where there is a good Y-model, the results here agree with the literature (Bang and Robins 2005; Kang and Schafer 2007) that the DR estimator does not perform better than the regression estimator, but could improve the performance of $\bar{y}_{IPW}$ obtained from the R-model alone. Compared to the two-step estimator $\bar{y}_{2sts}(Z,X)$, the DR estimator $\bar{y}_{2sts}(Z,Z)$ has the same IPW-weights, but differ with respect to the extra calibration variables in $Z \backslash X_{em}$ for Y_{em} and $Z \backslash X_{un}$ for Y_{un}. However, this makes little difference since the extra variables do not have any appreciable association with the respective outcome variable.

Therefore the one-step MGR estimator $\bar{y}_{MGR}$ seems reasonable for the Norwegian LFS, among the options considered here. The auxiliary variables may be selected with respect to several key Y-variables. It is the simplest in production, and has the lowest variance, although, in this case, the differences compared with the two-step alternatives are minor. We note that the existing production method in the LFS is essentially the same as subclass reweighting based on post-stratification by sex, age, and registered employment. It performs similarly to $\bar{y}_{MGR}$ for both Y_{em} and Y_{un}, with somewhat smaller adjustment of the point estimates, but also smaller variance for Y_{un}. Therefore, the key to improve the existing method must be to find other auxiliary variables in the statistical register system, as more administrative data are being made available, that are more predictive of the unemployment status Y_{un}. The MGR can be used instead of the post-stratification if the number of auxiliary variables increases.

Table 6. Association with (R, $Y_{en,}$ Y_{he}), selected†, B in 10^{-2}.

Auxiliary variable	λ_{cR}	$\lambda_{cY_{en}}$	$\lambda_{cY_{he}}$	B_{en} (RE)	B_{he} (RE)
Age	0.00†	0.02†	0.01†	0.01 (0.98)	0.12 (0.96)
Sex	0.00†	0.00	0.00	0.08 (0.98)	0.03 (0.99)
Region	0.00†	0.00	0.00	0.15 (0.98)	0.05 (0.99)
Family type	0.00	0.02†	0.00	−0.13 (1.00)	0.00 (1.00)
Birth country	0.01†	0.02†	0.00	−0.43 (1.05)	−0.04 (1.02)
Education	0.04†	0.01	0.01†	−0.39 (1.08)	−0.37 (1.13)
Marital status	0.01†	0.03†	0.01†	−0.31 (1.02)	0.07 (0.97)
Income	0.03†	0.03†	0.02†	−0.67 (1.08)	−0.42 (1.12)
Household income	0.02†	0.06†	0.02†	−0.93 (1.08)	−0.34 (1.10)

5.2. The SILC

For the SILC, we use data from the 2015 sample, where the response rate is 57% and the net sample size is about 9,200. We focus on two binary Y-variables: whether people find it difficult to make ends meet and whether they have poor health conditions, denoted by Y_{en} and Y_{he}, respectively.

The association measures of each available covariate with R, Y_{en} and Y_{he} are given in Table 6, together with B and RE by the respective subclass reweighting. Here, we are in a situation of only low association with both the outcome variables and nonresponse across the board. The covariates selected for the R-model and the two Y-models are marked (by †) in Table 6. No interaction terms are selected for any of the models based on these data. As in the case of LFS, largely the same variables are selected for both Y-models, denoted by X_{en} and X_{he}, respectively, and each of them includes the covariates that have the highest association with either Y_{en} or Y_{he}. The R-model includes all the available covariates (Z), except for family type that resembles marital status. While X_{he} is entirely nested in Z, X_{en} is almost so, except for family type.

The different estimators and their associated s.e.'s (in parentheses) are given in Table 7. Compared to the baseline estimates, reweighting leads to upwards adjustments for both Y_{en} and Y_{he}, and increases the variance in all the cases. Again, as exemplified in Subsection 4.3, the adjustment of the point estimate can be large, several times the s.e.'s here, despite

Table 7. SILC estimates (s.e.) in 10^{-2}, year 2015, $V = Z \vee X_{en}$.

Auxiliary for (IPW, MGR)	Mean of Y_{en}, $\bar{y} = 11.20$ (0.39)		
	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step Estimator
(Z, X_{en}) (V, V)	13.05 (0.44) 13.05 (0.44)	14.16 (0.46) 14.22 (0.46)	$\bar{y}_{2sts} = 14.68$ (0.48) $\bar{y}_{DR} = 14.65$ (0.48)
Auxiliary for (IPW, MGR)	Mean of Y_{he}, $\bar{y} = 6.07$ (0.30)		
	One-step $\bar{y}_{IPW}$	One-step $\bar{y}_{MGR}$	Two-step Estimator
(Z, X_{he}) (Z, Z)	6.83 (0.35) "	6.99 (0.35) 6.98 (0.37)	$\bar{y}_{2sts} = 7.00$ (0.36) $\bar{y}_{DR} = 6.94$ (0.38)

the low association with both Y and R; whereas low association with Y does increase the variance. For both Y-variables, it can be seen that the one-step MGR and the two-step estimators are closer to each other than the one-step IPW estimators. In particular, the IPW estimators do not have larger variances, compared to any of the alternatives that include calibration towards the selected population auxiliary totals. Notice that using RPS with five groups reduces the variance of $\bar{y}_{IPW}$ slightly, and it may somewhat change the point estimate, for example we would have $\bar{y}_{en} = 12.75$ (0.43) and $\bar{y}_{he} = 6.85$ (0.34) instead.

Regarding the three reweighting approaches, the results suggest similar conclusions for the SILC and for the LFS. The DR estimator using (V, V), for $V = Z \vee X$, does not offer any noticeable advantage compared to the traditional two-step approach using (Z, X) for the SILC. Nor does the two-step approach $\bar{y}_{2sts}$ using (Z, X) offer any trustworthy advantage over the one-step MGR using X. The variance of $\bar{y}_{2sts}$ is slightly larger than that of $\bar{y}_{MGR}$ for both Y-variables. The adjustment of the point estimate is similar in the case of Y_{he}, and about one s.e. larger by $\bar{y}_{2sts}$ for Y_{en}. However, given the low association of the available covariates with nonresponse, the R-model is hardly the true nonresponse model. Indeed, given the low association with the Y-variables, it seems possible that the difference in the adjusted point estimates may be spurious.

The situation here, where one may only achieve low association with Y, may very well happen in many countries that have fewer auxiliary variables available than Norway. It is often possible to find additional sample covariates that have higher association with nonresponse. For instance, given the rotating panel design of the SILC, one may introduce the Panel Response Status (PRS) as in Examples 3 and 4 in Subsection 4.3, which has a higher association with R ($\lambda_{cR} = 0.20$), but almost no association with the two Y-variables ($\lambda_{cY_{en}} = 0.00$, and $\lambda_{cY_{he}} = 0.00$). The variable PRS has three categories indicating whether an individual is a previous respondent, previous nonrespondent, or is a new sample unit. Adding PRS as an extra covariate to Z given in Table 6 yields Z^* for the R-model.

The new one-step IPW and two-step estimators using Z^* for the R-model are given in Table 8. The 500 bootstrap resamples are generated with the same design as the SILC, but are further stratified by whether or not an individual is a new sample unit. The most notable feature in Table 8 is that all the reweighting estimators produce greater point-estimate adjustments, but also considerably larger variances, compared to the corresponding estimators without PRS in Table 7. A simple explanation is that PRS enhances the association with R without increasing the association with the two Y-variables. On the one hand, it is highly likely that the baseline $\bar{y}$ underestimates both proportions, since all the reweighting methods produce upwards adjustments. On the other hand, it is unclear whether the bias of any adjusted estimator may have gone from negative to positive, and the increased variances certainly suggest a heightened risk of introducing spurious adjustments.

Table 8. SILC estimates (s.e.) in 10^{-2}, with Z^* for R-model.

	One-step $\bar{y}_{IPW}$	Two-step $\bar{y}_{2sts}$	Two-step $\bar{y}_{DR}$
Mean of Y_{en}: $\bar{y} = 11.20$ (0.39) Mean of Y_{he}: $\bar{y} = 6.07$ (0.29)	14.39 (0.65) 7.09 (0.43)	15.57 (0.66) 7.14 (0.43)	15.43 (0.63) 7.04 (0.44)

The existing production method of the SILC is reweighting by about 200 subclasses, which are formed by cross-classifying several of the auxiliary variables considered here. Stablising the variance of estimation is therefore an important aspect for improvement. This speaks against including variables like PRS, because the affected estimators would have considerably larger variances. Recall that X_{en} and X_{he} are essentially nested in Z (Table 6). A possible resolution is to settle for a common set of variables, denoted by Q, and choose between the IPW and MGR estimators based on an overall assessment of their efficiency for different Y-variables. Two initial choices for Q are (i) the intersection $Q_0 = Z \wedge X_{en} \wedge X_{he}$, and (ii) the union $Q_1 = Z \vee X_{en} \vee X_{he}$. In addition, one can explore any of the 32 possible Q between Q_0 and Q_1, and obtain the corresponding IPW and MGR estimates that are given in Table 9.

We observe the same pattern in Table 9 as previously, given low association with Y: the auxiliary variables Q that yield greater adjustment of the point estimates also lead to larger

Table 9. SILC estimates (s.e.) in 10^{-2}, with different auxiliary variables.

	Mean of Y_{en}		Mean of Y_{he}	
Variables	IPW (s.e.)	MGR (s.e.)	IPW (s.e.)	MGR (s.e.)
Q_0	12.68 (0.43)	13.27 (0.43)	6.61 (0.33)	6.81 (0.34)
Q_0, region	12.65 (0.43)	13.24 (0.43)	6.62 (0.33)	6.81 (0.34)
Q_0, sex	12.68 (0.43)	13.27 (0.43)	6.61 (0.33)	6.80 (0.34)
Q_0, birth country	12.79 (0.43)	14.06 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, education	12.91 (0.44)	13.37 (0.43)	6.83 (0.34)	6.99 (0.35)
Q_0, family type	12.70 (0.43)	13.37 (0.43)	6.61 (0.33)	6.81 (0.34)
Q_0, region, sex	12.65 (0.43)	13.24 (0.43)	6.62 (0.33)	6.81 (0.34)
Q_0, region, birth country	12.76 (0.43)	14.02 (0.46)	6.59 (0.33)	6.78 (0.35)
Q_0, region, education	12.89 (0.44)	13.36 (0.43)	6.84 (0.34)	7.00 (0.35)
Q_0, region, family type	12.67 (0.43)	13.35 (0.43)	6.63 (0.33)	6.81 (0.34)
Q_0, sex, birth country	12.79 (0.43)	14.06 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, sex, education	12.90 (0.44)	13.37 (0.43)	6.84 (0.34)	6.99 (0.35)
Q_0, sex, family type	12.70 (0.43)	13.37 (0.43)	6.61 (0.33)	6.80 (0.34)
Q_0, birth country, education	13.07 (0.44)	14.18 (0.46)	6.81 (0.34)	7.00 (0.37)
Q_0, birth country, family type	12.82 (0.43)	14.16 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, education, family type	12.91 (0.44)	13.45 (0.43)	6.82 (0.34)	6.99 (0.35)
Q_0, region, sex, birth country	12.76 (0.43)	14.02 (0.46)	6.59 (0.33)	6.78 (0.35)
Q_0, region, sex, education	12.88 (0.44)	13.36 (0.43)	6.85 (0.35)	7.00 (0.35)
Q_0, region, sex, family type	12.67 (0.43)	13.35 (0.43)	6.63 (0.33)	6.81 (0.34)
Q_0, region, birth country, education	13.06 (0.44)	14.14 (0.46)	6.81 (0.34)	6.97 (0.36)
Q_0, region, birth country, family type	12.78 (0.43)	14.12 (0.47)	6.59 (0.33)	6.78 (0.35)
Q_0, region, education, family type	12.89 (0.44)	13.44 (0.43)	6.84 (0.34)	7.00 (0.35)
Q_0, sex, birth country, education	13.07 (0.44)	14.18 (0.46)	6.82 (0.35)	7.00 (0.37)
Q_0, sex, birth country, family type	12.82 (0.43)	14.16 (0.46)	6.59 (0.33)	6.80 (0.35)
Q_0, sex, education, family type	12.90 (0.44)	13.45 (0.43)	6.84 (0.34)	6.99 (0.35)
Q_0, birth country, education, family type	13.07 (0.44)	14.26 (0.46)	6.81 (0.34)	6.99 (0.37)
Q_0, region, sex, birth country, education	13.05 (0.44)	14.15 (0.46)	6.83 (0.35)	6.98 (0.37)
Q_0, region, sex, birth country, family type	12.78 (0.43)	14.12 (0.47)	6.59 (0.33)	6.78 (0.35)
Q_0, region, sex, education, family type	12.88 (0.44)	13.44 (0.43)	6.85 (0.35)	7.00 (0.35)
Q_0, region, birth country, education, family type	13.06 (0.44)	14.22 (0.46)	6.81 (0.34)	6.97 (0.36)
Q_0, sex, birth country, education, family type	13.07 (0.44)	14.26 (0.46)	6.82 (0.35)	7.00 (0.37)
Q_1	13.05 (0.44)	14.22 (0.46)	6.82 (0.35)	6.98 (0.37)

variances. The simplest choice here appears to be Q_0, which achieves the minimum s.e.'s for both the IPW and MGR estimators for both the Y-variables. Adding extra auxiliary variables does not improve the efficiency, but it may be accepted in practice, if benchmarking towards the extra variable is considered necessary and the induced adjustment and variance are deemed reasonable. For example, region may be added to Q_0 to produce consistent regional estimates without losing efficiency or significantly affecting the point estimates.

6. Conclusions

Two interdependent decisions are required when reweighting for unit nonresponse: auxiliary variable selection and weight adjustment method. The following conclusions emerge from the review and empirical appraisal above.

When selecting the auxiliary variables, it is always useful to increase the association with the outcome variable, but seeking higher association with nonresponse is not necessarily helpful. In particular, one can achieve large useful adjustment of the point estimate and reduce the variance at the same time, provided high association with the outcome variable but only low association with nonresponse. While it is often possible to find variables that are primarily associated with nonresponse but not the outcome variables, such as the variable PRS in the LFS and SILC, caution would be necessary regarding such variables, because they tend to inflate the variance and heighten the risk of spurious adjustment, as has been demonstrated empirically in Subsections 4.3 and 5.2.

Regarding weight adjustment, the choice of method does matter, for example between the one-step IPW and MGR estimators, especially when there exist strong auxiliary variables for the outcome available, as for the employment variable in the LFS. In particular, it would be unwise *only* to consider the IPW (or RPS) estimator based on a nonresponse model, when high association with the outcome variable is available. Provided weak auxiliary variables for the outcome variable, bigger adjustment of the point estimate is often accompanied by an increasing variance, by either the IPW or MGR estimator. Limiting the loss of efficiency and avoiding spurious adjustment may be the priority in such situations. Thus, it is important to pay attention not only to the size of adjustment of the point estimate by the weight adjustment method, but also the effects of reweighting on the variance of estimation, whether the given auxiliary variables are strong or weak.

Finally, regarding the three main reweighting approaches identified in Section 1, we found no evidence in the situations examined that supports an uncritical general adoption of either the two-step approach. Neither the traditional nor the DR two-step approach yields any gains empirically for the Norwegian LFS and SILC. Since the 'true' nonresponse model envisaged for a two-step approach cannot be identified based on the observed data, whether the available auxiliary variables have low or high association with nonresponse, it makes sense to choose based on cross-examination of the alternatives in a given situation.

7. References

Andersson, P.G. and C.-E. Särndal. 2016. "Calibration for nonresponse treatment: in one or two steps?" *Statistical Journal of the IAOS* 32: 1–7.

Bang, H. and J.M. Robins. 2005. “Doubly robust estimation in missing data and causal inference models.” *Biometrics* 61: 962–972. DOI: https://doi.org/10.1111/j.1541-0420.2005.00377.x.

Bethlehem, J.G. 1988. “Reduction of nonresponse bias through regression estimation.” *Journal of Official Statistics* 4: 251–260. Available at: https://search.proquest.com/openview/90feaa08b1f293c8544d8576e6d94436/1?pq-origsite=gscholar&cbl=105444 (accessed February 2020).

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken: Wiley.

Brick, M.J. 2013. “Unit nonresponse and weighting adjustments: A critical review.” *Journal of Official Statistics* 29: 329–353. DOI: https://doi.org/10.2478/jos-2013-0026.

Canty, A. and A. Davison. 1999. “Resampling-based Variance Estimation for Labour Force Surveys.” *The Statistician* 48: 379–391.

Carpenter, J.R., M.G. Kenward, and S. Vansteelandt. 2006. “A comparison of multiple imputation and doubly robust estimation for analyses with missing data.” *J. R. Statist. Soc. A* 169: 571–584. DOI: https://doi.org/10.1111/j.1467-985X.2006.00407.x.

Cassel, C.M., C.-E. Särndal, and J.H. Wretman. 1983. “Some uses of statistical models in connection with the nonresponse problem.” In *Incomplete Data in Sample Surveys*, edited by W.G. Madow, I. Olkin, and D.B. Rubin, vol. 3, 143–160. New York: Academic Press.

Cochran, W. 1953. *Sampling Techniques*. New York: Wiley.

De Leeuw, E. and W. de Heer. 2002. “Trends in household survey nonresponse – a longitudinal and international comparison.” In *Survey Nonresponse*, edited by R. Groves, D. Dillman, J. Eltinge, and R.J.A. Little, 41–54. New York: Wiley. Available at: https://www.wiley.com/en-us/Survey+Nonresponse-p-9780471396277.

Eltinge, J.L. and I.S. Yansaneh. 1997. “Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey.” *Survey Methodology* 23: 33–40. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X19970013103 (accessed February 2020).

Haziza, D. and É. Lesage. 2016. “A discussion of weighting procedures for unit nonresponse.” *Journal of Official Statistics* 32: 129–145. DOI: https://doi.org/10.1515/jos-2016-0006.

Kalton, G. and I. Flores-Cervantes. 2003. “Weighting methods.” *Journal of Official Statistics* 19: 81–97. Available at: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/weighting-methods.pdf (accessed February 2020).

Kalton, G. and D. Kasprzyk. 1986. “The treatment of missing survey data.” *Survey Methodology* 12: 1–16. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198600114404 (accessed February 2020).

Kang, J.D.Y. and J.L. Schafer. 2007. “Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.” *Statistical Science* 22: 523–539. DOI: https://doi.org/10.1214/07-STS227.

Kim, J. and D. Haziza. 2014. “Doubly robust inference with missing data in survey sampling.” *Statistica Sinica* 24: 375–394. DOI: http://doi.org/10.5705/ss.2012.005.

Kott, P.S. and D. Liao. 2015. “One step or two? Calibration weighting from a complete list frame with nonresponse.” *Survey Methodology* 41: 165–181. Available at:

https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14172-eng.htm (accessed February 2020).

Little, R.J.A. 1986. "Survey nonresponse adjustments for estimates of means." *International Statistical Review* 54: 139–157. DOI: https://doi.org/10.2307/1403140.

Little, R.J. and D.B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R.J. and S. Vartivarian. 2005. "Does weighting for nonresponse increase the variance of survey means?" *Survey Methodology* 31: 161–168. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20050029046 (accessed February 2020).

Lunceford, J.K. and M. Davidian. 2004. "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study." *Statist. Med.* 23: 2937–2960. DOI: https://doi.org/10.1002/sim.1903.

Lundström, S. and C.-E. Särndal. 1999. "Calibration as a Standard Method for Treatment of Nonresponse." *Journal of Official Statistics* 15: 305–327. DOI: https://doi.org/10.1002/sim.1903.

Meyer, B.D., W.K.C. Mok, and J.X. Sullivan. 2015. "Household surveys in crisis." *Journal of Economic Perspectives* 29: 199–226. Available at: https://pubs.aeaweb.org/doi/pdf/10.1257/jep.29.4.199 (accessed September 2019).

Oh, H.L. and F.S. Scheuren. 1983. "Weighting adjustments for unit nonresponse." In *Incomplete Data in Sample Surveys*, edited by W.G. Madow, I. Olkin, and D.B. Rubin, vol. 2, 143–184. New York: Academic Press.

Robins, B.J.M. and N. Wang. 2000. "Inference for imputation estimators." *Biometrika* 87: 113–124. Available at: https://pdfs.semanticscholar.org/3214/538c562c010bdb-b055582e04fc3a4d7b6ce9.pdf (accessed February 2020).

Robins, J.M., A. Rotnitzky, and L.P. Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *J. Am. Statist. Ass.* 89: 846–866. DOI: https://doi.org/10.1080/01621459.1994.10476818.

Särndal, C.-E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Särndal, C.-E. and S. Lundström. 2008. "Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator." *Journal of Official Statistics* 24: 167–191. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/assessing-auxiliary-vectors-for-control-of-nonresponse-bias-in-the-calibration-estimator.pdf (accessed February 2020).

Särndal, C.-E. and S. Lundström. 2010. "Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias." *Survey Methodology* 36: 131–144. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010002/article/11376-eng.pdf?st=XFmzYjVz (accessed February 2020).

Schouten, B. 2007. "A selection strategy for weighting variables under a not-missing-at-random assumption." *Journal of Official Statistics* 23: 51–68. Available at: https://pdfs.semanticscholar.org/634f/ca032e7c166eef11ee65b84192e5aca1e039.pdf (accessed February 2020).

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. Chichester, UK: Wiley.

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. Available at: https://www.statistics.no/a/histstat/ano/ano_io73_02.pdf (accessed February 2020).

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. Available at: https://www.statistics.no/a/histstat/ano/ano_io73_02.pdf (accessed February 2020).

Thomsen, I. and L.-C. Zhang. 2001. “The effects of using administrative registers in economic short term statistics: The Norwegian Labour Force Survey as a case study.” *Journal of Official Statistics* 17: 285–294. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/the-effects-of-using-administrative-registers-in-economic-short-term-statistics-the-norwegian-labour-force-survey-as-a-case-study.pdf (accessed February 2020).

Zhang, L.-C. 1999. “A note on post-stratification when analyzing binary survey data subject to nonresponse.” *Journal of Official Statistics* 15: 329–334. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-note-on-post-stratification-when-analyzing-binary-survey-data-subject-to-nonresponse.pdf (accessed February 2020).

Zhang, L.-C. 2005. “On the bias in gross labour flow estimates due to nonresponse and mis-classification.” *Journal of Official Statistics* 21: 591–604. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/on-the-bias-in-gross-labour-flow-estimates-due-to-nonresponse-and-misclassification.pdf (accessed February 2020).

Zhang, L.-C., I. Thomsen, and Ø. Kleven. 2013. “On the use of auxiliary and para-data for dealing with non-sampling errors in household surveys.” *International Statistical Review* 81: 270–288. DOI: https://doi.org/10.1111/insr.12009.

Received January 2018
Revised November 2018
Accepted November 2019

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 173–196, http://dx.doi.org/10.2478/JOS-2020-0009

A Procedure for Estimating the Variance of the Population Mean in Rejective Sampling

Marius Stefan[1] *and Michael A. Hidiroglou*[2]

Rejective sampling was first introduced by Hájek in 1964 as a way to select a sample consisting uniquely of distinct units. If n denotes the fixed sample size, the n units are drawn independently with probabilities that may vary from unit to unit and the samples in which all units are not distinct are rejected. More generally, in rejective sampling, we select repeated samples according to a basic sampling design until a selected sample meets a specified balancing tolerance. Given a set of auxiliary variables, we consider a procedure in which the probability sample is rejected unless the sample mean of the auxiliary variables is within a specified distance of its corresponding population mean. The procedure represents an alternative to the well-known balanced cube method. In this article, we propose an estimator of the variance under the rejective sampling design. We also present the results of a Monte Carlo simulation study.

Key words: Balanced sampling; rejective sampling; normality assumption.

1. Introduction

Auxiliary data are commonly used these days in National Statistical Offices. The resulting estimators are either regression or calibration based. The regression-based procedures are chosen so as to improve the reliability of the estimators of the parameters of interest. The calibration-based procedures ensure that the weighted totals (means) of the auxiliary data exactly agree with their population totals (means). A deficiency of these procedures is that the final weight, the product of the original design weight times a factor accounting for the auxiliary data, can be negative. One way to avoid negative weights is to use the weight bounding algorithms given in Huang and Fuller (1978) or in Deville and Särndal (1992). Another way is to use the cube method given in Deville and Tillé (2004) or the rejective procedure developed by Fuller (2009). These methods will eliminate samples in which the final weights associated with some sampled units are negative. Comparisons between the cube and the rejective procedure have been given in Legg and Yu (2010) and Chauvet et al. (2017).

In this article, we focus on the rejective procedure developed by Fuller (2009) for estimating a population mean, $\bar{Y}_U = N^{-1} \sum_{i \in U} y_i$, where N is the population size and y denotes a characteristic of interest. A number of procedures closely related to Fuller's (2009) rejective procedure can be found in the literature. Hájek (1981, 66) gives a formal definition of rejective sampling resulting from a procedure in which a Poisson sample is

[1] Polytehnica University of Bucarest, Splaiul Independentei nr. 313, Bucarest, RO-060042, Romania. Email: mastefan@gmail.com
[2] Statistics Canada, Canada. Email: hidirog@yahoo.ca

rejected unless it contains exactly n units. The Hájek articles (1964, 1981) discuss the analysis of such samples extensively.

Denote as $p^b(s)$ the probability of selecting a specified sample s. The associated first order and second order inclusion probabilities are $\pi_i^b = \sum_{s \ni i} p^b(s)$ and $\pi_{ij}^b = \sum_{s \ni i,j} p^b(s)$, where $\pi_{ij}^b = \pi_i^b$ for $i = j$. A sample s^b selected from U with π_i^b as its first order selection probabilities is a *basic sample:* this term was introduced by Fuller (2009). The superscript b stands for *basic sampling design.* The expectation and variance operators under the basic sampling design will be denoted by $E_b(\cdot)$ and $V_b(\cdot)$.

We assume that a vector of auxiliary data, say $\boldsymbol{x}$, is available for each unit in the population. Let $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$ be the value of $\boldsymbol{x}$ for unit i in the population U. The known population mean is $\bar{\boldsymbol{X}}_U = N^{-1} \sum_{i \in U} \boldsymbol{x}_i$. For a given sample s^b, selected via the basic sampling design, $\hat{\bar{\boldsymbol{X}}}_{HT}^b = N^{-1} \sum_{i \in s^b} \boldsymbol{x}_i / \pi_i^b$ is the associated Horvitz-Thompson estimator of the population mean $\bar{\boldsymbol{X}}_U$. Its population variance is

$$V_b(\hat{\bar{\boldsymbol{X}}}_{HT}^b) = \frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} \Delta_{ij}^b \frac{\boldsymbol{x}_i}{\pi_i^b} \frac{\boldsymbol{x}_j^T}{\pi_j^b}$$

where $\Delta_{ij}^b = \pi_{ij}^b - \pi_i^b \pi_j^b$. The variance-covariance matrix $V_b(\hat{\bar{\boldsymbol{X}}}_{HT}^b)$ is assumed to be invertible.

The sample mean $\hat{\bar{\boldsymbol{X}}}_{HT}^b$ can be quite far away from the population mean $\bar{\boldsymbol{X}}_U$. The sample mean of samples selected with Deville and Tillé (2004)'s cube method or Fuller's (2009) rejective procedure will be "close" to the population mean.

In this article, we use Fuller's (2009) rejective procedure for selecting a sample. It is constructed based on a "distance" variable Q defined as

$$Q = (\hat{\bar{\boldsymbol{X}}}_{HT}^b - \bar{\boldsymbol{X}}_U)^T \left(V_b(\hat{\bar{\boldsymbol{X}}}_{HT}^b) \right)^{-1} (\hat{\bar{\boldsymbol{X}}}_{HT}^b - \bar{\boldsymbol{X}}_U). \tag{1}$$

A sample $s^b \subset U$ is initially selected using the basic sampling design. This sample is retained only if

$$Q \leq \gamma^2 \tag{2}$$

where $\gamma^2 > 0$ is a pre-specified constant. If the Q associated with the sample does not satisfy inequality (2), another sample s^b is selected. The process stops when inequality (2) is satisfied, and that sample is retained. Samples s^b that satisfy (2) will be denoted as s^r, where r stands for rejective sampling. The term *rejective sampling design* will be associated with samples s^r, selected using rule (2). The expectation and variance operators associated with the rejective sampling design will be respectively denoted by $E_r(\cdot)$ and $V_r(\cdot)$.

Given inequality (2), the distance between the sample mean $\hat{\bar{\boldsymbol{X}}}_{HT}^b$ and population mean $\bar{\boldsymbol{X}}_U$ is controlled via γ^2: the smaller γ^2 is, the closer $\hat{\bar{\boldsymbol{X}}}_{HT}^b$ is to $\bar{\boldsymbol{X}}_U$. An approximate rejection rate can be set by suitably selecting the value of γ^2. High rejection rates could provide high reductions in the variance. On the other hand, low rejection rates may not reduce the variance by a large amount, but provide sufficient comfort that a very poor sample will not be selected.

A sample is balanced on a vector of auxiliary variables, $\boldsymbol{x}$, if the following equations are satisfied:

$$\hat{\bar{\boldsymbol{X}}}_{HT}^b = \bar{\boldsymbol{X}}_U. \tag{3}$$

Balancing can be thought of as calibration at the design stage. Deville and Tillé (2004)'s cube method attempts to select balanced samples with pre-determined first order inclusion probabilities. Although the inclusion probabilities are exactly satisfied with the cube method, it may not be possible to satisfy equality in balancing Equation (3). Therefore, in cube sampling one has not control on the (possible) discrepancy between $\hat{\bar{X}}^{b}_{HT}$ and $\bar{X}_U$.

Given that γ is not zero, a rejective sample s^r selected using criterion (2) does not satisfy (3). A rejective sample s^r will satisfy (3) only when $\gamma \rightarrow 0$, or equivalently when the rejection rate tends to 100%. Therefore, a sample s^r is only approximately balanced, but the discrepancy between $\hat{\bar{X}}^{b}_{HT}$ and $\bar{X}_U$ can be controlled via the balancing tolerance γ. The drawback of rejective sampling is that the inclusion probabilities are usually unknown.

The weight associated with the basic sampling design for units belonging to a given rejective sample s^r is

$$w^{b}_{i,HT} = \frac{1}{\pi^{b}_{i}}. \tag{4}$$

An estimator of the population mean $\bar{Y}_U$ that uses these weights is

$$\hat{\bar{Y}}^{b}_{HT} = \frac{1}{N}\sum_{i \in s^r} w^{b}_{i,\mathrm{HT}} y_i. \tag{5}$$

The weighted estimator $\hat{\bar{Y}}^{b}_{HT}$ based on weights $w^{b}_{i,HT}$ is not a Horvitz-Thompson estimator as it is constructed with a rejective sample s^r. Chauvet et al. (2017) pointed out that $\hat{\bar{Y}}^{b}_{HT}$ can be biased given the rejective sampling procedure. Its rejective bias is given by:

$$B_r(\hat{\bar{Y}}^{b}_{HT}) = \frac{1}{N}\sum_{i \in U}\left(\frac{\pi^{r}_{i}}{\pi^{b}_{i}} - 1\right) y_i \tag{6}$$

where $\pi^{r}_{i} = \Sigma_{s^r \ni i} p^r(s^r)$ and $p^r(s^r)$ is the probability of selecting a specified sample s^r via the rejective sampling design. Thus, the bias can be large if some of the π^{r}_{i}/π^{b}_{i} ratios are unusually large. For Poisson sampling and simple random sampling without replacement π^{r}_{i}/π^{b}_{i} will be fairly close to 1.

The unknown first order inclusion probabilities π^{r}_{i} associated with the rejective sampling design may or may not be equal to π^{b}_{i} for $i = 1, \ldots, N$. The probability $p^r(s^r)$ of selecting a specific rejective sample s^r can be computed exactly if N and n are sufficiently small to enumerate all possible samples. It is then possible to compute the inclusion probabilities π^{r}_{i} and $\pi^{r}_{ij} = \Sigma_{s^r \ni i,j} p^r(s^r)$. The inclusion probabilities π^{r}_{i} and π^{r}_{ij} can be approximated via Monte Carlo methods if it is not possible to enumerate all samples. However, as Legg and You (2010) point out, simulating enough samples for a large population to give a precise estimate of the inclusion probability for each pair of units is impractical.

Another possibility is to approximate them. Recently, Chauvet el al. (2017) approximated the first order probability of inclusion π^{r}_{i} via the Edgeworth expansion for a basic sampling design that used Poisson sampling. Using these approximations, an approximately unbiased estimator for the population mean can be constructed given that rejective sampling was used to select the sample. They did not, however, approximate the joint selection inclusion probabilities π^{r}_{ij} necessary for measuring the precision of the population mean estimator under rejective sampling.

Let $\boldsymbol{x}^*$ be a vector of auxiliary data available at the estimation stage. The $\boldsymbol{x}^*$ vector is not necessarily identical to $\boldsymbol{x}$ used at the design stage in the definition of Q. We will suppose that $x \subseteq \boldsymbol{x}^*$.

There are a number of ways to use the auxiliary data $\boldsymbol{x}^*$. The GREG regression estimator, Särndal et al. (1992, chap. 6) is the one that we chose, as it is widely used. It is given by

$$\begin{aligned} \hat{\hat{Y}}^b_{GREG} &= \hat{\hat{Y}}^b_{HT} + (\bar{\boldsymbol{X}}^*_U - \hat{\hat{\boldsymbol{X}}}^{*b}_{HT})^T \hat{\boldsymbol{\beta}}^b_{GREG} \\ &= \frac{1}{N}\sum_{i\in s^r} w^b_{i,\mathrm{GREG}} y_i \end{aligned} \tag{7}$$

where

$$\begin{aligned} \hat{\boldsymbol{\beta}}^b_{GREG} &= \left(\sum_{i\in s^r} \frac{\boldsymbol{x}^*_i \boldsymbol{x}^{*T}_i}{\pi^b_i}\right)^{-1} \sum_{i\in s^r} \frac{\boldsymbol{x}^*_i y_i}{\pi^b_i} \quad \text{and} \\ w^b_{i,GREG} &= \frac{1}{\pi^b_i}\left\{1 + N(\bar{\boldsymbol{X}}^*_U - \hat{\hat{\boldsymbol{X}}}^{*b}_{HT})^T \left(\sum_{j\in s^r} \frac{\boldsymbol{x}^*_j \boldsymbol{x}^{*T}_j}{\pi^b_j}\right)^{-1} \boldsymbol{x}^*_i\right\} \end{aligned} \tag{8}$$

with $\hat{\hat{\boldsymbol{X}}}^{*b}_{HT} = N^{-1}\sum_{i\in s^r} \boldsymbol{x}^*_i/\pi^b_i$ and $\bar{\boldsymbol{X}}^*_U = N^{-1}\sum_{i\in U} \boldsymbol{x}^*_i$.

Fuller (2009) used the optimal estimator given by:

$$\hat{\hat{Y}}^b_{OPT} = \bar{\boldsymbol{X}}^{*T}_U \hat{\boldsymbol{\beta}}^b_{OPT} \tag{9}$$

where

$$\hat{\boldsymbol{\beta}}^b_{OPT} = \left(\sum_{i\in s^r} \frac{\phi^b_i \boldsymbol{x}^*_i \boldsymbol{x}^{*T}_i}{(\pi^b_i)^2}\right)^{-1} \sum_{i\in s^r} \frac{\phi^b_i \boldsymbol{x}^*_i y_i}{(\pi^b_i)^2}$$

and the ϕ^b_i's are constants determined by the design. These constants are $\phi^b_i = (1 - \pi^b_i), i = 1, \ldots, N$ for Poisson sampling, and $\phi^b_i = (N_h - 1)^{-1}(N_h - n_h)$ for the i^{th} element belonging to the h^{th} stratum for a stratified sampling design. The estimator $\hat{\hat{Y}}^b_{OPT}$ is design consistent under the basic procedure, (Fuller 2009), if $\mathrm{cov}(\hat{\hat{\boldsymbol{X}}}^{*bT}_{HT}, \hat{\hat{Y}}^b_{HT}) = O(n^{-1})$ and if there exists a vector $\boldsymbol{c}$ such that

$$\frac{\phi^b_i \boldsymbol{x}^{*T}_i \boldsymbol{c}}{(\pi^b_i)^2} = \frac{1}{\pi^b_i}. \tag{10}$$

Fuller (2009) proved that estimator $\hat{\hat{Y}}^b_{OPT}$ constructed with the rejective sample has the same limiting variance as the regression estimator that uses the first and second order inclusion probabilities associated with the basic selection procedure. Fuller et al. (2017) proposed a bootstrap procedure as an alternative way to estimate the variance of $\hat{\hat{Y}}^b_{OPT}$. The method, suggested for Poisson samples, also performs well with rejective Poisson samples.

We focus on estimating the variance of an estimator of $\bar{Y}_U$, say $\hat{\theta}$, given that rejective sampling has taken place. The variance estimator is a plug-in estimator obtained from a result for the rejective variance of $\hat{\theta}$ based on a normality assumption. The rejective variance estimator for $\hat{\theta}$ is expected to perform well if its distribution is normal or approximately normal. The theory will be applied to estimate the variance of $\hat{\theta} = \hat{\bar{Y}}^b_{HT}$ and $\hat{\theta} = \hat{\bar{Y}}^b_{GREG}$ under the rejective sampling knowing that, under fairly broad regularity conditions, the limiting distributions of these estimators are normal.

The article is structured as follows. In Section 2 we obtain $V_r(\hat{\theta})$ assuming that the joint distribution of $\hat{\theta}$ and the vector $\hat{\bar{\boldsymbol{X}}}^b_{HT} - \bar{\boldsymbol{X}}_U$ follows a multivariate normal distribution under the basic sampling design. We will also show under the normality assumption that $\hat{\theta}$ is unbiased under the rejective sampling design if $\hat{\theta}$ is unbiased under the basic sampling design. In Section 3 we show how an estimator $\hat{V}_r(\hat{\theta})$ of the rejective variance of $\hat{\theta}$ can be obtained. Section 4 provides the results of a simulation study that evaluates $V_r(\hat{\theta})$ and its estimator $\hat{V}_r(\hat{\theta})$. In this simulation, we focus on the weighted estimators $\hat{\bar{Y}}^b_{HT}$ and $\hat{\bar{Y}}^b_{GREG}$ defined in Equations (5) and (7) respectively. We considered two basic procedures: simple random sampling without replacement (SRSWOR) and Bernoulli sampling without replacement (BernWOR). Finally, Section 5 contains the concluding remarks.

2. Rejective Mean and Variance of $\hat{\boldsymbol{\theta}}$ Under the Normality Assumption

Recall that we denoted the mean and variance of an estimator $\hat{\theta}$ of $\bar{Y}_U$ under the rejective sampling as $E_r(\hat{\theta})$ and $V_r(\hat{\theta})$ respectively. The population parameters $E_r(\hat{\theta})$ and $V_r(\hat{\theta})$ are based on the unknown probabilities associated with the rejective sample. However, it is possible to express them in terms of the basic sampling distribution and Q given by (1). That is, $E_r(\hat{\theta})$ and $V_r(\hat{\theta})$ are set equal to the conditional mean and variance of $\hat{\theta}$ conditioned by $Q \leq \gamma^2$:

$$E_r(\hat{\theta}) = E_b(\hat{\theta}|Q \leq \gamma^2) \text{ and } V_r(\hat{\theta}) = V_b(\hat{\theta}|Q \leq \gamma^2). \tag{11}$$

We decompose the middle component of Q, the variance-covariance matrix $\left(V_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})\right)^{-1}$, using the Cholesky decomposition. That is,

$$\left(V_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})\right)^{-1} = \boldsymbol{P}^T\boldsymbol{P} \tag{12}$$

where $\boldsymbol{P}^T$ is a $p \times p$ lower triangular matrix that is invertible.

Next, define the p-dimensional vector $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$,

$$\boldsymbol{Z} = \boldsymbol{P}(\hat{\bar{\boldsymbol{X}}}^b_{HT} - \bar{\boldsymbol{X}}_U). \tag{13}$$

Using $\boldsymbol{Z}$, the quadratic form Q given by (1) can alternatively be written as:

$$Q = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^{p} Z_i^2 \tag{14}$$

where Z_i is the i^{th} component of the p-dimensional vector $\boldsymbol{Z}$.

Define the $p+1$ dimensional vector $\boldsymbol{W}$ as $\hat{\theta}$ augmented with $\boldsymbol{Z}$: that is $\boldsymbol{W} = (\hat{\theta}, \boldsymbol{Z}^T)^T$. Using (14), the conditional mean and the conditional variance in (11) are respectively given by

$$E_r(\hat{\theta}) = E_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \leq \gamma^2) \text{ and } V_r(\hat{\theta}) = V_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \leq \gamma^2)$$

The conditional mean and variance can be evaluated using the multivariate distribution of $\boldsymbol{W}$. We assume that the sampling distribution of $\boldsymbol{W}$ under the basic sampling design is a multivariate normal distribution of dimension $p + 1$. That is $\boldsymbol{W} \sim MVN_{p+1}(\boldsymbol{\mu}_w; \boldsymbol{\Sigma}_w)$ with $\boldsymbol{\mu}_w = E_b(\boldsymbol{W})$ and $\boldsymbol{\Sigma}_w = V_b(\boldsymbol{W})$. It follows from (13) that $E_b\,(\boldsymbol{Z}) = \boldsymbol{0}$ and $\boldsymbol{\mu}_w^T = (\mu_\theta, \boldsymbol{0}^T)$, where $\mu_\theta = E_b(\hat{\theta})$.

Since the matrix $\boldsymbol{P}$ is non-singular, Equation (12) can be expressed as $\boldsymbol{P}V_b(\hat{\hat{\boldsymbol{X}}}{}_{HT}^b)\boldsymbol{P}^T = \boldsymbol{I}_p$ where $\boldsymbol{I}_p$ is the identity matrix of order p. It follows that the variance-covariance matrix $V_b\,(\boldsymbol{Z})$ of $\boldsymbol{Z}$ under the basic sampling design is the identity matrix $\boldsymbol{I}_p$. Since $\boldsymbol{Z}$ is a component of $\boldsymbol{W}$, it follows that, under the basic sampling design, $\boldsymbol{Z}$ is distributed as $\boldsymbol{Z} \sim MVN_p\,(\boldsymbol{0}; \boldsymbol{I}_p)$ and its density function is $f_{\boldsymbol{Z}}(\boldsymbol{z}) = (\sqrt{2\pi})^{-p}e^{-\frac{1}{2}z^Tz}$.

Denote as $\sigma^2 = V_b(\hat{\theta})$ the variance of $\hat{\theta}$, and $\boldsymbol{\sigma}_{z\theta} = \text{cov}_b\,(\boldsymbol{Z},\hat{\theta})$ as the covariance between the random vector $\boldsymbol{Z}$ and the $\hat{\theta}$, under the basic sampling design. We have:

$$\sigma_{z\theta} = \text{cov}_b(\boldsymbol{Z}, \hat{\theta}) = \boldsymbol{P}\,\text{cov}_b(\hat{\hat{\boldsymbol{X}}}{}_{HT}^b, \hat{\theta}) \tag{15}$$

where $\text{cov}_b(\hat{\hat{\boldsymbol{X}}}{}_{HT}^b,\hat{\theta}) = (\text{cov}_b(\hat{\hat{X}}{}_{1HT}^b,\hat{\theta}), \ldots, \text{cov}_b(\hat{\hat{X}}{}_{pHT}^b,\hat{\theta}))^T$ with $\hat{\hat{X}}{}_{kHT}^b = N^{-1}\sum_{i\in s^b} x_{ik}/\pi_i^b$.

The variance-covariance matrix $\boldsymbol{\Sigma}_w$ and its inverse are given by:

$$\boldsymbol{\Sigma}_w = \begin{pmatrix} \sigma_\theta^2 & \boldsymbol{\sigma}_{z\theta}^T \\ \boldsymbol{\sigma}_{z\theta} & \boldsymbol{I}_p \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_w^{-1} = \begin{pmatrix} a^{-1} & -a^{-1}\boldsymbol{\sigma}_{z\theta}^T \\ -a^{-1}\boldsymbol{\sigma}_{z\theta} & \boldsymbol{M} \end{pmatrix} \tag{16}$$

where

$$a = \sigma_\theta^2 - \boldsymbol{\sigma}_{z\theta}^T\boldsymbol{\sigma}_{z\theta} \text{ and } \boldsymbol{M} = \boldsymbol{I}_p + \frac{1}{a}\sigma_{z\theta}\sigma_{z\theta}^T.$$

The determinant of matrix $\boldsymbol{\Sigma}_w$, $|\boldsymbol{\Sigma}_w|$, is equal to a.

Given the above preliminaries, we can now spell out our main result concerning the rejective mean $E_r(\hat{\theta})$ and the rejective variance $V_r(\hat{\theta})$ of estimator $\hat{\theta}$ using the normality of $\boldsymbol{W}$.

Theorem 1. Assume that the basic sampling distribution of vector $\boldsymbol{W}$ follows a multivariate normal distribution: that is $\boldsymbol{W} \sim MVN_{p+1}(\boldsymbol{\mu}_w; \boldsymbol{\Sigma}_w)$. Given that $\hat{\hat{\boldsymbol{X}}}{}_{HT}^b$ satisfies inequality (2), the conditional mean and variance of $\hat{\theta}$ are:

$$\text{i.}\quad E_r(\hat{\theta}) = E_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \leq \gamma^2) = \mu_\theta \tag{17}$$

and

$$\text{ii.}\quad V_r(\hat{\theta}) = V_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \leq \gamma^2) = \sigma_\theta^2 - \boldsymbol{\sigma}_{z\theta}^T\boldsymbol{\sigma}_{z\theta}\left(1 - \frac{\int_A z_1^2 f_{\boldsymbol{Z}}(\boldsymbol{z})d\boldsymbol{z}}{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})d\boldsymbol{z}}\right) \tag{18}$$

where $A = \{(z_1, \ldots, z_p) \in \mathbb{R}^p | z_1^2 + \ldots + z_p^2 \leq \gamma^2\}$, and z_1 is the first component of $\boldsymbol{z}^T = (z_1, \ldots, z_p)$.

Proof: See Appendix A (Subsection 6.1).

Deville and Tillé (2005) used a similar normality assumption on the distribution of an augmented vector to evaluate the variance in the case of balanced sampling. Deville and Tillé (2005) obtained four alternative approximations for the variance of the Horvitz-Thompson (Horvitz and Thompson 1952) under balanced sampling that allowed them to construct variance estimators that do not depend on second order inclusion probabilities.

It follows from part i. of Theorem 1 that $\hat{\theta}$ will be unbiased under the rejective sampling design if $\hat{\theta}$ is an unbiased estimator of the population mean under the basic sampling design. Part ii. of Theorem 1 provides a formula for computing the variance of $\hat{\theta}$ under the rejective sampling and the normality assumption. Notice that if the normality assumption $\boldsymbol{W} \sim MVN_{p+1}(\boldsymbol{\mu}_w; \boldsymbol{\Sigma}_w)$ only holds approximately, then Equations (17) and (18) respectively represent approximations of $E_r\,(\hat{\theta})$ and $V_r\,(\hat{\theta})$.

Next, we show how the integrals in Equation (18) can be computed.

Proposition 1: For a positive integer $n \geq 0$, let $J_n(\gamma)$ the integral given by $J_n(\gamma) = \int_0^{\gamma} r^n e^{-\frac{r^2}{2}} dr$.

i. The integral $J_n(\gamma)$ obeys the following recursive relation

$$J_{n+1}(\gamma) = nJ_{n-1}(\gamma) - \gamma^n e^{-\frac{\gamma^2}{2}}, \text{where } n \geq 1. \tag{19}$$

The first two $J_n(\gamma)$ values are computed as: $J_0(\gamma) = \sqrt{2\pi}(\Phi(\gamma) - 0.5)$ and $J_1(\gamma) = 1 - e^{-\frac{\gamma^2}{2}}$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

ii. Given A and $f_{\boldsymbol{Z}}(\boldsymbol{z})$ as defined in Theorem 1, we have that:

$$1 - \frac{\int_A z_1^2 f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z}}{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z}} = g(\gamma, p) \tag{20}$$

where $g(\gamma, p) = \frac{\gamma^p e^{-\frac{\gamma^2}{2}}}{pJ_{p-1}(\gamma)}$.

Proof: See Appendix B (Subsection 6.2).

Using (12) and (15), the product $\boldsymbol{\sigma}_{z\theta}^T \boldsymbol{\sigma}_{z\theta}$ is given by:

$$\boldsymbol{\sigma}_{z\theta}^T \boldsymbol{\sigma}_{z\theta} = \text{cov}_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b, \hat{\theta})^T \left(V_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b)\right)^{-1} \text{cov}_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b, \hat{\theta}).$$

Recall that $\sigma_\theta^2 = V_b(\hat{\theta})$. Using (18) and (20), $V_r(\hat{\theta})$ can be expressed as:

$$V_r(\hat{\theta}) = V_b(\hat{\theta}) - \text{cov}_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b, \hat{\theta})^T \left(V_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b)\right)^{-1} \text{cov}_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b, \hat{\theta}) g(\gamma, p). \tag{21}$$

The variance matrix $V_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b)$ and its inverse are positive definite. The quadratic form $\boldsymbol{u}^T \left(V_b(\hat{\hat{\boldsymbol{X}}}_{HT}^b)\right)^{-1} \boldsymbol{u}$ is greater or equal to zero for any vector $\boldsymbol{u}$. Since $g(\gamma, p)$ is positive, the second term in Equation (21) is greater or equal to zero, and it follows that $V_r(\hat{\theta}) \leq V_b(\hat{\theta})$. If an estimator $\hat{\theta}$ satisfies the conditions of Theorem 1, Equation (21) shows that its rejective variance is equal or smaller than its basic variance.

Table 1. Values of $g(\gamma, p)$ as a function of p and Rejection Rate (RR).

	Rejection Rate (RR)		
p	**0.80**	**0.90**	**0.95**
1	0.978	0.994	0.998
2	0.892	0.948	0.974
3	0.810	0.887	0.931
4	0.745	0.830	0.885
5	0.692	0.782	0.842

Notice that $\lim_{\gamma\to\infty} J_0(\gamma) = \sqrt{\pi/2}$ and $\lim_{\gamma\to\infty} J_1(\gamma) = 1$. Using these limits and the recursive formula given by (19), it follows that $\lim_{\gamma\to\infty} J_p(\gamma)$ is finite for any fixed p. This implies that $\lim_{\gamma\to\infty} g(\gamma, p) = 0$ and that $\lim_{\gamma\to\infty} V_r(\hat{\theta}) = V_b(\hat{\theta})$. This means that when γ is large, the basic and the rejective sampling plans are similar, and consequently there is little gain in precision in using rejective sampling.

Since $\lim_{\gamma\to 0} J_0(\gamma) = \lim_{\gamma\to 0} J_1(\gamma) = 0$, it follows that $\lim_{\gamma\to 0} J_p(\gamma) = 0$ for any fixed p. Hence:

$$\lim_{\gamma\to 0} g(\gamma, p) = \lim_{\gamma\to 0} \frac{\gamma^p}{pJ_{p-1}(\gamma)} = \lim_{\gamma\to 0} \frac{p\gamma^{p-1}}{p\gamma^{p-1}e^{-\frac{\gamma^2}{2}}} = 1. \tag{22}$$

Using Equations (21) and (22), the minimum rejective variance of $\hat{\theta}$, $V_r^{\min}(\hat{\theta})$, is attained when γ tends to zero. This minimum is:

$$V_r^{\min}(\hat{\theta}) = V_b(\hat{\theta}) - \mathrm{cov}_b(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})^T \left(V_b(\hat{\hat{\boldsymbol{X}}}^b_{HT})\right)^{-1} \mathrm{cov}_b(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta}). \tag{23}$$

The rejective variance $V_r(\hat{\theta})$ is a function of $g(\gamma, p)$. Table 1 illustrates the behavior of $g(\gamma, p)$ as a function of p and the rejection rate (RR). An approximate rejection rate can be set using the quantiles of the $\chi^2(p)$ distribution. Recall that γ^2 given in Equation (2) is a pre-specified constant. As γ tends to zero, the rejection rate tends to 100% for a fixed p. On the other hand, for a fixed rejection rate, γ increases as p increases.

The results given in Table 1 support Equation (22). That is, for a given p, as the rejection rate increases, or equivalently as γ tends to zero, $g\ (\gamma, p)$, tends to 1.

For a given rejection rate, the term $g(\gamma, p)$ decreases as p increases, or equivalently as γ tends to infinity. This implies that increasing the number of variables in the distance variable Q defined by (1) does not necessarily result in reductions of the rejective variance $V_r(\hat{\theta})$. In order to decrease $V_r(\hat{\theta})$, one has to make sure that the variance-covariance term $\mathrm{cov}_b(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})^T \left(V_b(\hat{\hat{\boldsymbol{X}}}^b_{HT})\right)^{-1} \mathrm{cov}_b(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ is large.

3. Estimation of $V_r(\hat{\boldsymbol{\theta}})$

In this section we construct an estimator $\hat{V}_r(\hat{\theta})$ of the rejective variance $V_r(\hat{\theta})$. It is obtained by plugging into Equation (21) the estimators of its components under the basic sampling design.

Three population parameters in Equation (21) are to be estimated. They are: i. the variance-covariance matrix $V_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})$; ii. the variance $V_b(\hat{\theta})$; and iii. the covariance vector $\operatorname{cov}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$. These parameters correspond to the basic sampling design. They are estimated using inclusion probabilities π_i^b and π_{ij}^b associated with units that belong to the rejective sample s^r.

An estimator of $V_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})$ is

$$\hat{V}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}) = \frac{1}{N^2} \sum_{i \in s^r} \sum_{j \in s^r} \frac{\Delta_{ij}^b}{\pi_{ij}^b} \frac{\boldsymbol{x}_i}{\pi_i^b} \frac{\boldsymbol{x}_j^T}{\pi_j^b}. \tag{24}$$

Let $\hat{V}_b(\hat{\theta})$ and $\widehat{\operatorname{cov}}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ be the respective estimators of $V_b(\hat{\theta})$ and $\operatorname{cov}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ under the basic design. Then, an estimator for $V_b(\hat{\theta})$ is

$$\hat{V}_r(\hat{\theta}) = \hat{V}_b(\hat{\theta}) - \widehat{\operatorname{cov}}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})^T \left(\hat{V}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}) \right)^{-1} \widehat{\operatorname{cov}}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta}) g(\gamma, p). \tag{25}$$

Remark 1: The plug-in estimator $\hat{V}_r(\hat{\theta})$ is obtained by replacing the unknown parameters $V_b(\hat{\theta})$, $\operatorname{cov}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ and $V_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})$ by their respective estimators $\hat{V}_b(\hat{\theta})$, $\widehat{\operatorname{cov}}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ and $\hat{V}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})$. They are unbiased under the basic procedure. That is, if one was to average their values over the set of samples s^b, their expectation would be $V_b(\hat{\theta})$, $\operatorname{cov}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ and $V_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})$ respectively. However, given that we compute their values with the rejective samples s^r, some bias may occur.

Remark 2: Part i. of Theorem 1 can be applied to each of the three estimators, $\hat{V}_b(\hat{\theta})$, $\widehat{\operatorname{cov}}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ and $\hat{V}_b(\hat{\bar{\boldsymbol{X}}}^b_{HT})$. If their basic sampling distribution is approximately normal, they will be approximately unbiased when they are computed using the s^r samples. This in turn will provide an estimator $\hat{V}_r(\hat{\theta})$ of $V_r(\hat{\theta})$ that will be approximately unbiased. Results in Section 4 illustrate that the bias of $\hat{V}_r(\hat{\theta})$ decreases as n increases.

4. Simulation Study

In this section we report on a Monte Carlo simulation that evaluates the Equation (21) for the rejective variance $V_r(\hat{\theta})$ and its estimator $\hat{V}_r(\hat{\theta})$, when $\hat{\theta} = \hat{\bar{Y}}^b_{HT}$ and $\hat{\theta} = \hat{\bar{Y}}^b_{GREG}$. We considered two basic sampling designs with equal probabilities: one of fixed sample size and the other one with random sample size. These two basic sampling designs are:

1. SRSWOR of fixed sample size n with inclusion probabilities $\pi_i^b = n/N$, and
2. BernWOR of mean sample size n with equal inclusion probabilities $\pi_i^b = n/N$.

Next, we describe how the population variances and their estimators are obtained for the two estimators $\hat{\bar{Y}}^b_{HT}$ and $\hat{\bar{Y}}^b_{GREG}$ and the two sampling designs. We denote by $V_{hb}(\hat{\bar{\boldsymbol{X}}}^b_{HT})$, $h = 1, 2$ the variance-covariance matrix of $\hat{\bar{\boldsymbol{X}}}^b_{HT} = N^{-1} \sum_{i \in s^b} \boldsymbol{x}_i / \pi_i^b$ under the basic sampling design. The subscript h is set to 1 when SRSWOR is used as the basic sampling design. It is set to 2 when BernWOR is used as the basic sampling design. For SRSWOR, the matrix $V_{1b}(\hat{\bar{\boldsymbol{X}}}^b_{HT})$ is $(1-f)n^{-1}(N-1)^{-1} \sum_{i \in U} (\boldsymbol{x}_i - \bar{\boldsymbol{X}}_U)(\boldsymbol{x}_i - \bar{\boldsymbol{X}}_U)^T$. For BernWOR, $V_{2b}(\hat{\bar{\boldsymbol{X}}}^b_{HT})$ is $(1-f)n^{-1}N^{-1} \sum_{i \in U} \boldsymbol{x}_i \boldsymbol{x}_i^T$, where $f = n/N$ (see Särndal et al. 1992, Result 5.4.1., 170).

The estimated population mean $\hat{\hat{\boldsymbol{X}}}^b_{HT}$ was based on the two-dimensional vector $\boldsymbol{x}_k = (x_{1k}, x_{2k})^T$ for SRSWOR, to ensure that the matrix $V_{1b}(\hat{\hat{\boldsymbol{X}}}^b_{HT})$ would be non-singular. For BernWOR, $\hat{\hat{\boldsymbol{X}}}^b_{HT}$ was based on the three-dimensional vector $\boldsymbol{x}_k = (1, x_{1k}, x_{2k})^T$. The three-dimensional vector of auxiliary variables $\boldsymbol{x}^*_k = (1, x_{1k}, x_{2k})^T$ was used to construct $\hat{\hat{Y}}^b_{GREG}$ for both SRSWOR and BernWOR.

We computed the population variance $V_b(\hat{\theta})$ and the population covariance $\text{cov}_b(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ for the two basic sampling designs, SRSWOR with fixed sample size n and BernWOR with expected sample size n. We denote them as $V_{hb}(\hat{\theta})$ and $\text{cov}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ for $h = 1, 2,$ where $\hat{\theta}$ is either $\hat{\hat{Y}}^b_{HT} = N^{-1} \sum_{i \in s^r} y_i / \pi^b_i$ or $\hat{\hat{Y}}^b_{GREG} = \hat{\hat{Y}}^b_{HT} + (\bar{\boldsymbol{X}}^*_U - \hat{\hat{\boldsymbol{X}}}^{*b}_{HT})^T \hat{\boldsymbol{\beta}}^b_{GREG}$. Given that the population data, y and x^* are known, it is possible to compute an exact population variance for $\hat{\hat{Y}}^b_{HT}$ (see Särndal et al. 1992, Result 2.8.1., 43) and an exact population covariance $\text{cov}_b(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\hat{Y}}^b_{HT})$ (see Särndal et al. 1992, Result 5.4.1., 170). On the other hand, this is not possible for the regression estimator $\hat{\hat{Y}}^b_{GREG}$. Särndal et al. (1992, Result 6.6.1., 235) provide an approximate population variance for $\hat{\hat{Y}}^b_{GREG}$ based on a Taylor linearization of $\hat{\hat{Y}}^b_{GREG}$. The Taylor expansion can also be used to approximate $\text{cov}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\hat{Y}}^b_{GREG})$. However, these approximations are only reasonable for moderate to large values of n.

Given this drawback, we chose to compute Monte Carlo values for $V_{hb}(\hat{\hat{Y}}^b_{GREG})$ and $\text{cov}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\hat{Y}}^b_{GREG})$ by sampling a large number of basic samples s^b from the population U. Although we could have computed $V_{hb}(\hat{\hat{Y}}^b_{HT})$ and $\text{cov}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\hat{Y}}^b_{HT})$ exactly, we evaluated them using the large number of samples s^b selected from the population.

The estimators $\hat{V}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT})$, $\hat{V}_{hb}(\hat{\theta})$ and $\widehat{\text{cov}}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ were obtained via the standard theory in Särndal et al. (1992) for the Horvitz-Thompson and GREG estimators. They were computed using the data of the selected rejective samples s^r and the inclusion probabilities inherited from the basic sample design. Estimators $\hat{V}_{hb}(\hat{\hat{Y}}^b_{HT})$ are based on Result 2.8.1. (Särndal et al. 1992, 43). Estimators $\hat{V}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT})$ and $\widehat{\text{cov}}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\hat{Y}}^b_{HT})$ are based on Result 5.4.1. (Särndal et al. 1992, 170). For the GREG estimator, $\hat{V}_{hb}(\hat{\hat{Y}}^b_{GREG})$ and $\widehat{\text{cov}}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\hat{Y}}^b_{GREG})$ use Result 6.6.1. (Särndal et al. 1992, 235).

We now explain how the simulation was carried out. The dependent variable y was generated using the following linear model:

$$y_k = 1 + x_{1k} + 5x_{2k} + e_k, k = 1, \ldots, N \text{ with } e_k \sim N(0; 1) \tag{26}$$

where $N = 1{,}000$. The population of y-values generated by (26) has mean $\bar{Y}_U = 21.147$ and variance $S^2_{yU} = 234.72$. The population coefficient of determination associated with model (26) is $R^2 = 99.5\%$. The values of x_1 are generated using a normal distribution of mean 10 and variance 1. The values of x_2 are generated using a gamma distribution of mean 2 and variance 10. The differential mix of distributions for generating the independent variables x was chosen to illustrate how well Theorem 1 held under non-normal, asymmetric distributions.

We carried out two separate simulations. The objective of the first simulation was to compute accurately the population variance $V_{hb}(\hat{\theta})$ and covariance $\text{cov}_{hb}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ for $h = 1, 2$. We selected $L = 500{,}000$ basic samples, s^b, from the population U for each of the sampling procedures, SRSWOR ($h = 1$) and BernWOR ($h = 2$). For the l^{th} selected sample, we computed $\hat{\theta}^{(l)}$, where $\hat{\theta}^{(l)} = \hat{\hat{Y}}^{b(l)}_{HT}$ or $\hat{\theta}^{(l)} = \hat{\hat{Y}}^{b(l)}_{GREG}$. The resulting Monte Carlo variances and covariances are given by

$$V_{hb}^{MC1}(\hat{\theta}) = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}^{(l)} - E_{hb}^{MC1}(\hat{\theta}^{(l)})\right)^{2}$$

and

$$\text{cov}_{hb}^{MC1}(\hat{\hat{\boldsymbol{X}}}_{HT}^{b}, \hat{\theta}) = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{\hat{\boldsymbol{X}}}_{HT}^{b(l)} - E_{hb}^{MC1}(\hat{\hat{\boldsymbol{X}}}_{HT}^{b})\right)\left(\hat{\theta}^{(l)} - E_{hb}^{MC1}(\hat{\theta})\right)$$

where $E_{hb}^{MC1}(\hat{\theta}) = \sum_{l=1}^{L}\hat{\theta}^{(l)}/L$ and $E_{hb}^{MC1}(\hat{\hat{\boldsymbol{X}}}_{HT}^{b}) = \sum_{l=1}^{L}\hat{\hat{\boldsymbol{X}}}_{HT}^{b(l)}/L$ with $\hat{\theta}^{(l)}$ and $\hat{\hat{\boldsymbol{X}}}_{HT}^{b(l)}$ the respective values of $\hat{\theta}$ and $\hat{\hat{\boldsymbol{X}}}_{HT}^{b}$ for the l^{th} selected basic sample s^{b}.

The values of the rejective variance $V_{r}(\hat{\theta})$ are computed via Equation (21) that depends on $V_{hb}^{MC1}(\hat{\hat{\boldsymbol{X}}}_{HT}^{b})$, $V_{hb}^{MC1}(\hat{\theta})$ and $\text{cov}_{hb}^{MC1}(\hat{\hat{\boldsymbol{X}}}_{HT}^{b}, \hat{\theta})$. These variances are denoted as $V_{1r}(\hat{\theta})$ for SRSWOR and $V_{2r}(\hat{\theta})$ for BernWOR.

The objective of the second simulation was to test how accurate the rejective variance $V_{hr}(\hat{\theta})$ based on (21) and its estimator $\hat{V}_{hr}(\hat{\theta})$ were under the rejective sampling. The γ^{2} value was chosen so as to yield rejective samples with a 90% rejection rate for both SRSWOR and BernWOR. The γ^{2} value was chosen by trial and error for both of these sampling designs. For SRSWOR, a value of γ^{2} equal to 0.05 ($\gamma_{1}^{2} = 0.05$) resulted in a rejection rate of 90% of the basic samples s^{b}. The γ^{2} had to be increased to 0.22 ($\gamma_{2}^{2} = 0.22$) for BernWOR to obtain the 90% rejection rate. Noting that Q is approximately distributed as a $\chi^{2}(p)$ where p is the length of vector $\boldsymbol{x}_{k}$, we could have obtained the value of γ^{2} using the quantiles of χ^{2} (p). However, the trial and error approach proved to be better in getting exact 90% rejection rates.

Given the required γ^{2} for each of the basic sampling procedures, SRSWOR and BernWOR, we selected $L = 500{,}000$ rejective samples s^{r} from a large number of basic samples s^{b}, based on criterion (2). For each selected rejective sample s^{r}, we computed $\hat{V}_{hb}(\hat{\hat{\boldsymbol{X}}}_{HT}^{b})$, $\hat{V}_{hb}(\hat{\theta})$ and $\widehat{\text{cov}}_{hb}(\hat{\hat{\boldsymbol{X}}}_{HT}^{b}, \hat{\theta})$, for $h = 1, 2$. The estimated values of $V_{r}(\hat{\theta})$, denoted by $\hat{V}_{1r}(\hat{\theta})$ for SRSWOR and by $\hat{V}_{2r}(\hat{\theta})$ for BernWOR, were obtained using Equation (25) of Section 3.

For the ℓ^{th} selected rejective sample let:

- $\hat{\theta}^{(\ell)}$ be the value of estimator $\hat{\theta}$ for the ℓ^{th} sample,
- $\hat{V}_{hr}(\hat{\theta})^{(\ell)}$ be the value of estimator $\hat{V}_{hr}(\hat{\theta})$, $h = 1, 2$ for the ℓ^{th} sample, and
- $\hat{V}_{hb}(\hat{\theta})^{(\ell)}$ be the value of estimator $\hat{V}_{hb}(\hat{\theta})$, $h = 1, 2$ for the ℓ^{th} sample.

For $h = 1, 2$ the Monte Carlo expectation and the Monte Carlo variance of estimator $\hat{\theta}$ are computed as:

$$E_{hr}^{MC2}(\hat{\theta}) = \frac{1}{L}\sum_{\ell=1}^{L}\hat{\theta}^{(\ell)} \text{ and } V_{hr}^{MC2}(\hat{\theta}) = \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{\theta}^{(\ell)} - E_{hr}^{MC2}(\hat{\theta})\right)^{2}.$$

The Monte Carlo expectation of estimators $\hat{V}_{hr}(\hat{\theta})$ and $\hat{V}_{hb}(\hat{\theta})$ are computed as:

$$E_{hr}^{MC2}\left(\hat{V}_{hr}(\hat{\theta})\right) = \frac{1}{L}\sum_{\ell=1}^{L}\hat{V}_{hr}(\hat{\theta})^{(\ell)} \text{ and } E_{hr}^{MC2}\left(\hat{V}_{hb}(\hat{\theta})\right) = \frac{1}{L}\sum_{\ell=1}^{L}\hat{V}_{hb}(\hat{\theta})^{(\ell)}.$$

4.1. Rejective Sampling Using SRSWOR

The following observations can be made from Table 2. The estimators $\hat{\bar{Y}}^b_{HT}$ and $\hat{\bar{Y}}^b_{GREG}$ are virtually unbiased under the rejective sampling design. This follows because the Monte Carlo expectation of $\hat{\bar{Y}}^b_{HT}$ and $\hat{\bar{Y}}^b_{GREG}$, represented by $E^{MC2}_{1r}(\hat{\theta})$, is quite close to the true population mean $\bar{Y}_U = 21.147$, for all values of n.

We compare the various combinations of population variances $V^{MC1}_{1b}(\hat{\theta})$, $V^{MC2}_{1r}(\hat{\theta})$ and $V_{1r}(\hat{\theta})$.

The use of rejective sampling results in gains in terms of population variance. This follows by comparing the variance $V^{MC1}_{1b}(\hat{\theta})$ under SRSWOR to the Monte Carlo variance $V^{MC2}_{1r}(\hat{\theta})$ under rejective sampling. The gains are quite large for the $\hat{\bar{Y}}^b_{HT}$ estimator: this makes sense, as we have drawn samples whose mean, $\hat{\bar{\boldsymbol{X}}}^b_{HT}$, is quite close to the population mean $\hat{\bar{\boldsymbol{X}}}_U$. On the other hand, for the regression estimator $\hat{\bar{Y}}^b_{GREG}$, the gains are not as large since it uses auxiliary data that are well correlated with y.

The value of $V_{1r}(\hat{\theta})$ is compared to the Monte Carlo variance of $\hat{\theta}$ under rejective sampling, $V^{MC2}_{1r}(\hat{\theta})$. Recall that the two components of $V_{1r}(\hat{\theta})$, defined by Equation (21), were obtained via simulation under SRSWOR. The value of $V_{1r}(\hat{\theta})$ is quite close to $V^{MC2}_{1r}(\hat{\theta})$ for $\hat{\theta} = \hat{\bar{Y}}^b_{HT}$. For $\hat{\theta} = \hat{\bar{Y}}^b_{GREG}$, the largest difference between $V_{1r}(\hat{\bar{Y}}^b_{GREG})$ and $V^{MC2}_{1r}(\hat{\bar{Y}}^b_{GREG})$ occurs when $n = 10$: in this case we have that $V_{1r}(\hat{\bar{Y}}^b_{GREG}) = 0.1792$ and $V^{MC2}_{1r}(\hat{\bar{Y}}^b_{GREG}) = 0.0947$. This means that for $n = 10$, the normality assumption of Theorem 1 is far from being satisfied. Theorem 1 is not applicable for small values of n. For moderate to large values of n, $V_{1r}(\hat{\bar{Y}}^b_{GREG})$ and $V^{MC2}_{1r}(\hat{\bar{Y}}^b_{GREG})$ are approximately equal.

The covariance in the second term of $V_{1r}(\hat{\bar{Y}}^b_{GREG})$ given by Equation (21) can be approximated using Result 6.6.1. (Särndal et al. 1992, 235). That is $\mathrm{cov}_{1b}(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\bar{Y}}^b_{GREG}) \approx (1-f)n^{-1}\boldsymbol{S}_{xE}$ where $\boldsymbol{S}_{xE} = (N-1)^{-1}(\sum_{i\in U} E_i\boldsymbol{x}_i - N\bar{E}\bar{\boldsymbol{X}}_U)$, $\boldsymbol{\beta}_{0,GREG} = \left(\sum_{i\in U} \boldsymbol{x}^*_i\boldsymbol{x}^{*T}_i\right)^{-1}\sum_{i\in U}\boldsymbol{x}^*_i y_i$, $E_i = y_i - \boldsymbol{x}^{*T}_i\boldsymbol{\beta}_{0,GREG}$ and $\bar{E} = N^{-1}\sum_{i\in U} E_i$. It can be shown that $\sum_{i\in U} E_i\boldsymbol{x}_i = 0$ and $\sum_{i\in U} E_i = 0$, using the system of equations $\sum_{i\in U}(y_i - \boldsymbol{x}^{*T}_i\boldsymbol{\beta}_{0,GREG})\boldsymbol{x}^*_i = 0$ and $\boldsymbol{x}_i \subset \boldsymbol{x}^*_i$. Hence, $\boldsymbol{S}_{xE} = 0$ and the second term in Equation (21) is approximately zero. This explains why the variances $V^{MC1}_{1b}(\hat{\bar{Y}}^b_{GREG})$ and $V_{1r}(\hat{\bar{Y}}^b_{GREG})$ are approximately equal for all values of n considered in Table 2. On the other hand, we noticed that $V_{1r}(\hat{\bar{Y}}^b_{GREG})$ and $V^{MC2}_{1r}(\hat{\bar{Y}}^b_{GREG})$ are getting closer as n increases. Consequently, the rejective variance of $\hat{\bar{Y}}^b_{GREG}$ tends to its basic variance if SRSWOR is the basic sampling design. This is not the case for $\hat{\bar{Y}}^b_{HT}$, as Table 2 clearly illustrates.

The estimator $\hat{V}_{1b}(\hat{\theta})$ is unbiased for $V_{1b}(\hat{\theta})$ under the SRSWOR basic sampling design. However, when its Monte Carlo mean is computed over the set of the rejective samples s^r, this estimator has some bias. This is readily observed from Table 2 by comparing the value of $E^{MC2}_{1r}[\hat{V}_{1b}(\hat{\theta})]$ to the variance $V^{MC1}_{1b}(\hat{\theta})$. For $n = 10$ the bias is large for both estimators, $\hat{\theta} = \hat{\bar{Y}}^b_{HT}$ and $\hat{\theta} = \hat{\bar{Y}}^b_{GREG}$. Under the rejective sampling, as n increases, the bias of $\hat{V}_{1b}(\hat{\theta})$ as an estimator of $V_{1b}(\hat{\theta})$ decreases (see Remark 2). A similar conclusion holds for $\widehat{\mathrm{cov}}_{1b}(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ as an estimator of $\mathrm{cov}_{1b}(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ under rejective sampling (numerical results not shown).

We next turn to the estimators of the rejective variance of $\hat{\theta}$. The variance of $\hat{\theta}$ under rejective sampling can be estimated by $\hat{V}_{1b}(\hat{\theta})$ or by $\hat{V}_{1r}(\hat{\theta})$. The first estimator, $\hat{V}_{1b}(\hat{\theta})$, ignores that rejective sampling has taken place. The second estimator, $\hat{V}_{1r}(\hat{\theta})$, obtained via estimators $\hat{V}_{1b}(\hat{\bar{\boldsymbol{X}}}^b_{HT})$, $\hat{V}_{1b}(\hat{\theta})$ and $\widehat{\mathrm{cov}}_{1b}(\hat{\bar{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$, accounts for the rejective sampling.

Table 2. Basic sampling design SRSWOR and RR $= 0.90$ ($\gamma_1^2 = 0.05$).

	Estimator							
	$\hat{\theta} = \hat{\bar{Y}}_{HT}^{b}$				$\hat{\theta} = \hat{\bar{Y}}_{GREG}^{b}$			
Summary statistics	$\boldsymbol{n = 10}$	$\boldsymbol{n = 20}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$	$\boldsymbol{n = 10}$	$\boldsymbol{n = 20}$	$\boldsymbol{n = 50}$	$\boldsymbol{n = 100}$
$E_{1r}^{MC2}(\hat{\theta})$	21.125	21.134	21.141	21.144	21.146	21.146	21.146	21.147
$V_{1b}^{MC1}(\hat{\theta})$	23.234	11.518	4.445	2.114	0.1794	0.0552	0.0191	0.0088
$V_{1r}^{MC2}(\hat{\theta})$	0.3869	0.1903	0.0736	0.0351	0.0947	0.0473	0.0182	0.0086
$V_{1r}(\hat{\theta})$ (eq. (21))	0.3833	0.1899	0.0734	0.348	0.1792	0.0552	0.0191	0.0088
$E_{1r}^{MC2}[\hat{V}_{1r}(\hat{\theta})]$	0.3244	0.1752	0.0714	0.0344	0.0738	0.0421	0.0175	0.0085
$E_{1r}^{MC2}[\hat{V}_{1b}(\hat{\theta})]$	20.207	10.741	4.351	2.090	0.0739	0.0421	0.0175	0.0085

Note: $\bar{Y}_U = 21.147$, $S_{yU}^2 = 234.72$ and $R^2 = 99.5\%$.

The estimated variance of $\hat{\hat{Y}}^b_{HT}$ can be computed as $\hat{V}_{1b}(\hat{\hat{Y}}^b_{HT})$ or as $\hat{V}_{1r}(\hat{\hat{Y}}^b_{HT})$. Given that the population variance of $\hat{\hat{Y}}^b_{HT}$ is $V^{MC2}_{1r}(\hat{\theta})$, $\hat{V}_{1b}(\hat{\hat{Y}}^b_{HT})$ is highly biased for all values of n. On the other hand, $\hat{V}_{1r}(\hat{\hat{Y}}^b_{HT})$ has a small bias since, under rejective sampling, estimators $\hat{V}_{1b}(\hat{\hat{\boldsymbol{X}}}^b_{HT})$, $\hat{V}_{1b}(\hat{\theta})$ and $\widehat{\text{cov}}_{1b}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ are biased for $V_{1b}(\hat{\hat{\boldsymbol{X}}}^b_{HT})$, $V_{1b}(\hat{\theta})$ and $\text{cov}_{1b}(\hat{\hat{\boldsymbol{X}}}^b_{HT}, \hat{\theta})$ respectively (see Remark 1). As n increases, the bias of these estimators tends to zero. It follows that estimator $\hat{V}_{1r}(\hat{\hat{Y}}^b_{HT})$ becomes unbiased for the rejective variance of $\hat{\hat{Y}}^b_{HT}$ as n becomes large.

There are two alternative estimators for estimating the variance of $\hat{\hat{Y}}^b_{GREG}$ under the rejective sampling:

i. $\hat{V}_{1b}(\hat{\hat{Y}}^b_{GREG})$ based on the standard theory and probabilities (π^b_i, π^b_{ij}) of the basic sampling procedure and
ii. $\hat{V}_{1r}(\hat{\hat{Y}}^b_{GREG})$ obtained using Equation (25).

From a bias point of view, they are very similar if we compare $E^{MC2}_{1r}[\hat{V}_{1b}(\hat{\hat{Y}}^b_{GREG})]$ to $E^{MC2}_{1r}[\hat{V}_{1r}(\hat{\hat{Y}}^b_{GREG})]$. This result is not surprising. We noticed that for the GREG estimator, its rejective and basic variances are getting closer. On the other hand, under rejective sampling, the bias of $\hat{V}_{1b}(\hat{\hat{Y}}^b_{GREG})$ as an estimator of $V_{1b}(\hat{\hat{Y}}^b_{GREG})$ tends to zero. It follows that the bias of $\hat{V}_{1b}(\hat{\hat{Y}}^b_{GREG})$ as an estimator of $V_{1r}(\hat{\hat{Y}}^b_{GREG})$ decreases as n becomes larger. Fuller (2009) proved that this would be the case for the optimal estimator, $\hat{\hat{Y}}^b_{OPT}$, defined in Equation (9).

A referee pointed out that the proposed variance estimator in (25) shows large bias for small sample sizes in the simulation study. He suggested that we consider alternative variance estimators proposed in Deville and Tillé (2005). These estimators, that we denote as DT, were developed for the variance of the Horvitz-Thompson estimator $\hat{\hat{Y}}^b_{HT}$ in the context of balanced sampling defined by Equation (3).

Using Deville and Tillé (2005)'s notation, the DT variance estimators are given by

$$\hat{V}^{DTi}_r(\hat{\theta}) = \frac{1}{N^2} \sum_{k \in s^r} c_{ki} (\breve{y}_k - \hat{\breve{y}}^*_k)^2, \quad i = 1, \ldots, 5 \tag{27}$$

where $\hat{\breve{y}}^*_k = \hat{\breve{\boldsymbol{x}}}^T_k (\sum_{l \in s^r} c_{li} \breve{\boldsymbol{x}}_l \breve{\boldsymbol{x}}^T_l)^{-1} \sum_{l \in s^r} c_{li} \breve{\boldsymbol{x}}_l \breve{y}_l$, with $\breve{\boldsymbol{x}}_k = \boldsymbol{x}_k / \pi_k$ and $\breve{y}_k = y_k / \pi_k$.

Deville and Tillé (2005) proposed five c_k's labeled as $c_{ki}, i = 1, \ldots, 5$. The c_k's are:

i. $c_{k1} = 1 - \pi_k$
ii. $c_{k2} = (1 - \pi_k) \frac{n}{n-p}$ where p is the length of vector $\boldsymbol{x}_k$
iii. $c_{k3} = (1 - \pi_k) \frac{\sum_{k \in s^r} (1 - \pi_k)}{\sum_{k \in s^r} D_{kk1}}$ where $D_{kk1} = c_{k1} - c^2_{k1} \breve{\boldsymbol{x}}^T_k \left(\sum_{l \in s^r} c_{l1} \breve{\boldsymbol{x}}_l \breve{\boldsymbol{x}}^T_l \right)^{-1} \breve{\boldsymbol{x}}_k$
iv. $c_{k4} = \frac{b_{k4}}{\pi_k} \frac{n}{n-p} \frac{N-p}{N}$ where b_{k4} are solutions to the nonlinear equations

$$\pi_k (1 - \pi_k) = b_{k4} - b^2_{k4} \breve{\boldsymbol{x}}^T_k \left(\sum_{l \in U} b_{l4} \breve{\boldsymbol{x}}_l \breve{\boldsymbol{x}}^T_l \right)^{-1} \breve{\boldsymbol{x}}_k, k = 1, \ldots, N$$

and

v. c_{k5} where the c_{k5}'s are solutions to the nonlinear equations

$$1 - \pi_k = c_{k5} - c_{k5}^2 \breve{\boldsymbol{x}}_k^T \left(\sum_{l \in s^r} c_{l5} \breve{\boldsymbol{x}}_l \breve{\boldsymbol{x}}_l^T \right)^{-1} \breve{\boldsymbol{x}}_k, k = 1, \ldots, n$$

We followed Matei and Tillé (2005) to solve the nonlinear equations by the fixed point technique using a single iteration.

We denote our estimator given by (25) as $\hat{V}_r^{SH}(\hat{\theta})$. We analyzed the performance of the DT variance estimators, and compared them to our own, for $\hat{\theta} = \hat{\hat{Y}}_{HT}^b$, via a third simulation study using SRSWOR as the basic sampling design. Since under SRSWOR, $\hat{V}_r^{DT2}(\hat{\theta}) = \hat{V}_r^{DT3}(\hat{\theta})$, we computed $\hat{V}_r^{DTi}(\hat{\theta})$ for $i = 1, 2, 4, 5$. We used $\boldsymbol{x}_k = (x_{1k}, x_{2k})$, implying that $p = 2$.

We generated a large number ($L = 500{,}000$) of rejective samples s^r to compute the Monte Carlo rejective variance as:

$$V_r^{MC3}(\hat{\theta}) = \frac{1}{L} \sum_{l=1}^{L} (\hat{\theta}^{(l)} - \hat{\hat{\theta}})^2$$

with $\hat{\hat{\theta}} = L^{-1} \sum_{l=1}^{L} \hat{\theta}^{(l)}$ where $\hat{\theta}^{(l)}$ is the ℓ^{th} value of $\hat{\theta}$.

Let $\hat{V}_r$ be one of the five variance estimators ($\hat{V}_r^{SH}(\hat{\theta})$ and $\hat{V}_r^{DTi}(\hat{\theta})$, $i = 1, 2, 4, 5$) to be compared in the simulation study, and $\hat{V}_r^{(l)}$ its ℓ^{th} value. The variance estimators are compared via the Relative Bias (RB):

$$RB(\hat{V}_r) = \frac{E_r^{MC3}(\hat{V}_r)}{V_r^{MC3}(\hat{\theta})} - 1$$

where the Monte Carlo mean of $\hat{V}_r$ is computed as $E_r^{MC3}(\hat{V}_r) = L^{-1} \sum_{l=1}^{L} \hat{V}_r^{(l)}$.

We considered two more values for the rejection rate: 0.99 and 0.997. When the basic sampling design is SRSWOR, a rejection rate of 99% is obtained for $\gamma_3^2 = 0.003$ whereas a rejection rate of 99.7% is obtained for $\gamma_4^2 = 0.0009$. The results of the relative bias of the five variance estimators under the three rejective rates, 0.90, 0.99 and 0.997, are given in Table 3.

Table 3. Relative bias (%) of the variance estimators for $\hat{\theta} = \hat{\hat{Y}}_{HT}^b$: basic sampling design SRSWOR.

		$\hat{V}_r^{SH}(\hat{\theta})$	$\hat{V}_r^{DT1}(\hat{\theta})$	$\hat{V}_r^{DT2}(\hat{\theta})$	$\hat{V}_r^{DT4}(\hat{\theta})$	$\hat{V}_r^{DT5}(\hat{\theta})$
RR = 0.90 ($\gamma_1^2 = 0.05$)	***n* = 10**	−16.1	−79.9	−74.9	−75.0	−76.0
	***n* = 20**	−7.9	−77.5	−75.0	−75.0	−75.4
	***n* = 50**	−2.9	−75.9	−74.9	−74.9	−75.0
	***n* = 100**	−1.9	−75.4	−74.9	−74.9	−74.9
RR = 0.99 ($\gamma_3^2 = 0.003$)	***n* = 10**	−20.6	−31.5	−14.3	−14.4	−17.9
	***n* = 20**	−10.6	−23.6	−15.1	−15.1	−16.4
	***n* = 50**	−3.5	−17.7	−14.3	−14.3	−14.6
	***n* = 100**	−1.6	−16.1	−14.4	−14.4	−14.5
RR = 0.997 ($\gamma_4^2 = 0.0009$)	***n* = 10**	−22.0	−23.7	−4.6	−4.7	−8.6
	***n* = 20**	−11.3	−14.7	−5.2	−5.2	−6.6
	***n* = 50**	−5.3	−9.3	−5.5	−5.6	−5.9
	***n* = 100**	−2.7	−6.9	−5.0	−5.0	−5.1

Note: $\bar{Y}_U = 21.147$, $S_{yU}^2 = 234.72$ and $R^2 = 99.5\%$.

The relative bias of the four DT variance estimators differ across rejection rates and sample sizes. The relative bias of these estimators decreases as the rejection rate increases. When the rejection rate is 90%, none of the DT variance estimators have a relative bias that is smaller than the SH variance estimator. All four DT variance estimators display very similar relative bias, ranging from -79.9% to -74.9% for all sample sizes considered.

The relative bias of the DT estimators decreases when the rejection rate is increased to 99%. DT2, DT4 and DT5 have relative biases that are similar. Their bias is smaller than the one associated with SH only for $n = 10$. DT1 has relative bias that is consistently higher than the one associated with SH.

The DT estimators have the smallest relative bias when the rejection rate is 99.7%. This is not surprising as the balancing Equation (3) is closely satisfied at such a high rejection rate. Once more, DT1 has the largest relative bias amongst the DT variance estimators. The other DT variance estimators have smaller relative bias than the SH estimator when n is 10 or 20. For n equal to 50 or 100, the SH estimator has the smallest relative bias.

4.2. *Rejective Sampling Using BernWOR*

In Table 4, we present the results obtained when the basic procedure is BernWOR and the rejection rate is 90%. The results in the table are with respect to the expected sample size.

The Monte Carlo expectation of $\hat{\theta} = \hat{\hat{Y}}^b_{HT}$ and $\hat{\theta} = \hat{\hat{Y}}^b_{GREG}$ represented by $E^{MC2}_{2r}(\hat{\theta})$ is quite close to the true population mean $\bar{Y}_U = 21.147$, for all values of n.

BernWOR adds extra variation because the sample size is random. For the HT estimator, the basic variance $V^{MC1}_{2b}(\hat{\hat{Y}}^b_{HT})$ under BernWOR is approximately three times larger than $V^{MC1}_{1b}(\hat{\hat{Y}}^b_{HT})$ under SRSWOR. For the GREG estimator, $V^{MC1}_{2b}(\hat{\hat{Y}}^b_{GREG})$ and $V^{MC1}_{1b}(\hat{\hat{Y}}^b_{GREG})$ are much closer except for $n = 10$. In this case, $V^{MC1}_{2b}(\hat{\hat{Y}}^b_{GREG})$ is 0.3098 as opposed to 0.1794 for $V^{MC1}_{1b}(\hat{\hat{Y}}^b_{GREG})$. Due to the small expected sample size ($n = 10$), the value $V^{MC1}_{2b}(\hat{\hat{Y}}^b_{GREG}) = 0.3098$ was computed using BernWOR samples s^b of size larger than 5. There is still a large difference between 0.3098 and 0.1794 compared to the corresponding differences associated to larger expected sample sizes considered in the simulation. This shows that it is fairly inappropriate to use a GREG estimator in samples with very few observations when the GREG estimator is based on a three-dimensional vector $(1, x_{1\mathrm{k}}, x_{2k})^T$.

The gains in terms of variance due to the rejective sampling design are larger for SRSWOR as opposed to BernWOR for $\hat{\theta} = \hat{\hat{Y}}^b_{HT}$. This is observed by comparing the ratios $V^{MC1}_{1b}(\hat{\hat{Y}}^b_{HT})/V^{MC2}_{1r}(\hat{\hat{Y}}^b_{HT})$ in Table 2 for SRSWOR to the ratios $V^{MC1}_{2b}(\hat{\hat{Y}}^b_{HT})/V^{MC2}_{2r}(\hat{\hat{Y}}^b_{HT})$ in Table 4 for BernWOR. The ratio $V^{MC1}_{1b}(\hat{\hat{Y}}^b_{HT})/V^{MC2}_{1r}(\hat{\hat{Y}}^b_{HT})$ is approximately equal to 60, whereas the ratio $V^{MC1}_{2b}(\hat{\hat{Y}}^b_{HT})/V^{MC2}_{2r}(\hat{\hat{Y}}^b_{HT})$ is approximately equal to 20 for any sample size in the simulation. For $\hat{\theta} = \hat{\hat{Y}}^b_{GREG}$, the gains in terms of variance are similar for SRSWOR and BernWOR.

Note that $V_{2r}(\hat{\theta})$ and $V^{MC2}_{2r}(\hat{\theta})$, for $\hat{\theta} = \hat{\hat{Y}}^b_{HT}$ and $\hat{\theta} = \hat{\hat{Y}}^b_{GREG}$, are getting closer as n increases.

For $\hat{\hat{Y}}^b_{GREG}$, its rejective and basic variances are getting closer as n increases. This observation is in line with what happens when SRSWOR is the initial sampling design. A similar argument can be used to support this observation for moderate to large values of n. The population covariance $\mathrm{cov}_{2b}(\hat{\hat{X}}^b_{HT}, \hat{\hat{Y}}^b_{GREG})$ is approximately equal to

Table 4. Basic sampling design BernWOR and RR = 0.90 ($\gamma_2^2 = 0.22$).

	Estimator							
	$\hat{\theta} = \hat{\bar{Y}}_{HT}^b$ Expected sample sizes				$\hat{\theta} = \hat{\bar{Y}}_{GREG}^b$ Expected sample sizes			
Summary statistics	***n* = 10**	***n* = 20**	***n* = 50**	***n* = 100**	***n* = 10**	***n* = 20**	***n* = 50**	***n* = 100**
$E_{2r}^{MC2}(\hat{\theta})$	21.936	21.037	21.109	21.127	21.144	21.145	21.146	21.147
$V_{2b}^{MC1}(\hat{\theta})$	67.490	33.404	12.946	6.136	0.3098	0.0628	0.0196	0.0089
$V_{2r}^{MC2}(\hat{\theta})$	3.390	1.4321	0.5896	0.2796	0.0966	0.0478	0.0184	0.0087
$V_{2r}(\hat{\theta})$ (eq.(21))	3.023	1.4964	0.5799	0.2748	0.3097	0.0627	0.0196	0.0089
$E_{2r}^{MC2}[\hat{V}_{2r}(\hat{\theta})]$	2.745	1.4237	0.5700	0.2725	0.0665	0.0401	0.0172	0.0084
$E_{2r}^{MC2}[\hat{V}_{2b}(\hat{\theta})]$	61.744	31.894	12.743	6.0875	0.0667	0.0402	0.0172	0.0084

Note: $\bar{Y}_U = 21.147$, $S_{yU}^2 = 234.72$ and $R^2 = 99.5\%$.

$(1-f)n^{-1}N^{-1}\sum_{i\in U} E_i \boldsymbol{x}_i$. This approximation is obtained by applying Result 6.6.1., in Särndal et al. (1992, 235) to the case of the covariance between the GREG and the HT estimators when the sampling design is BernWOR. Since $\sum_{i\in U} E_i \boldsymbol{x}_i = 0$, it follows that the second term in Equation (21) approaches zero as n increases.

The population variance $V_{2r}(\hat{\theta})$ is estimated by $\hat{V}_{2r}(\hat{\theta})$. Comparing the Monte Carlo mean of the resulting estimator $E_{2r}^{MC2}[\hat{V}_{2r}(\hat{\theta})]$ to the Monte Carlo variance $V_{2r}^{MC2}(\hat{\theta})$, we see that large differences occur for $n = 10$ for both $\hat{\bar{Y}}_{HT}^{b}$ and $\hat{\bar{Y}}_{GREG}^{b}$. As n increases the bias of these estimators tends to zero.

5. Conclusion

The use of rejective sampling brings about reductions in variance for both the Horvitz-Thompson estimator and the GREG regression estimator. The reduction is more significant for the Horvitz-Thompson than for the regression estimator.

We obtained an exact formula for the rejective variance $V_r(\hat{\theta})$ of a population mean estimator $\hat{\theta}$ by assuming that under the basic sampling design, $\hat{\theta}$ and its associated auxiliary data mean $\hat{\bar{\boldsymbol{X}}}_{HT}^{b}$, have a joint normal distribution. This result was obtained by conditioning on $Q \leq \gamma^2$. This allowed us to avoid computing (approximating) the unknown rejective inclusion probabilities π_i^r and π_{ij}^r.

If the normal distribution is only approximate, $V_r(\hat{\theta})$ represents an approximation of the true rejective variance. There are estimators $\hat{\theta}$ of the population mean $\bar{Y}_U$ whose distribution tends to normality. Our result is applicable to such estimators in samples with sufficiently large n.

An estimator $\hat{V}_r(\hat{\theta})$ for the rejective variance was obtained by replacing the unknown parameters in $V_r(\hat{\theta})$ by estimators associated with the basic sampling design and based on the rejective sample.

A simulation study was undertaken to evaluate the accuracy of $V_r(\hat{\theta})$ and the bias of its estimator $\hat{V}_r(\hat{\theta})$. This was carried out for two estimators ($\hat{\theta} = \hat{\bar{Y}}_{HT}^{b}$ and $\hat{\theta} = \hat{\bar{Y}}_{GREG}^{b}$) and two basic sampling designs (SRSWOR and BernWOR). The empirical results show that if the normality assumption is approximately respected, the formula for $V_r(\hat{\theta})$ performs well for moderate to large values of n. For small sample sizes, the proposed variance estimator $\hat{V}_r(\hat{\theta})$ is biased but the bias reduces as n increases.

6. Appendix

6.1. *Appendix A: Proof of Theorem 1*

i. Denote by $f_{\boldsymbol{W}}(\boldsymbol{w})$ the density function of vector $\boldsymbol{W}$ where $\boldsymbol{w} = (x, z)$ and $z = (z_1, \ldots, z_p)$.

The marginal density of $\boldsymbol{Z}$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_W(x, z)dx = \frac{1}{(\sqrt{2\pi})^p} e^{-\frac{z^T z}{2}}$$

Using the formula for the inverse of matrix $\boldsymbol{\Sigma}_{\boldsymbol{W}}$, the determinant $|\boldsymbol{\Sigma}_{\boldsymbol{w}}| = a$ and the definition of matrix $\boldsymbol{M}$, $\boldsymbol{M} = \boldsymbol{I}_p + a^{-1}\sigma_{z\theta}\sigma_{z\theta}^T$, the density function $f_{\boldsymbol{W}}(\boldsymbol{w})$ of vector $\boldsymbol{W}$ is:

$$
\begin{aligned}
f_{\boldsymbol{W}}(\boldsymbol{w}) &= \frac{1}{(\sqrt{2\pi})^{p+1}\sqrt{|\boldsymbol{\Sigma}_w|}}\, e^{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_w)^T\boldsymbol{\Sigma}_w^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_w)} \\
&= \frac{1}{\sqrt{2\pi a}}\, e^{-\frac{1}{2a}\left[(x-\mu_\theta)-z^T\sigma_{z\theta}\right]^2} f_{\boldsymbol{Z}}(z)
\end{aligned}
\tag{A1}
$$

The conditional mean $E_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \le \gamma^2)$ can alternatively be rewritten as $E_b(\hat{\theta}|\boldsymbol{Z} \in \boldsymbol{A})$, where $\boldsymbol{A}$ is the random event defined as $\boldsymbol{A} = \{\omega|Z_1(\omega)^2 + \ldots + Z_p(\omega)^2 \le \gamma^2\}$. In order to compute $E_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \le \gamma^2)$ and $E_b(\hat{\theta}^2|\boldsymbol{Z}^T\boldsymbol{Z} \le \gamma^2)$ one needs to find the density function $f_{\hat{\theta}|\boldsymbol{Z}\in\boldsymbol{A}}(x)$ of the conditional variable $\hat{\theta}|\boldsymbol{Z} \in \boldsymbol{A}$.

The cumulative probability function of $\hat{\theta}|\boldsymbol{Z} \in \boldsymbol{A}$ is given by:

$$
F_{\hat{\theta}|\boldsymbol{Z}\in\boldsymbol{A}}(x) = P(\hat{\theta} \le x|\boldsymbol{Z} \in \boldsymbol{A}) = \frac{P(\hat{\theta} \le x \cap \boldsymbol{Z} \in \boldsymbol{A})}{P(\boldsymbol{Z} \in \boldsymbol{A})} = \frac{\int_{-\infty}^{x}\int_A f_{\boldsymbol{W}}(t,z)\,dz\,dt}{\int_A f_{\boldsymbol{Z}}(z)\,dz}
$$

where A is the set defined in Theorem 1.

The conditional density function $f_{\hat{\theta}|\boldsymbol{Z}\in\boldsymbol{A}}(x)$ is obtained by differentiating $F_{\hat{\theta}|\boldsymbol{Z}\in\boldsymbol{A}}(x)$ with respect to x:

$$
f_{\hat{\theta}|\boldsymbol{Z}\in\boldsymbol{A}}(x) = \frac{dF_{\hat{\theta}|\boldsymbol{Z}\in\boldsymbol{A}}(x)}{dx} = \frac{\int_A f_{\boldsymbol{W}}(x,z)\,dz}{\int_A f_{\boldsymbol{Z}}(z)\,dz}
$$

As a consequence, the conditional mean $E_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \le \gamma^2)$ can be computed as:

$$
E_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \le \gamma^2) = \frac{\int_{-\infty}^{\infty} x\left(\int_A f_{\boldsymbol{W}}(x,z)\,dz\right)dx}{\int_A f_{\boldsymbol{Z}}(z)\,dz} = \frac{\int_A\left(\int_{-\infty}^{\infty} x f_{\boldsymbol{W}}(x,z)\,dx\right)dz}{\int_A f_{\boldsymbol{Z}}(z)\,dz}
\tag{A2}
$$

The integral $\int_{-\infty}^{\infty} x f_{\boldsymbol{W}}(x,z)\,dx$ in (A2) can be computed as follows:

$$
\begin{aligned}
\int_{-\infty}^{\infty} x f_{\boldsymbol{W}}(x,z)\,dx &= \int_{-\infty}^{\infty} (x-\mu_\theta) f_{\boldsymbol{W}}(x,z)\,dx + \mu_\theta \int_{-\infty}^{\infty} f_{\boldsymbol{W}}(x,z)\,dx \\
&= \int_{-\infty}^{\infty} (x-\mu_\theta) f_{\boldsymbol{W}}(x,z)\,dx + \mu_\theta f_{\boldsymbol{Z}}(z)
\end{aligned}
$$

The first term is obtained by replacing $f_{\boldsymbol{w}}\,(x,z)$ given in (A1):

$$
\begin{aligned}
\int_{-\infty}^{\infty} (x-\mu_\theta) f_{\boldsymbol{W}}(x,z)\,dx &= \frac{1}{\sqrt{2\pi a}} \int_{-\infty}^{\infty} (x-\mu_\theta)\, e^{-\frac{1}{2a}\left[(x-\mu_\theta)-z^T\boldsymbol{\sigma}_{z\theta}\right]^2} dx f_{\boldsymbol{Z}}(z) \\
&= (z^T\boldsymbol{\sigma}_{z\theta}) f_{\boldsymbol{Z}}\,(z)
\end{aligned}
$$

In the last equation we used the value of integral:

$$\frac{1}{\sqrt{2\pi a}}\int_{-\infty}^{\infty} y e^{-\frac{1}{2a}\left(y-z^T\sigma_{z\theta}\right)^2} dy = \boldsymbol{z}^T\boldsymbol{\sigma}_{z\theta}$$

Hence,

$$\int_{-\infty}^{\infty} x f_{\boldsymbol{W}}(x,z)dx = (z^T\boldsymbol{\sigma}_{z\theta} + \mu_\theta) f_{\boldsymbol{Z}}(z) \tag{A3}$$

Replacing (A3) in (A2), we obtain:

$$E_b(\hat{\theta}|\boldsymbol{Z}^T\boldsymbol{Z} \leq \gamma^2) = \frac{\int_A (z^T\boldsymbol{\sigma}_{z\theta} + \mu_\theta) f_{\boldsymbol{Z}}(z)dz}{\int_A f_{\boldsymbol{Z}}(z)dz} = \frac{\int_A (z^T\boldsymbol{\sigma}_{z\theta}) f_{\boldsymbol{Z}}(z)dz}{\int_A f_{\boldsymbol{Z}}(z)dz} + \mu_\theta = \mu_\theta \tag{A4}$$

In (A4) we use that $\int_A z_i f_{\boldsymbol{Z}}(z)dz = 0$ which implies $\int_A (z^T\boldsymbol{\sigma}_{z\theta}) f_{\boldsymbol{Z}}(z)dz = 0$ (see (B7) below).

ii A similar argument holds for $E_b(\hat{\theta}^2|\boldsymbol{Z}^T\boldsymbol{Z} \leq \gamma^2)$:

$$E_b(\hat{\theta}^2|\boldsymbol{Z}^T\boldsymbol{Z} \leq \gamma^2) = \frac{\int_{-\infty}^{\infty} x^2 \left(\int_A f_{\boldsymbol{W}}(x,z)dz\right)dx}{\int_A f_{\boldsymbol{Z}}(z)dz} = \frac{\int_A \left(\int_{-\infty}^{\infty} x^2 f_{\boldsymbol{W}}(x,z)dx\right)dz}{\int_A f_{\boldsymbol{Z}}(z)dz} \tag{A5}$$

The numerator of (A5) can be computed as

$$\int_{-\infty}^{\infty} x^2 f_{\boldsymbol{W}}(x,z)dx = \int_{-\infty}^{\infty} (x-\mu_\theta)^2 f_{\boldsymbol{W}}(x,z)dx + 2\mu_\theta \int_{-\infty}^{\infty} (x-\mu_\theta) f_{\boldsymbol{W}}(x,z)dx$$

$$+ \mu_\theta^2 \int_{-\infty}^{\infty} f_{\boldsymbol{W}}(x,z)dx \tag{A6}$$

We now compute each term of (A6). For the first term we replace the density function $f_{\boldsymbol{W}}(x,z)$ by Equation (A1):

$$\int_{-\infty}^{\infty} (x-\mu_\theta)^2 f_{\boldsymbol{W}}(x,z)dx = \frac{1}{\sqrt{2\pi a}}\int_{-\infty}^{\infty} (x-\mu_\theta)^2 e^{-\frac{1}{2a}\left[(x-\mu_\theta)-z^T\sigma_{z\theta}\right]^2} dx f_{\boldsymbol{Z}}(z)$$

$$= [a + (z^T\boldsymbol{\sigma}_{z\theta})^2] f_{\boldsymbol{Z}}(z) \tag{A7}$$

In the above equation we used the value of integral

$$\frac{1}{\sqrt{2\pi a}}\int_{-\infty}^{\infty} y^2 e^{-\frac{1}{2a}\left[y-z^T\sigma_{z\theta}\right)^2} dy = a + (z^T\boldsymbol{\sigma}_{z\theta})^2$$

For the second term of (A6) we use (A3):

$$2\mu_\theta \int_{-\infty}^{\infty} (x-\mu_\theta) f_{\mathbf{W}}(x,z)dx = 2\mu_\theta(z^T\sigma_{z\theta}) f_{\mathbf{Z}}(z) \tag{A8}$$

The third term in (A6) is

$$\mu_\theta^2 \int_{-\infty}^{\infty} f_{\mathbf{W}}(x,z)dx = \mu_\theta^2 f_{\mathbf{Z}}(z) \tag{A9}$$

Replacing (A7), (A8) and (A9) into (A6) one obtains:

$$\int_{-\infty}^{\infty} x^2 f_{\mathbf{W}}(x,z)dx = [a + (z^T\boldsymbol{\sigma}_{z\theta})^2 + 2\mu_\theta(z^T\boldsymbol{\sigma}_{z\theta}) + \mu_\theta^2] f_{\mathbf{Z}}(z) \tag{A10}$$

Replacing (A10) into (A5) and using again $\int_A (z^T\boldsymbol{\sigma}_{z\theta}) f_{\mathbf{Z}}(z)dz = 0$, we obtain:

$$E_b(\hat{\theta}^2 | \mathbf{Z}^T\mathbf{Z} \leq \gamma^2) = a + \mu_\theta^2 + \frac{\int_A (z^T\boldsymbol{\sigma}_{z\theta})^2 f_{\mathbf{Z}}(z)dz}{\int_A f_{\mathbf{Z}}(z)dz}. \tag{A11}$$

The conditional variance follows from (A4) and (A11):

$$V_b(\hat{\theta} | \mathbf{Z}^T\mathbf{Z} \leq \gamma^2) = a + \frac{\int_A (z^T\boldsymbol{\sigma}_{z\theta})^2 f_{\mathbf{Z}}(z)dz}{\int_A f_{\mathbf{Z}}(z)dz} \tag{A12}$$

Now, by symmetry, $\int_A z_i^2 f_{\mathbf{Z}}(z)dz = \int_A z_1^2 f_{\mathbf{Z}}(z)dz$ and $\int_A z_i z_j f_{\mathbf{Z}}(z)dz = 0$ for $i \neq j$ (see (B8) below). It follows that:

$$\int_A (z^T\boldsymbol{\sigma}_{z\theta})^2 f_{\mathbf{Z}}(z)dz = (\boldsymbol{\sigma}_{z\theta}^T\boldsymbol{\sigma}_{z\theta}) \int_A z_1^2 f_{\mathbf{Z}}(z)dz \tag{A13}$$

Using (A12), (A13) and the definition of *a*, $a = \sigma_\theta^2 - \boldsymbol{\sigma}_{z\theta}^T\boldsymbol{\sigma}_{z\theta}$, the conditional variance is obtained:

$$V_b(\hat{\theta} | \mathbf{Z}^T\mathbf{Z} \leq \gamma^2) = \sigma_\theta^2 - \boldsymbol{\sigma}_{z\theta}^T\boldsymbol{\sigma}_{z\theta}\left(1 - \frac{\int_A z_1^2 f_{\mathbf{Z}}(z)dz}{\int_A f_{\mathbf{Z}}(z)dz}\right)$$

and this proves ii.

6.2. Appendix B: Proof of Proposition 1

i. We have that:

$$J_{n+1}(\gamma) = \int_0^\gamma r^n \left(-e^{-\frac{r^2}{2}}\right)' dr = r^n e^{-\frac{r^2}{2}}\Big|_\gamma^0 + n\int_0^\gamma r^{n-1} e^{-\frac{r^2}{2}} dr = nJ_{n-1}(\gamma) - \gamma^n e^{-\frac{\gamma^2}{2}}$$

and relation (19) is proved.

The first two "J" terms are:

$$J_0(\gamma) = \frac{\sqrt{2\pi}}{\sqrt{2\pi}} \int_0^{\gamma} e^{-\frac{r^2}{2}} dr = \sqrt{2\pi}(\Phi(\gamma) - \Phi(0)) = \sqrt{2\pi}(\Phi(\gamma) - 0.5),$$

$$J_1(\gamma) = \int_0^{\gamma} \left(-e^{-\frac{r^2}{2}}\right)' dr = e^{-\frac{r^2}{2}}\Bigg|_{\gamma}^{0} = 1 - e^{-\frac{\gamma^2}{2}}$$

ii. We use the transformation to p-spherical coordinates. For a point $z \in A$ with $z = (z_1, \ldots, z_p)$, its p-spherical coordinates are $(r, \theta_1, \ldots, \theta_{p-1})$, where:

$$\begin{aligned} z_1 &= r\cos\theta_1 \\ z_2 &= r\sin\theta_1\cos\theta_2 \\ z_3 &= r\sin\theta_1\sin\theta_2\cos\theta_3 \\ &\vdots \\ z_{p-1} &= r\sin\theta_1\sin\theta_2 \ldots \sin\theta_{p-2}\cos\theta_{p-1} \\ z_p &= r\sin\theta_1\sin\theta_2 \ldots \sin\theta_{p-2}\sin\theta_{p-1} \end{aligned} \tag{B1}$$

with $r \in [0, \gamma]$, $\theta_1 \in [0, \pi], \ldots, \theta_{p-2} \in [0, \pi]$ and $\theta_{p-1} \in [0, 2\pi]$. Using (B1), we have that $z^T z = r^2$.

Let us denote by B the set defined as the Cartesian product

$$B = [0, \gamma] \times [0, \pi] \times \ldots \times [0, \pi] \times [0, 2\pi]$$

The transformation (B1) maps the set A into the set B. The Jacobian of the transformation (B1) is given by:

$$dz_1 dz_2 \ldots dz_p = r^{p-1}\sin^{p-2}\theta_1 \sin^{p-3}\theta_2 \ldots \sin\theta_{p-2}\; dr\, d\theta_1 \ldots d\theta_{p-2} d\theta_{p-1} \tag{B2}$$

For a positive integer $n \geq 0$, let us denote by T_n the integral

$$T_n = \int_0^{\pi} \sin^n \theta\, d\theta$$

It can be shown that T_n obeys the following recursive relationship:

$$nT_n = (n-1)T_{n-2} \tag{B3}$$

Using (B1) and (B2), the integrals $\int_A f_Z(z)dz$ and $\int_A z_1^2 f_Z(z)dz$ are given by:

$$\int_A f_Z(z)dz = \frac{1}{(\sqrt{2\pi})^p}\int_B e^{-\frac{1}{2}r^2} r^{p-1}\sin^{p-2}\theta_1 \sin^{p-3}\theta_2 \ldots \sin\theta_{p-2}drd\theta_1\ldots d\theta_{p-2}d\theta_{p-1} \Rightarrow$$

$$\int_A f_Z(z)dz = \frac{2\pi}{(\sqrt{2\pi})^p}J_{p-1}(\gamma)T_{p-2}T_{p-3}\ldots T_1 \tag{B4}$$

and

$$\int_A z_1^2 f_Z(z)dz = \frac{1}{(\sqrt{2\pi})^p}\int_B e^{-\frac{1}{2}r^2} r^{p+1}\sin^{p-2}\theta_1\cos^2\theta_1\sin^{p-3}\theta_2\ldots\sin\theta_{p-2}drd\theta_1\ldots d\theta_{p-1} \Rightarrow$$

$$\int_A z_1^2 f_Z(z)dz = \frac{2\pi}{(\sqrt{2\pi})^p}J_{p+1}(\gamma)(T_{p-2}-T_p)T_{p-3}\ldots T_1 \tag{B5}$$

Using (B3), one gets that $(T_{p-2} - T_p) = T_{p-2}/p$ and replacing in (B5) it follows that

$$\int_A z_1^2 f_Z(z)dz = \frac{2\pi}{p(\sqrt{2\pi})^p}J_{p+1}(\gamma)T_{p-2}T_{p-3}\ldots T_1 \tag{B6}$$

From (B4) and (B6) it results that

$$1 - \frac{\int_A z_1^2 f_Z(z)dz}{\int_A f_Z(z)dz} = 1 - \frac{J_{p+1}(\gamma)}{pJ_{p-1}(\gamma)}$$

Then, the recursive formula (19) obtained for $J_n(\gamma)$ is used and Equation (20) follows.

In Appendix A, we stated without proof that $\int_A z_1 f_Z(z)dz$ and $\int_A z_1 z_2 f_Z(z)dz$ were null. We now proceed to prove this. Since $\int_0^\pi \sin^{p-2}\theta_1 \cos\theta_1 \, d\theta_1 = 0$, it follows that:

$$\int_A z_1 f_Z(z)dz = \frac{1}{(\sqrt{2\pi})^p}\int_B e^{-\frac{1}{2}r^2} r^p \sin^{p-2}\theta_1\cos\theta_1\sin^{p-3}\theta_2\ldots\sin\theta_{p-2}drd\theta_1\ldots d\theta_{p-1} \Rightarrow$$

$$\int_A z_1 f_Z(z)dz = \frac{2\pi}{(\sqrt{2\pi})^p}J_p(\gamma)T_{p-3}\ldots T_1\int_0^\pi \sin^{p-2}\theta_1\cos\theta_1 d\theta_1 = 0 \tag{B7}$$

Also,

$$\int_A z_1 z_2 f_Z(z)dz = \frac{1}{(\sqrt{2\pi})^p}\int_B e^{-\frac{1}{2}r^2} r^{p+1}\sin^{p-1}\theta_1\cos\theta_1\sin^{p-3}\theta_2\cos\theta_2\ldots\sin\theta_{p-2}drd\theta_1\ldots d\theta_{p-1} \Rightarrow$$

$$\int_A z_1 z_2 f_Z(z)dz = \frac{2\pi}{(\sqrt{2\pi})^p}J_{p+1}(\gamma)T_{p-4}\ldots T_1\int_0^\pi \sin^{p-1}\theta_1\cos\theta_1 d\theta_1\int_0^\pi \sin^{p-3}\theta_2\cos\theta_2 d\theta_2 = 0 \tag{B8}$$

By symmetry, $\int_A z_i f_Z(z)dz = \int_A z_1 f_Z(z)dz = 0$ and $\int_A z_i z_j f_Z(z)dz = \int_A z_1 z_2 f_Z(z)dz = 0, i \neq j$.

7. References

Chauvet, G., D. Haziza, and E. Lesage. 2017. "Examining Some Aspects of Balanced Sampling in Surveys." *Statistica Sinica* 27: 313–334. DOI: http://dx.doi.org/10.5705/ss.2013.244.

Deville, J.-C. and C.E. Särndal. 1992. "Calibration estimators in survey sampling." *Journal of the American Statistical Association* 87: 376–382. DOI: http://doi.org/10.2307/2290268.

Deville, J.-C. and Y. Tillé. 2004. "Efficient balanced sampling: The cube method." *Biometrika* 91: 893–912. DOI: http://doi.org/10.1093/biomet/91.4.893.

Deville, J.-C. and Y. Tillé. 2005. "Variance approximation under balanced sampling." *Journal of statistical planning and inference* 128: 569–591. DOI: https://doi.org/10.1016/j.jspi.2003.11.011.

Fuller, W.A. 2009. "Some design properties of a rejective sampling procedure." *Biometrika* 96: 933–944. DOI: https://doi.org/10.1093/biomet/asp042.

Fuller, W.A., J.C. Legg, and Y. Li. 2017. "Bootstrap variance estimation for rejective sampling." *Journal of the American Statistical Association* 112: 1562–1570. DOI: https://doi.org/10.1080/01621459.2016.1222285.

Hájek, J. 1964. "Asymptotic theory of rejective sampling with varying probabilities from a finite population." *Ann. Math. Statist.* 35: 1491–1523. DOI: https://doi.org/10.1214/aoms/1177700375.

Hájek, J. 1981. *Sampling from a finite population*. Statistics: Textbooks and Monographs 37. New York: Marcel Dekker Inc.

Horvitz, D.G. and D.J. Thompson. 1952. "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association* 47: 663–685. DOI: https://doi.org/10.2307/2280784.

Huang, E.T. and W.A. Fuller. 1978. "Nonnegative regression estimation for sample survey data." *Proceedings of the Social Statistics Section, American Statistical Association*. Alexandria, VA, 300–305. Available at: https://lib.dr.iastate.edu/rtd/6460 (accessed February 2020).

Legg, J.C. and C.L. Yu. 2010. "A comparison of sample set restriction procedures." *Survey Methodology* 36: 69–79. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X201000111249 (accessed February 2020).

Matei, A. and Y. Tillé. 2005. "Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size." *Journal of Official Statistics* 21(4): 543–570. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/evaluation-of-variance-approximations-and-estimators-in-maximum-entropy-sampling-with-unequal-probability-and-fixed-sample-size.pdf (accessed February 2020).

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

Received June 2018
Revised March 2019
Accepted June 2019

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 197–223, http://dx.doi.org/10.2478/JOS-2020-0010

Fully Bayesian Benchmarking of Small Area Estimation Models

Junni L. Zhang[1] *and John Bryant*[2]

Estimates for small areas defined by social, demographic, and geographic variables are increasingly important for official statistics. To overcome problems of small sample sizes, statisticians usually derive model-based estimates. When aggregated, however, the model-based estimates typically do not agree with aggregate estimates (benchmarks) obtained through more direct methods. This lack of agreement between estimates can be problematic for users of small area estimates. Benchmarking methods have been widely used to enforce agreement. Fully Bayesian benchmarking methods, in the sense of yielding full posterior distributions after benchmarking, can provide coherent measures of uncertainty for all quantities of interest, but research on fully Bayesian benchmarking methods is limited. We present a flexible fully Bayesian approach to benchmarking that allows for a wide range of models and benchmarks. We revise the likelihood by multiplying it by a probability distribution that measures agreement with the benchmarks. We outline Markov chain Monte Carlo methods to generate samples from benchmarked posterior distributions. We present two simulations, and an application to English and Welsh life expectancies.

Key words: Small domain estimation; Bayesian hierarchical model; area-level model; life expectancy.

1. Introduction

Small area estimation is the problem of obtaining estimates for many areas or domains defined by social, demographic and geographic variables where the number of observations in an area can be small. It has many practical applications, from monitoring unemployment to the targeting of anti-poverty programs (Pfeffermann 2013; Rao and Molina 2015), and is increasingly important for official statistics. In the United States, for instance, county-level estimates of poverty rates from the Small Area Income and Poverty Estimates (SAIPE) program are used to allocate federal funding (U.S. Census Bureau 2014). In areas where the number of observations is small, 'direct' methods, such as estimating rates by dividing the number of events in the area by the population at risk, perform poorly. Small area estimation models compensate for small sample sizes by exploiting additional information, such as covariate data or values from similar areas.

Because of their practical importance, small area estimates often receive extensive public scrutiny. This scrutiny typically includes a consistency check: estimates for small areas should agree with aggregate estimates for large areas, which are generally obtained

[1] National School of Development, Center for Statistical Science and Center for Data Science, Peking University, Beijing, 100871 China. Email: junnizhang@163.com
[2] Bayesian Demography Limited, Christchurch, New Zealand. Email: john@bayesiandemography.com

using direct methods. Model-based estimates of the number of poor people in each county, for instance, should add up to direct estimates of the number of poor people in the state. Minor discrepancies may be tolerated, but major discrepancies undermine the credibility of the estimates. Moreover, if estimates are used to allocate funding, discrepancies create grounds for dispute. Many statistical offices and funding bodies accordingly have a "one- figure" policy, whereby estimates in different tables describing the same phenomenon must all agree with each other (De Waal 2016, 232). The U.S. Census Bureau, for instance, adjusts county-level small area estimates to agree with state-level ones as part of the SAIPE program (U.S. Census Bureau 2014). Within the field of small area estimation, the aggregate estimates are referred to as benchmarks, and techniques for forcing small area estimates to agree with the benchmarks are known as benchmarking (Pfeffermann 2013).

Many existing methods for benchmarking treat benchmarks as a type of constraint. The methods differ, however, in the way that the constraints are interpreted, and in the way that the constraints are incorporated into the estimation procedures. Some methods follow a two-step procedure: first estimating the small area models, and then modifying the resulting point estimators to satisfy the benchmarking constraints (You et al. 2004; Datta et al. 2011; Berg and Fuller 2009; Berg et al. 2012; Fabrizi et al. 2012; Steorts and Ghosh 2013; Fabrizi et al. 2014; Ghosh et al. 2015). Some methods treat benchmarks as constraints on the underlying small area parameters and estimate the small area models under these constraints (Pfeffermann and Barnard 1991; Pfeffermann and Tiller 2006; Fabrizi et al. 2012; Pfeffermann et al. 2014). Some methods estimate the small area models in a way that the benchmarking constraints are satisfied for point estimators of the small area parameters (You and Rao 2002, 2003; Wang et al. 2008; You et al. 2013; Bell et al. 2013; Ranalli et al. 2018).

Most methods, including all of the ones cited above, focus on obtaining point estimates of small area parameters and associated uncertainty measures. Some Bayesian benchmarking methods, however, provide probability distributions for small area parameters (Toto and Nandram 2010; Nandram et al. 2011; Nandram and Sayit 2011; Vesper 2013). These methods are fully Bayesian in the sense that they yield a full posterior distribution for all unknown quantities after benchmarking. On this definition of fully Bayesian benchmarking, methods such as those of You et al. (2004) and Datta et al. (2011), which derive posterior distributions without benchmarking but provide point estimators after benchmarking, are not fully Bayesian. The advantage of having a full posterior distribution is that it automatically provides measures of uncertainty for all model parameters, small area parameters, and derived quantities.

In this article, we present an approach to fully Bayesian benchmarking that can be applied to a wide range of small area models. We treat benchmarks as estimates for underlying aggregate parameters. To measure agreement with the benchmarks, we specify a probability distribution for the benchmarks conditional on the aggregate parameters. We revise the likelihood function by multiplying the original likelihood function by the probability distribution for the benchmarks. Multiplying the revised likelihood function by the prior distribution then yields the benchmarked posterior distribution.

In the main body of the article, we focus on 'area-level' models, as opposed to 'unit-level' models (Rao and Molina 2015). Area-level models relate small area direct estimators to area-specific covariates. The Fay-Herriot model (Fay and Herriot 1979), for instance, is a popular area-level model used for the estimation of small area means. Unit-level models relate the unit values of an outcome variable to unit-specific covariates. The World Bank or ELL method (Elbers et al. 2003), for instance, is a widely used method for estimating small area poverty indicators, in which a unit-level model is fitted using survey data, and then applied to census data to obtain values of the outcome for all units. In the online Supplemental data (see Section 5) we discuss how our methods could be extended to unit-level models.

We implement our approach using Markov chain Monte Carlo (MCMC) methods. The methods are designed to work with complicated models that would be difficult to benchmark using previous fully Bayesian benchmarking approaches.

Our approach accommodates multiple benchmarks, and benchmarks that are nonlinearly related to small-area quantities. There is little previous research on nonlinear benchmarks: exceptions are Datta et al. (2011) and Fabrizi et al. (2012). In the application section, we estimate age-specific mortality rates benchmarked to life expectancies, which are nonlinearly related to the age-specific rates.

Our approach also allows control over the degree of agreement between model-based estimates and benchmarks. In some applications, users require exact agreement between small areas estimates and benchmarks, while in others, they may tolerate minor discrepancies. We refer to methods that achieve complete agreement as exact benchmarking, and methods that allow discrepancies as inexact benchmarking. Almost all previous methods have implemented exact benchmarking. Exceptions include Bell et al. (2013, Section 2), Nandram and Sayit (2011), and Vesper (2013).

The rest of the article is organized as follows. Section 2 describes our approach, including an outline of the associated MCMC methods. Section 3 compares our approach with previous approaches. Section 4 uses two simulation studies to illustrate the effect of benchmarking on the performance of small area models. Section 5 applies our methods to the problem of estimating district-level life expectancy in England and Wales. Section 6 summarizes the advantages of our methods.

2. A Fully Bayesian Approach to Benchmarking

2.1. Conceptual Framework

We start with a standard setup for the fully Bayesian estimation of area-level models. The aim is to estimate area-level parameters $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_n\}^\top$, such as means, rates, or probabilities, on the n areas defined by a multiway classification constructed from variables such as age, sex, and region. The data are area-level observations $\boldsymbol{y} = \{y_1, \ldots, y_n\}$. In a hierarchical Bayesian model, the likelihood is $p(\boldsymbol{y} \mid \boldsymbol{\gamma})$, the prior distribution is $p(\boldsymbol{\phi})p(\boldsymbol{\gamma} \mid \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a vector of hyperparameters, and the posterior distribution is

$$p(\boldsymbol{\gamma}, \boldsymbol{\phi} \mid \boldsymbol{y}) \propto p(\boldsymbol{\phi})p(\boldsymbol{\gamma} \mid \boldsymbol{\phi})p(\boldsymbol{y} \mid \boldsymbol{\gamma}). \tag{1}$$

The prior may itself have a complicated hierarchical structure. Throughout the article, we use Roman letters to denote data, and use Greek letters to denote parameters.

We extend this setup to incorporate benchmarking. The statistician carrying out the small area estimation is provided with a set of benchmarks $\boldsymbol{m} = \{m_1, \ldots, m_d\}^\top$, with d much less than n. The benchmarks are pre-existing summary statistics at a more aggregate level than $\boldsymbol{y}$. If $\boldsymbol{y}$ is numbers of people in the labor force disaggregated by age, sex, and education level, for example, then $\boldsymbol{m}$ might be estimates of labor force participation rates disaggregated only by sex. If $\boldsymbol{y}$ is death counts disaggregated by age, sex and region, then $\boldsymbol{m}$ might be estimates of life expectancy by sex and region. The statistician is required to make estimates of the area-level parameters $\boldsymbol{\gamma}$ agree with the benchmarks $\boldsymbol{m}$.

The benchmarks could be calculated from $\boldsymbol{y}$, or from other data sources. Within the small area estimation literature, benchmarking where $\boldsymbol{m}$ is calculated from $\boldsymbol{y}$ itself is known as internal benchmarking, and benchmarking where $\boldsymbol{m}$ is calculated from other sources is known as external benchmarking (e.g., Bell et al. 2013).

Decisions on whether to benchmark, on which statistics to benchmark to, on whether to use internal or external benchmarking, and on the degree of agreement required between small area estimates and benchmarks, are typically determined by the institutional setting and the specifics of the application. Statistical agencies often have a policy of using direct methods for aggregate measures where sample sizes are large, and using model-based methods to disaggregate further, with the requirement that model-based estimates agree with aggregate ones (Little 2012). In other words, statistical agencies require statisticians to perform internal benchmarking.

If the small area estimates will be used to allocate funding, then exact benchmarking may be required, to avoid surpluses or shortfalls. In contrast, if the main users of small area estimates are researchers and policy analysts, then some discrepancies between small area estimates and aggregates estimates may be acceptable.

We distinguish between the benchmarks and the underlying parameters that they estimate. Let $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_d\}^\top$ denote the parameters that the benchmarks $\boldsymbol{m}$ estimate. Vector $\boldsymbol{\psi}$ is derived from $\boldsymbol{\gamma}$ through a deterministic benchmarking function $\boldsymbol{\psi} = \boldsymbol{f}(\boldsymbol{\gamma})$, which consists of d components $\psi_j = f_j(\boldsymbol{\gamma})$, $j = 1, \ldots, d$. For each benchmarking parameter ψ_j, let δ_j denote the set of areas i such that γ_i contributes to ψ_j. We require that the δ_j do not overlap, in that each area i belongs to at most one δ_j. This restriction is commonly used in applications of benchmarking.

The components of the benchmarking function are typically linear, so that

$$\psi_j = \sum_{i=1}^{n} b_{ij} \gamma_i, \qquad i = 1, \ldots, n, j = 1, \ldots, d, \tag{2}$$

where the b_{ij} are known constants and $b_{ij} = 0$ for $i \notin \delta_j$. Equivalently, $\boldsymbol{\psi} = \boldsymbol{B}^\top \boldsymbol{\gamma}$ where $\boldsymbol{B}$ is a $n \times d$ matrix of b_{ij}. For example, if $\boldsymbol{\gamma}$ is labor force participation rates by age, sex and education level, and $\boldsymbol{\psi}$ is labor force participation rates by sex, then δ_j consists of all areas associated with sex j, and $b_{ij} = w_i / \sum_{i' \in \delta_j} w_{i'}$, where w_i is the population count for area i. However, the components of the benchmarking function may also be nonlinear. For example, if $\boldsymbol{\gamma}$ is mortality rates by age, sex and region, and $\boldsymbol{\psi}$ is life expectancy by sex and region, then δ_j consists of all areas associated with each combination j of sex and region, and f_j is a nonlinear deterministic function of $\{\gamma_i : i \in \delta_j\}$ (Preston et al. 2001, chap. 3). In the above formulation, we have assumed that there is only one set of benchmarks corresponding to mutually exclusive sets of small areas. In the Supplementary data

(Section 6) we discuss how our approach can be extended to allow for multiple sets of benchmarks, for instance, with one set of benchmarks estimating labor force participation rates by sex, and a second set of benchmarks estimating labor force participation rates by age.

To measure agreement with the benchmarks, we specify a probability distribution for the benchmarks conditional on the aggregate parameters, $p^{[m\,|\,\psi]}(\boldsymbol{m}\,|\,\boldsymbol{\psi}) = p^{[m\,|\,\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$. We then multiply the original likelihood $p(\boldsymbol{y}\,|\,\gamma)$ by this distribution. The modified likelihood $p(\boldsymbol{y}\,|\,\boldsymbol{\gamma})p^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$ is a compromise between the original likelihood and the requirement to agree with the benchmarks. The component $p^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$ pulls the original likelihood towards the benchmarks. For values of $\boldsymbol{\gamma}$ yielding larger (smaller) values for $p^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma}))$, the original likelihood is inflated (deflated).

In the special case of external benchmarking where $\boldsymbol{m}$ comes from completely separate data sources from $\boldsymbol{y}$ and where $p^{[m\,|\,\psi]}$ describes the sampling distribution of $\boldsymbol{m}$ given $\boldsymbol{\psi}$, the revised likelihood gives the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{m}$ given the parameters $\boldsymbol{\gamma}$. But in external benchmarking where $p^{[m\,|\,\psi]}$ is not equal to the sampling distribution of $\boldsymbol{m}$ given $\boldsymbol{\psi}$, or in internal benchmarking, $p^{[m\,|\,\psi]}$ cannot be interpreted as a standard component of the likelihood, but rather as a device for enforcing the extra requirement to agree with the benchmarks.

With the revised likelihood, the benchmarked posterior distribution is given by

$$p(\boldsymbol{\gamma}, \boldsymbol{\phi}\,|\,\boldsymbol{y}, \boldsymbol{m}) \propto p(\boldsymbol{\phi})p(\boldsymbol{\gamma}\,|\,\boldsymbol{\phi})p(\boldsymbol{y}\,|\,\boldsymbol{\gamma})^{[m/\psi]}(\boldsymbol{m}\,|\boldsymbol{f}(\boldsymbol{\gamma})). \tag{3}$$

In external benchmarking, a possible alternative approach is to incorporate the benchmarks into the prior. Under this approach, conditional on the benchmarks $\boldsymbol{m}$, the parameters $\boldsymbol{\psi}$ are assumed to have a prior distribution $p^{[\psi\,|\,m]}(\boldsymbol{\psi}\,|\,\boldsymbol{m})$. There is a second prior $p^*(\boldsymbol{\psi})$, implied by $p(\boldsymbol{\phi})p(\boldsymbol{\gamma}\,|\,\boldsymbol{\phi})$ and $\boldsymbol{\psi} = \boldsymbol{f}(\boldsymbol{\gamma})$. The two priors $p^{[\psi\,|\,m]}(\boldsymbol{\psi}\,|\,\boldsymbol{m})$ and $p^*(\boldsymbol{\psi})$ need to be combined. This can be regarded as a special case of Bayesian melding proposed by Poole and Raftery (2000). Poole and Raftery (2000) note that the problem of combining priors is addressed by the literature on combining expert judgements, with a standard method being logarithm pooling, which leads to the pooled prior distribution for $\boldsymbol{\psi}$,

$$\tilde{p}(\boldsymbol{\psi}\,|\,\boldsymbol{m}) \propto \left[p^*(\boldsymbol{\psi})\right]^{\alpha}\left[p^{[\psi|m]}(\boldsymbol{\psi}\,|\,\boldsymbol{m})\right]^{1-\alpha}. \tag{4}$$

for some value $0 < \alpha < 1$. However, Equation (4) needs to be inverted, through a complicated procedure, to the parameter space for $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ to yield a pooled prior distribution $\tilde{p}(\boldsymbol{\gamma}, \boldsymbol{\phi}\,|\,\boldsymbol{m})$. Simulating from the corresponding posterior distribution, $\tilde{p}(\boldsymbol{\gamma}, \boldsymbol{\phi}\,|\,\boldsymbol{m})p(\boldsymbol{y}\,|\,\gamma)$, can be difficult with complicated models. Furthermore, logarithm pooling has undesirable properties for probability calculations (O'Hagan et al. 2006, Subsection 9.2.2.).

Under external benchmarking, our approach corresponds to treating benchmarks as data and incorporating them into the likelihood. This approach is also related to the literature on combining expert judgements, in particular Morris (1974), Morris (1977), Lindley et al. (1979), Lindley (1983), Roback and Givens (2001), and Albert et al. (2012), who argue for treating expert judgements as data, and for building models of the accuracy of these judgements. This approach avoids the limitations of logarithm pooling.

2.2. Exact Benchmarking

Under exact benchmarking, model-based estimates are required to agree perfectly with the benchmarks. We interpret perfect agreement to mean that

$$p^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}(\boldsymbol{m}/\boldsymbol{\psi}) = \begin{cases} 1 & \text{if } \boldsymbol{m} = \boldsymbol{\psi}; \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

This interpretation of exact benchmarking is effectively the same as the one adopted by Pfeffermann and Barnard (1991), Pfeffermann and Tiller (2006), Fabrizi et al. (2012), and Pfeffermann et al. (2014), all of whom take frequentist approaches, and Nandram and Sayit (2011), who take a fully Bayesian approach. These methods all treat the benchmarks as constraints on the small area parameters.

When Equation (5) is plugged into Equation (3), the benchmarked posterior distribution becomes a singular distribution concentrated on the region $\{(\boldsymbol{\gamma},\boldsymbol{\phi}) : \boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}\}$. Every $\boldsymbol{\gamma}$ in the posterior distribution satisfies the restriction $\boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}$. Therefore, any point estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, such as the posterior mean or posterior median, satisfies $\boldsymbol{f}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{m}$. We show how samples can be generated from the singular posterior distribution in Subsection 2.5.

2.3. Inexact Benchmarking

Under inexact benchmarking, $p^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}$ is a non-degenerate distribution. The statistician can define $p^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}$ so that it operationalizes the definition required by the particular institutional setting. For example, if it is required that most discrepancies are smaller than a given tolerance a, such that $\Pr(\,|m_j - \psi_j| < a \,|\, \psi_j) \geq q$ for $j = 1, \ldots, d$, then it may be appropriate to specify $p^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}$ as

$$m_j \overset{\text{ind}}{\sim} \mathrm{N}\left(\psi_j, \left(\frac{a}{z_{(1-q)/2}}\right)^2\right), \tag{6}$$

where $\overset{\text{ind}}{\sim}$ indicates independent distributions, and $z_{(1-q)/2}$ is the upper $(1 - q)/2$ quantile of a standard normal distribution.

In some applications, the sampling distribution of $\boldsymbol{m}$ given $\boldsymbol{\psi}$, $p_{\text{sample}}^{[\boldsymbol{m}|\boldsymbol{\psi}]}$, is known, and it may be appropriate to incorporate the sampling distribution into the measure of agreement. The measure can be customized by including a discrepancy parameter λ, with smaller values of λ enforcing greater agreement. We illustrate with two examples.

In the first example, the data $\boldsymbol{y}$ are obtained from a survey, and the benchmarks $\boldsymbol{m}$ are direct estimates calculated from $\boldsymbol{y}$, with standard errors $\boldsymbol{s}$. If the survey was implemented well, then $\boldsymbol{m}$ should be unbiased for $\boldsymbol{\psi}$, and $\boldsymbol{s}$ should be approximately correct. If each m_j is derived from a large number of observations, then we can assume that, conditional on the ψ_j, each m_j is independently normally distributed with mean ψ_j and standard deviation s_j. The sampling distribution $p_{\text{sample}}^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}$ is given by

$$p_{\text{sample}}^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}(\boldsymbol{m}\,|\,\boldsymbol{\psi}) \propto \exp\left(-\sum_{j=1}^{d} \frac{(m_j - \psi_j)^2}{2s_j^2}\right). \tag{7}$$

Incorporating a discrepancy parameter $0 < \lambda \leq 1$ into Equation (7) yields

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi}) \propto \exp\left(-\sum_{j=1}^{d} \frac{(m_j - \psi_j)^2}{2\lambda s_j^2}\right). \tag{8}$$

When $\lambda \rightarrow 0$, Equation (8) converges to Equation (5). Hence exact benchmarking is a limiting case of inexact benchmarking.

In the second example, the data are counts of events that follow Poisson distributions. We have $y_i \sim \text{Poisson}(w_i \gamma_i)$, where w_i is the known exposure for area i. Let $v_j = \sum_{i \in \delta_j} w_i$ denote the total exposure associated with δ_j. Then $\psi_j = \sum_{i \in \delta_j} w_i \gamma_i / v_j$, and its estimate is $m_j = \sum_{i \in \delta_j} y_i / v_j$ where $\sum_{i \in \delta_j} y_i \sim \text{Poisson}\left(\sum_{i \in \delta j} w_i \gamma_i\right) \sim \text{Poisson}(v_j \psi_j)$. The sampling distribution $p_{\text{sample}}^{[\boldsymbol{m} \lceil \boldsymbol{\psi}]}$ is given by

$$p_{\text{sample}}^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi}) \propto \prod_{j=1}^{d} \text{Poisson}(v_j m_j \mid v_j \psi_j). \tag{9}$$

Incorporating a discrepancy parameter $0 < \lambda \leq 1$ into (9) yields

$$p_{\text{sample}}^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\mathbf{m} \mid \boldsymbol{\psi}) \propto \prod_{j=1}^{d} \text{Poisson}(\lambda v_j m_j \mid \lambda v_j \psi_j), \tag{10}$$

with convergence to exact benchmaking as $\lambda \rightarrow 0$.

In the above examples of $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}$, we have assumed conditional independence of m_j's given the underlying parameters ψ_j's, and used simple models for $p(m_j \mid \psi_j)$. This is similar to assuming conditional independence of y_i's given γ_i's and using simple models for $p(y_i \mid \gamma_i)$. Unconditionally, the m_j's can have complicated correlations, such as correlations between neighbouring time points or age groups. Such correlations are captured by the prior model on the underlying benchmarking parameters $\boldsymbol{\psi}$, which is implied by the prior model on $\boldsymbol{\gamma}$, $p(\boldsymbol{\phi})p(\boldsymbol{\gamma} \mid \boldsymbol{\phi})$, and the equality $\boldsymbol{\psi} = \boldsymbol{f}(\boldsymbol{\gamma})$. The prior model on $\boldsymbol{\gamma}$ typically uses a complicated hierarchical structure to model relationship between the underlying parameters, such as similarities between neighbouring time points or age groups.

The appropriate value for the discrepancy parameter λ in any particular application depends on the sizes of discrepancies between model-based estimates and benchmarks that can be tolerated in that application. As we discuss in Subsection 2.6, the effects of benchmarking on performance measures such as accuracy are difficult to predict. One possible approach to setting λ is to fit a model several times with alternative values for λ, and use the highest value that gives acceptable levels of discrepancy.

2.4. An Illustrative Analytical Example

To illustrate the benchmarked posterior distribution, we present an example in which the distribution can be derived in closed form. The data $\boldsymbol{y} = \{y_i, \ldots, y_{\text{n}}\}^\top$ are generated and modelled using

$$y_i \overset{\text{ind}}{\sim} \text{N}(\gamma_i, \sigma^2) \tag{11}$$

$$\gamma_i \overset{\text{ind}}{\sim} \text{N}(\mu_0, \tau^2), \tag{12}$$

where μ_0, σ^2 and τ^2 are known. There is a single benchmark $m = \sum_{i=1}^{n} w_i y_i$ estimating benchmarking parameter $\psi = \sum_{i=1}^{n} w_i \gamma_i$, where the w_i's are a set of weights satisfying $\sum_{i=1}^{n} w_i = 1$. Let $\boldsymbol{w} = (w_1, \ldots, w_{\mathrm{n}})^{\top}$ denote the vector of weights, and $\mathbf{1}_n$ a vector of n ones. Then $\boldsymbol{w}^{\top}\boldsymbol{y} = m$, $\boldsymbol{w}^{\top}\boldsymbol{\gamma} = \psi$, and $\boldsymbol{w}^{\top}\mathbf{1}_n = 1$. Under exact benchmarking, $p^{[\mathbf{m}\,|\,\boldsymbol{\psi}]}$ is given by Equation (5). Under inexact benchmarking, the sampling distribution $p^{[\boldsymbol{m}\,|\,\boldsymbol{\psi}]}_{\text{sample}}$ is given by

$$p^{[m\,|\,\psi]}_{\text{sample}}\big(m \mid \psi\big) \sim \mathrm{N}\big(\psi, \big(\boldsymbol{w}^{\top}\boldsymbol{w}\big)\sigma^2\big). \tag{13}$$

We incorporate a discrepancy parameter λ into (13) and arrive at

$$p^{[m\,|\,\psi]}\big(m \mid \psi\big) \sim \mathrm{N}\big(\psi, \lambda\big(\boldsymbol{w}^{\top}\boldsymbol{w}\big)\sigma^2\big), \tag{14}$$

where $0 < \lambda \leq 1$.

Let $\boldsymbol{I}_{\mathrm{n}}$ be the $n \times n$ identity matrix. As shown in the Supplemental data (Section 1), Equations (5), (11), (12) and (14) yield posterior distributions for $\boldsymbol{\gamma}$ that are multivariate normal under no benchmarking (NB), exact benchmarking (EB), and inexact benchmarking (IB), with means and variances

$$\boldsymbol{\mu}^{\mathrm{NB}} = -\frac{\sigma^2}{\sigma^2+\tau^2}\mathbf{1}_n\mu_0 + \frac{\tau^2}{\sigma^2+\tau^2}\boldsymbol{y}, \tag{15}$$

$$\boldsymbol{\Sigma}^{\mathrm{NB}} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\boldsymbol{I}_n, \tag{16}$$

$$\boldsymbol{\mu}^{\mathrm{EB}} = \frac{\sigma^2}{\sigma^2+\tau^2}\left[\mathbf{1}_n - \frac{1}{\boldsymbol{w}^{\top}\boldsymbol{w}}\boldsymbol{w}\right]\mu_0 + \frac{\sigma^2}{\sigma^2+\tau^2}\frac{1}{\boldsymbol{w}^{\top}\boldsymbol{w}}\boldsymbol{w}m + \frac{\tau^2}{\sigma^2+\tau^2}\boldsymbol{y}, \tag{17}$$

$$\boldsymbol{\Sigma}^{\mathrm{EB}} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\left[\boldsymbol{I}_n - \frac{1}{\boldsymbol{w}^{\top}\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}^{\top}\right], \tag{18}$$

$$\boldsymbol{\mu}^{\mathrm{IB}} = \left[1 - \frac{\tau^2}{\lambda\sigma^2+(\lambda+1)\tau^2}\right]\boldsymbol{\mu}^{\mathrm{NB}} + \frac{\tau^2}{\lambda\sigma^2+(\lambda+1)\tau^2}\boldsymbol{\mu}^{\mathrm{EB}}, \tag{19}$$

$$\boldsymbol{\Sigma}^{\mathrm{IB}} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\left[\boldsymbol{I}_n - \frac{\tau^2}{\big(\lambda\sigma^2+(\lambda+1)\tau^2\big)\boldsymbol{w}^{\top}\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}^{\top}\right]. \tag{20}$$

With no benchmarking, the posterior mean for γ_i equals the observation y_i shrunk towards the prior mean μ_0. With exact benchmarking, the posterior mean is instead shrunk towards a linear combination of the prior mean μ_0 and the benchmark m. With inexact benchmarking, the posterior mean is a compromise between the means under no benchmarking and exact benchmarking. Benchmarking reduces posterior variance in this setting, with exact benchmarking leading to larger reductions than inexact benchmarking.

2.5. A General MCMC Approach to Sampling from a Benchmarked Posterior Distribution

In practical applications, closed form expressions for the benchmarked posterior distribution Equation (3) are seldom available, and posterior inference must be carried out via simulation. We outline a general MCMC strategy for sampling from Equation (3).

We first discuss the case where the components of the benchmarking function are linear. Under exact benchmarking, Equation (3) is a singular distribution concentrated on the region $\{(\boldsymbol{\gamma}, \boldsymbol{\phi})\colon \boldsymbol{B}^\top\boldsymbol{\gamma} = \boldsymbol{m}\}$. We obtain draws from this singular distribution by choosing an initial value $\boldsymbol{\gamma}^{(0)}$ that satisfies $\boldsymbol{B}^\top\boldsymbol{\gamma}^{(0)} = \boldsymbol{m}$, and then repeatedly iterating through the following steps:

E1. Update $\boldsymbol{\gamma}^{(t)} \mid \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{y}$ subject to $\boldsymbol{B}^\top\boldsymbol{\gamma}^{(t)} = \boldsymbol{m}$.

E2. Update $\boldsymbol{\phi}^{(t)} \mid \boldsymbol{\gamma}^{(t)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{y}$.

Step E2 can be done using standard methods. Step E1 ensure that the constraint $\boldsymbol{B}^\top\boldsymbol{\gamma} = \boldsymbol{m}$ continues to be satisfied. It is carried out as follows.

An area i_1 is randomly selected from $\{1, \ldots, n\}$. If area i_1 does not belong to any δ_j, so that γ_{i_1} is not subject to any benchmarking constraint, then γ_{i_1} is updated using standard methods. If area i_1 is the only area in an δ_j, then γ_{i_1} is fully determined by the benchmark m_j and is not updated. Otherwise, γ_{i_1} is updated through a Metropolis-Hastings step. A proposal $\gamma^*_{i_1}$ is generated from $J\left(\gamma^*_{i_1} \mid \gamma^{(t-1)}_{i_1}\right)$. Then another area i_2 from the same δ_j that i_1 belongs to is randomly selected, and $\gamma^*_{i_2}$ is obtained as $\gamma^*_{i_2} = \gamma^{(t-1)}_{i_2} + b_{i_1j}/b_{i_2j}\left(\gamma^{(t-1)}_{i_1} - \gamma^*_{i_1}\right)$, which ensures that $b_{i_1j}\gamma^*_{i_1} + b_{i_2j}\gamma^*_{i_2} = b_{i_1j}\gamma^{(t-1)}_{i_1} + b_{i_2j}\gamma^{(t-1)}_{i_2}$. Setting $\gamma^*_i = \gamma^{(t-1)}_i$ for $i \notin \{i_1, i_2\}$ yields a proposed value γ^* for which $\boldsymbol{B}^\top\boldsymbol{\gamma}^* = \boldsymbol{m}$ continues to be satisfied.

To calculate the joint proposal density $J\left(\left(\gamma^*_{i_1}, \gamma^*_{i_2}\right) \mid \left(\gamma^{(t-1)}_{i_1}, \gamma^{(t-1)}_{i_2}\right)\right)$, we need to take account of the fact that we could have arrived at $\left(\gamma^*_{i_1}, \gamma^*_{i_2}\right)$ in one of two ways: by drawing $\gamma^*_{i_1}$ and then calculating $\gamma^*_{i_2}$ or by drawing $\gamma^*_{i_2}$ and then calculating $\gamma^*_{i_1}$. The resulting Metropolis-Hastings ratio is

$$r = \left[\frac{p\left(\boldsymbol{\gamma}^* \mid \boldsymbol{\phi}\right)p\left(\boldsymbol{y} \mid \boldsymbol{\gamma}^*\right)}{p\left(\boldsymbol{\gamma}^{(t-1)} \mid \boldsymbol{\phi}\right)p\left(\boldsymbol{y} \mid \boldsymbol{\gamma}^{(t-1)}\right)}\right] \times \frac{J\left(\gamma^{(t-1)}_{i_1} \mid \gamma^*_{i_1}\right) + \left|b_{i_1j}/b_{i_2j}\right| J\left(\gamma^{(t-1)}_{i_2} \mid \gamma^*_{i_2}\right)}{J\left(\gamma^*_{i_1} \mid \gamma^{(t-1)}_{i_1}\right) + \left|b_{i_1j}/b_{i_2j}\right| J\left(\gamma^*_{i_2} \mid \gamma^{(t-1)}_{i_2}\right)}. \qquad (21)$$

Since the benchmarked posterior distribution Equation (3) under exact benchmarking is a singular distribution, it is not immediately obvious that the above Metropolis-Hastings algorithm has the desired convergence property. The Supplemental data (Subsection 2.1) provides a proof that under exact benchmarking, the stationary distribution of chains produced by the above algorithm is indeed Equation (3).

Under inexact benchmarking, the algorithm for sampling from Equation (3) is

I1. Update $\boldsymbol{\gamma}^{(t-\frac{1}{2})} \mid \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{y}$ subject to the constraint $\boldsymbol{B}^\top\boldsymbol{\gamma}^{(t-1)} = \boldsymbol{\psi}^{(t-1)}$, where $\boldsymbol{\psi}^{(t-1)} = \boldsymbol{B}^\top\boldsymbol{\gamma}^{(t-1)}$.

I2. Update $\boldsymbol{\gamma}^{(t)} \mid \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\gamma}^{(t-\frac{1}{2})}, \boldsymbol{y}$ with no constraint.

I3. Update $\boldsymbol{\phi}^{(t)} \mid \boldsymbol{\gamma}^{(\mathrm{t})}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{y}$.

Step I1 is similar to step E1 with exact benchmarking, and step I3 is the same as step E2 with exact benchmarking. Step I2 can be carried out using Metropolis-Hastings updates similar to those for an unbenchmarked model, except that the density $p^{[\boldsymbol{m}/\boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{B}^{\top}\boldsymbol{\gamma})$ needs to be accounted for in the Metropolis-Hastings ratio. Step I1 is not strictly necessary, but speeds up the exploration of the parameter space when $p^{[\boldsymbol{m}/\boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{B}^{\top}\boldsymbol{\gamma})$ is tightly concentrated around the hyperplane defined by $\boldsymbol{m} \mid \boldsymbol{B}^{\top}\boldsymbol{\gamma}$. The Supplemental data (Subsection 2.2) provides a proof that under inexact benchmarking, the stationary distribution of chains produced by the above algorithm is the benchmarked posterior distribution in Equation (3).

When the components of the benchmarking function are nonlinear, under inexact benchmarking Equation (3) is a singular distribution concentrated on the region $\{(\boldsymbol{\gamma},\boldsymbol{\phi}) : \boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}\}$. There is generally no efficient way to implement a step similar to E1 or I1 which ensures that the constraint $\boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{m}$ or $\boldsymbol{f}(\boldsymbol{\gamma}) = \boldsymbol{\psi}^{(t-1)}$ continues to be satisfied. Instead we use steps I2 and I3 for inexact benchmarking, and approximate exact benchmarking by using inexact benchmarking with discrepancy parameter λ close to zero.

We have implemented our general MCMC approaches with a specific family of area-level hierarchical models:

$$y_i \mid \gamma_i, \sigma^2 \overset{\text{ind}}{\sim} G\left(\gamma_i, w_i, \sigma^2\right), \tag{22}$$

$$g\left(\gamma_i\right) \mid \boldsymbol{\beta}, \tau^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}, \tau^2\right). \tag{23}$$

In Equation (22), y_i is an observation for area i within a multiway classification, G denotes the normal, Poisson or binomial distribution, w_i is a known weight, exposure or number of trials, and σ^2 is a variance, used only with the normal distribution. In Equation (23), g is the identity, log or logit link function. The transformed values $g(\gamma_i)$ are modelled using a structure similar to analysis of variance. Vector $\boldsymbol{\beta}$ contains batches of coefficients representing main effects and interactions formed from the cross-classifying dimensions. Vector $\boldsymbol{x}_i$ is a vector consisting of ones and zeros indicating which main effects and interactions are associated with each area i. We place no restrictions on the prior for $\boldsymbol{\beta}$, and it will typically have a complicated hierarchical structure. The Supplemental data (Section 3) gives details of the specific MCMC samplers.

We have written R packages implementing the models, which can be obtained from github.com/statisticsnz/R. The family of models included in the packages can accommodate a wide range of real applications, and the packages are user-friendly, making it easy for practitioners to implement the fully Bayesian benchmarking approached presented in this article.

2.6. *The Effects of External and Internal Benchmarking on Model Performance*

External benchmarking allows information from the external data sources to be incorporated into the analysis. When $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}$ is constructed from a correctly specified $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}_{\text{sample}}$, external benchmarking should, on average, improve model performance as measured by criteria such as accuracy and coverage.

The effect of internal benchmarking on accuracy and coverage is more ambiguous. Internal benchmarking entails using data $\boldsymbol{y}$ twice: once when calculating benchmarks $\boldsymbol{m}$, and again in $p(\boldsymbol{y} \mid \boldsymbol{\gamma})$. If a correctly-specified model is subject to internal benchmarking, then it is no longer correctly specified. Performance on accuracy and coverage can be expected to suffer. Previous studies with frequentist and empirical Bayes approaches have confirmed that this is indeed the case: when the unbenchmarked model is correctly specified, benchmarking typically reduces accuracy and coverage (Pfeffermann and Barnard 1991; Wang et al. 2008; Datta et al. 2011; Bell et al. 2013).

When the unbenchmarked model is correctly specified, methods of benchmarking that enforce stronger forms of agreement with the benchmarks can be expected to perform worse in terms of accuracy and coverage. For example, exact benchmarking can be expected to have poorer accuracy and coverage than inexact benchmarking.

In real applications, however, the model is almost always misspecified. When the model is misspecified, the effects of internal benchmarking on model performance are uncertain. Previous simulation studies suggest that, depending on the details of the data and model, internal benchmarking can sometimes improve performance (Pfeffermann and Tiller 2006; Nandram et al. 2011; Pfeffermann 2013; Vesper 2013; Ranalli et al. 2018).

Given the uncertainty about the effect of benchmarking on model performance, we suggest that benchmarking not be seen as a method for protecting against model misspecification. Instead, analysts should use standard model-checking tools such as posterior predictive checks (Gelman et al. 2014, chap. 6) to detect possible problems with their models, and adjust the models accordingly. Benchmarking should, rather, be seen as a method for achieving agreement between model-based estimates and benchmarks.

3. Comparison with Previous Approaches

3.1. An Alternative Interpretation of Exact Benchmarking

In our interpretation of exact benchmarking, set out in Equation (5), the entire posterior distribution must agree with the benchmarks. Under this approach, any standard point estimate derived from the posterior distribution, such as the posterior mean or posterior median, automatically agrees with the benchmarks.

Most previous approaches interpret exact benchmarking less strictly. Instead of working with full distributions, they work only with point estimates. They require a specific point estimate, $\hat{\gamma}^{\mathrm{Spe}}$, to agree with the benchmarks,

$$\boldsymbol{f}\left(\hat{\gamma}^{\mathrm{Spe}}\right) = \boldsymbol{m}. \tag{24}$$

You et al. (2004), Datta et al. (2011), and Ghosh et al. (2015), for example, obtain point estimates $\hat{\gamma}^{\mathrm{FB}}$ from a fully Bayesian model, and then adjust them to obtain a new set of estimates $\hat{\gamma}^{\mathrm{Spe}}$ that satisfy the benchmarking constraint. When the benchmarks are linear, one such estimator is the raked or ratio-adjusted estimator,

$$\hat{\gamma}_i^{\mathrm{Spe}} = \hat{\gamma}_i^{\mathrm{FB}} \frac{m_j}{\sum_{i'=1}^{n} b_{i'j} \hat{\gamma}_{i'}^{\mathrm{FB}}}. \tag{25}$$

The raked estimator is easy to implement, and is widely used in practice, but has been characterised as ad hoc (Ghosh et al. 2015). Datta et al. (2011) (henceforth DGSM) instead propose an estimator that minimizes the expected posterior loss based on a weighted squared error loss function $L(\boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \rho_i (\hat{\gamma}_i - \gamma_i)^2$, where ρ_i are known weights and $\hat{\gamma}$ satisfies $\boldsymbol{f}(\hat{\gamma}) = \boldsymbol{m}$. As DGSM point out, with the appropriate choice of ρ_i, the raked estimator can be derived as a special case of their estimator. Ghosh et al. (2015) argue that the squared error loss function may not be appropriate for estimating positive quantities such as income, and propose an estimator that minimizes the expected posterior loss based on a variant of the Kullback-Leibler loss function.

Our approach to exact benchmarking enforces stronger forms of agreement with the benchmarks. As discussed in Subsection 2.6, this can lead to poorer performance on criteria such as accuracy and coverage, when the unbenchmarked model is correctly specified. However, when the model is misspecified, depending on the details of the data and model, enforcing stronger forms of agreement may sometimes improve performance, as we illustrate in Subsection 4.1.

Requiring that the entire posterior distribution agree with the benchmarks, as we do with exact benchmarking, means that the statistician does not have to choose a particular loss function. Moreover, with the entire posterior distribution available, the statistician can obtain a posterior distribution for any function of the small area parameters (Gelman et al. 2014, 261–262). For instance, given a posterior distribution for county-level income levels, the statistician can derive a posterior distribution for county income rankings.

3.2. *Previous Fully Bayesian Benchmarking Approaches*

Toto and Nandram (2010) and Nandram et al. (2011) use fully Bayesian benchmarking on a model where the posterior distribution is multivariate normal, and where there is a single benchmark and a linear benchmarking function. Their model is specified at the unit level rather than the area level. We discuss unit models further in the Supplemental data, Section 4.

Nandram and Sayit (2011) benchmark an area-level beta-binomial hierarchical Bayesian model,

$$y_i \mid \gamma_i \overset{\text{ind}}{\sim} \text{Binomial}(w_i, \gamma_i), \tag{26}$$

$$\gamma_i \mid \mu, \tau \overset{\text{ind}}{\sim} \text{Beta}\left(\mu\tau, (1-\mu)\tau\right), \tag{27}$$

$$p(\mu, \tau) \propto \left(1+\tau^2\right)^{-1}, \quad 0 < \mu < 1, \quad \tau \geq 0. \tag{28}$$

Here w_i is a known number of trials for area i. The authors work with a single benchmarking parameter $\psi = \sum_{i=1}^{n} b_i \gamma_i$, where $b_i = w_i / \sum_{i'=1} w_{i'}$. Instead of incorporating the benchmark into the likelihood, Nandram and Sayit (2011) incorporate it into a prior distribution for ψ, $p(\psi) \sim \text{Beta}(m\tau_0, (1-m)\tau_0)$. The authors consider three scenarios for $p(\psi)$: (1) exact benchmarking, with $\tau_0 \to \infty$ and $p(\psi)$ a point mass at m; (2) inexact benchmarking, with τ_0 specified by the user; and (3) inexact benchmarking, with $m = 1/2$ and $\tau_0 = 2$, so that $p(\psi) \sim \text{Uniform}[0, 1]$.

Let $p_{\text{NB}}(\gamma_1, \ldots, \gamma_{n-1}, \gamma_n, \mu, \tau \mid \boldsymbol{y})$ denote the unbenchmarked posterior distribution for (γ, μ, τ). By using the identity $\gamma_n = \left(\psi - \sum_{i=1}^{n-1} b_i \gamma_i\right) / b_n$, Nandram and Sayit (2011) are able to work with $(\gamma_1, \ldots, \gamma_{n-1}, \psi)$ instead of $(\gamma_1, \ldots, \gamma_{n-1}, \gamma_n)$, and derive the benchmarked posterior distribution for ($\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau$) as

$$p(\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau \,|\, \mathbf{y}) \propto p(\psi) p_{\mathrm{NB}}\left(\gamma_1, \ldots, \gamma_{n-1}, \frac{1}{b_n}\left(\psi - \sum_{i=1}^{n-1} b_i \gamma_i\right), \mu, \tau \,|\, \mathbf{y}\right). \tag{29}$$

Nandram and Sayit (2011) use a Gibbs sampling algorithm to sample ($\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau$) from Equation (29). The full conditional distributions for γ_i ($i = 1, \ldots, n-1$) and ψ are both proportional to the product of two density functions, one being a truncated beta density and the other being a generalized beta density. Specialized algorithms are used to draw samples from these distributions. After obtaining samples for ($\gamma_1, \ldots, \gamma_{n-1}, \psi, \mu, \tau$), samples for γ_{n} are then obtained using the identity $\gamma_n = \left(\psi - \sum_{i=1}^{n-1} b_i \gamma_i\right)/b_n$.

Implementation of the approach used by Nandram and Sayit (2011) depends on the choice of which small area is labeled as area n and left out. This choice affects the specific posterior distribution derived in Equation (29) and hence affects the computational efficiency of the MCMC algorithms. Nandram and Sayit (2011) sort the areas in ascending order of y_i, with area n having the largest value of y_i. When there are multiple benchmarks, with this approach, one area needs to be left out in each δ_j. Poor choices may lead to poor computational efficiency. In contrast, our MCMC approach in Subsection 2.5 does not depend on the labeling of areas.

The approach of Nandram and Sayit (2011) is also difficult to generalize to nonlinear benchmarking functions. Even with a single benchmarking parameter $\psi = f(\gamma_1, \ldots, \gamma_n)$ where f is nonlinear, it can be difficult to write γ_n analytically as a function of $\gamma_1, \ldots, \gamma_{n-1}, \psi$. Therefore, it may not be possible to write out a benchmarked posterior distribution similar to Equation (29), or to draw samples from it. In contrast, our approach can accommodate nonlinear benchmarking functions.

Vesper (2013) works with the Fay-Herriot model (Fay and Herriot 1979):

$$y_i \,|\, \gamma_i \overset{\mathrm{ind}}{\sim} \mathrm{N}\left(\gamma_i, \sigma_i^2\right), \tag{30}$$

$$\gamma_i \,|\, \boldsymbol{\beta}, \tau^2 \overset{\mathrm{ind}}{\sim} \mathrm{N}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \tau^2\right), \tag{31}$$

$$p\left(\boldsymbol{\beta}, \tau^2\right) \propto \left(\tau^2\right)^{-a-1} e^{-b/\tau^2}, \tag{32}$$

where σ_i^2 is the variance of y_i and is assumed known, $\boldsymbol{x}_i$ is an observed vector of covariates for area i, and a and b are known constants. There is a single benchmarking parameter $\psi = \sum_{i=1}^n b_i \gamma_i$, and $p(\psi) \sim N\left(m, \sum_{i=1}^n b_i^2 \sigma_i^2\right)$. The benchmarked posterior distribution is similar to Equation (29) from Nandram and Sayit (2011). This approach is subject to the same implementation limitations as that of Nandram and Sayit (2011).

4. Two Simulation Studies

4.1. Estimation of Fertility Rates from Registration Data

We use simulated data on births to examine how benchmarking affects accuracy, coverage, and agreement between model-based estimates and benchmarks. We examine performance with correctly specified models and with misspecified models. All code and data for the simulation are available at github.com/bayesiandemography/fertsim.

The simulated data consist of counts of births and reproductive-age women. With the baseline ‘no change’ data sets, counts of births are generated using the model

$$y_{art} \overset{\text{ind}}{\sim} \text{Poisson}(w_{art}\gamma_{art}^{\text{Tr}}), \tag{33}$$

$$\log \gamma_{art}^{\text{Tr}} \overset{\text{ind}}{\sim} \text{N}(\beta_a^{\text{age-std}} + \beta_r^{\text{reg-Tr}} + \beta_t^{\text{time-Tr}}, \sigma_{\text{Tr}}^2), \tag{34}$$

where y_{art} is the number of births to women in age group $a \in \{15-19, \ldots, 40-44\}$ in region $r \in \{1, \ldots, 30\}$ during period $t \in \{1, 2, 3\}$, w_{art} is the corresponding person-years of exposure, and γ_{art}^{Tr} is the true underlying birth rate. We set $w_{art} = 300$ for all a, r, t, and set $\sigma_{\text{Tr}} = 0.1$. Age effects $\beta_a^{\text{age-std}}$ are taken from the ‘standard’ fertility schedule in Table 1; region effects have distribution $\beta_r^{\text{reg-Tr}} \overset{\text{ind}}{\sim} \text{N}(0, 0.1^2)$; and time effects have distribution $\beta_t^{\text{time-Tr}} \overset{\text{ind}}{\sim} \text{N}(0, 0.1^2)$.

To explore how benchmarking affects performance under model misspecification, we also construct ‘change in level’ and ‘change in distribution’ data sets by perturbing the ‘no change’ data set. The ‘change in level’ data set represents a sudden change in the level of fertility in a subset of regions. Birth rates and counts are identical to the ‘no change’ data set, except for areas in regions 26–30 during period 3, which we refer to as being ‘nonstandard’ areas. Log rates for the nonstandard areas are generated by adding log 0.2, to the existing log rates. Counts for nonstandard areas are obtained by drawing new values from Equation (33). The ‘change in distribution’ data set represents a sudden change in the age-pattern, rather than the level, of fertility. Birth rates and counts are again identical to the ‘no change’ data set except in regions 26–30 during period 3. Log rates for nonstandard areas are generated by replacing the ‘standard’ age effects from Table 1 with the ‘nonstandard’ age effects, and leaving region effects and time effects the same. Counts for nonstandard areas are obtained by drawing new values from Equation (33).

We simulate an analysis seeking to estimate γ_{art}^{Tr}. The simulated analysis model has likelihood

$$y_{art} \overset{\text{ind}}{\sim} \text{Poisson}(w_{art}\gamma_{art}), \tag{35}$$

and assumes that

$$\log \gamma_{art} \overset{\text{ind}}{\sim} \text{N}(\beta^0 + \beta_a^{\text{age}} + \beta_r^{\text{reg}} + \beta_t^{\text{time}}, \sigma^2). \tag{36}$$

Table 1. Age effects (exponentiated).

	Standard	Nonstandard
15–19	0.0288	0.0695
20–24	0.0713	0.1225
25–29	0.1083	0.0936
30–34	0.1210	0.0726
35–39	0.0653	0.0394
40–44	0.0127	0.0098
Total	0.4074	0.4074

Equation (36) is correctly specified for the ‘no change’ data sets. However, it is misspecified for the ‘change in level’ and ‘change in distribution’ data sets, since the true data-generating processes for these data sets contain interactions.

The intercept term in the simulated analysis model has a proper but diffuse prior, $\beta^0 \sim \mathrm{N}(0, 10^2)$. The region effect has a normal prior $\beta_r^{\mathrm{reg}} \sim \mathrm{N}(0, \tau_{\mathrm{reg}}^2)$, with a weakly informative half-t prior with seven degrees of freedom on the standard deviation (Gelman et al. 2008), $\tau_{\mathrm{reg}} \sim t_7^+(0, 0.25^2)$. The age effect has a ‘random walk with noise’ (Prado and West 2010, 119–120) prior,

$$\beta_a^{\mathrm{age}} \overset{\mathrm{ind}}{\sim} \mathrm{N}(\eta_a^{\mathrm{age}} \tau_{\mathrm{age}}^2) \tag{37}$$

$$\eta_a^{\mathrm{age}} \overset{\mathrm{ind}}{\sim} \mathrm{N}(\eta_{a-1}^{\mathrm{age}} \omega^2) \tag{38}$$

with $\eta_0^{\mathrm{age}} \sim \mathrm{N}(0, 10^2)$, $\tau_{\mathrm{age}} \sim t^+(0, 1)$, and $\omega \sim t_7^+(0, 1)$. The random walk with noise prior recognizes the tendency for neighbouring age groups to have similar values. The standard deviation parameter from Equation (36) has a weakly informative half-t prior, $\sigma \sim t_7^+(0, 0.25^2)$. The time effect has the same prior as the region effect.

The analysis model is fitted with (i) no benchmarking, (ii) exact benchmarking, and (iii) inexact benchmarking. Let $v_{rt} = \sum_a w_{art} = 1,800$. The benchmarks are region-time means $m_{rt} = \sum_a y_{art}/v_{rt}$, which estimate benchmarking parameters $\psi_{rt} = \sum_a w_{art}\gamma_{art}/v_{rt}$. Under exact benchmarking, $p^{[\mathbf{m} \mid \boldsymbol{\psi}]}(\mathbf{m} \mid \boldsymbol{\psi})$ is given by Equation (5). Under inexact benchmarking, $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi})$ is given by Equation (10), which becomes

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi}) \propto \prod_{r=1}^{30} \prod_{t=1}^{3} \mathrm{Poisson}(\lambda v_{rt} m_{rt} \mid \lambda v_{rt} \psi_{rt}). \tag{39}$$

We consider the case where $\lambda = 1$, which allows discrepancies between model-based estimates and benchmarks to vary in line with Poisson variation in birth counts. As discussed in Subsection 2.3, lower values for λ would lead to smaller discrepancies. We use the posterior means of γ_{art} and ψ_{rt} as point estimators.

We also adjust the posterior means of γ_{art} from the ‘no benchmarking’ case to obtain the raked estimator in Equation (25), and the DGSM estimator based on a weighted squared error loss function. Following Datta et al. (2011, 580) and Wang et al. (2008), we set ρ_{art}, the weight in the weighted squared error loss function, equal to the inverse of the estimated variance of the direct estimate y_{art}/w_{art}. Since these two estimators achieve exact benchmarking, the corresponding point estimators of ψ_{rt} are equal to m_{rt}.

We apply four performance measures. The first is

$$D_{rt} = E\left(\frac{|\hat{\psi}_{rt} - m_{rt}|}{m_{rt}}\right), \tag{40}$$

where $\hat{\psi}_{rt}$ is the point estimator of ψ_{rt}. This measure captures discrepancies (i.e., levels of disagreement) between the model-based estimates and benchmarks. With exact benchmarking under our approach, the raked estimator, and the DGSM estimator, D_{rt} always equals 0.

The second performance measure is the mean squared error from using the point estimator $\hat{\gamma}_{rt}$ to estimate $\gamma_{art}^{\mathrm{Tr}}$,

$$\mathrm{MSE}_{art} = \mathrm{E}(\hat{\gamma}_{art} - \gamma_{art}^{\mathrm{Tr}})^2. \tag{41}$$

This measure captures the accuracy of the point estimator.

The third measure, W_i^q, is the expected width of a $(1 - q) \times 100\%$ credible interval for γ_{art}. Let $\gamma_{art}^{q/2}$ and $\gamma_{art}^{1-q/2}$ be the $q/2$ and $1 - q/2$ quantiles for the posterior distribution of γ_{art}. Then

$$W_{art}^q = \mathrm{E}\left(\gamma_{art}^{1-q/2} - \gamma_{art}^{q/2}\right). \tag{42}$$

Values for W_i^q cannot be calculated for the raked and DGSM estimators, since these estimators do not come with measures of uncertainty.

The fourth measure, C_{art}^q, is the expected coverage rate of a $(1 - q) \times 100\%$ credible interval for $\gamma_{art}^{\mathrm{Tr}}$,

$$C_{art}^q = \Pr\left(\gamma_{art}^{q/2} \leq \gamma_{art}^{\mathrm{Tr}} \leq \gamma_{art}^{1-q/2}\right). \tag{43}$$

Again, values for C_{art}^q cannot be calculated for the raked and DGSM estimators.

We use $K = 100$ simulation replicates. As discussed in the Supplementary data, 100 replicates is enough to obtain stable estimates for the performance indicators we are interested in. At each replicate, results for no benchmarking, exact benchmarking, inexact benchmarking, raked estimators, and DGSM estimators are obtained for each of the ‘no change’, ‘change in distribution’, and ‘change in level’ data sets, yielding $5 \times 3 = 15$ sets of results. With the unbenchmarked model, the Gibbs sampler is run with four independent chains, each with 20,000 iterations. Every 40th draw from the final 10,000 iterations of each chain is recorded, yielding a combined total of 2,000 draws from the posterior distribution. With the benchmarked models, which converge more quickly, the number of iterations and thinning ratios are both reduced by a factor of five.

When calculating performance measures D_{rt}, MSE_{art}, W_{art}^q, and C_{art}^q, we use means across K replicates to approximate $\mathrm{E}(\cdot)$ or $\Pr(\cdot)$ in (40)–(43). These measures are calculated separately for each *rt* or *art*. Figure 1 summarizes the resulting distributions across *rt* or *art* using boxplots, where W_{art}^q and C_{art}^q are calculated for $q = 0.95$. The median for each distribution is printed above the corresponding notch in the boxplot.

The top row of Figure 1 gives results for the ‘no change’ data sets, where the analysis model is correctly specified. The model without benchmarking departs furthest from the benchmarks, but has the lowest MSE. The version of our model with exact benchmarking has the opposite strengths and weaknesses, agreeing exactly with the benchmarks (by construction), but having the highest MSE. The raked and DGSM estimators also obtain complete agreement, but with median MSE that is approximately 3–4% lower than the model with exact benchmarking. The model with inexact benchmarking is in an intermediate position, with moderate agreement and moderate MSE.

The models with no benchmarking, inexact benchmarking, and exact benchmarking have coverage rates close to the nominal 95%, but the model without benchmarking achieves this with the narrowest credible intervals. As noted above, the raked and DGSM estimators do not have uncertainty measures and therefore do not have coverage rates.

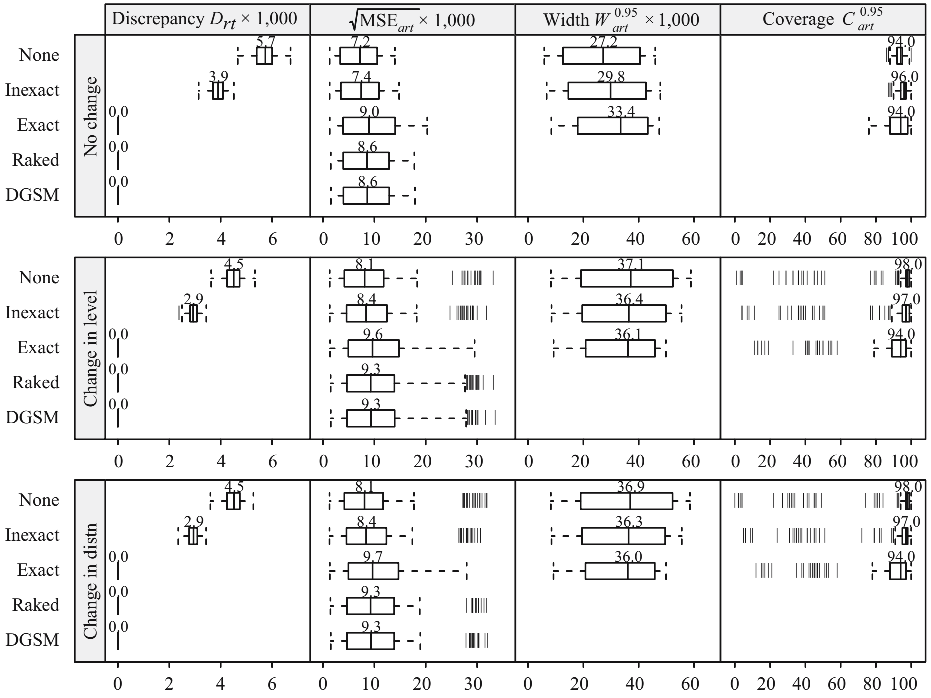


Fig. 1. Performance of models of fertility rates, by type of benchmarking and data set. The results are based on K = 100 replicates.

The second row of Figure 1 shows results for the ‘change in levels’ data sets. The rank order of the discrepancy measures and MSEs is preserved. The rank order of the width of credible intervals has changed, with the model without benchmarking having the widest credible intervals. In general, credible intervals are wider than in the ‘no change’ case, but coverage rates for a subset of areas are poor, with and without benchmarks. Similar results are obtained with the ‘change in distribution’ data sets.

Figure 2 shows results for the ‘change in levels’ data sets, but distinguishing between standard and non-standard areas. In the standard areas, performance is similar to the overall picture in Figure 1. In the nonstandard areas, our benchmarked models have smaller MSE than the non-benchmarked model, and the raked and DGSM estimators. Coverage rates for the nonstandard areas are poor for all three versions of our model, but the model with exact benchmarking has better coverage rates than the other two. As can be seen in Figure 3, a similar pattern is found with ‘change in distribution’ data sets.

4.2. *Estimation of Smoking Prevalence from Survey Data*

In the second simulation we compare the performance of benchmarked and non-benchmarked models when estimating finite-population quantities.

Ideally, the distinction between finite-population and super-population quantities should be reflected in the benchmarking procedures, so that, for instance, agreement with $\boldsymbol{m}$ is measured using the finite-population equivalent of $\boldsymbol{\psi}$. We have not done so, on pragmatic grounds. Using super-population quantities simplifies the MCMC computations, and when

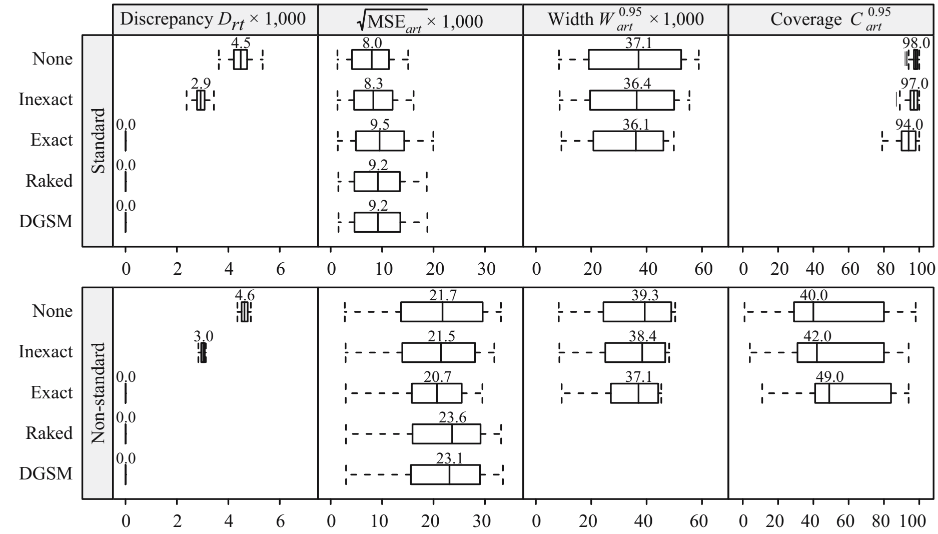


Fig. 2. Performance of model of fertility rates when applied to 'change in level' data sets, distinguishing between standard and nonstandard areas. The results are based on K = 100 replicates.

sample sizes are large, as they typically are for aggregate quantities such as $\boldsymbol{m}$ and $\boldsymbol{\psi}$, super-population quantities closely approximate finite-population ones.

The simulation uses artificial surveys of smoking prevalence constructed from real data from the 2013 New Zealand population census. The artificial surveys are generated by randomly drawing records from a file containing unit-level census data on the population aged 15 and over. The file contains information on age, sex, region within the country, income level, and whether the respondent currently smokes. The file excludes the 8% of people who did not answer the smoking question, leaving a total of 3.07 million records.

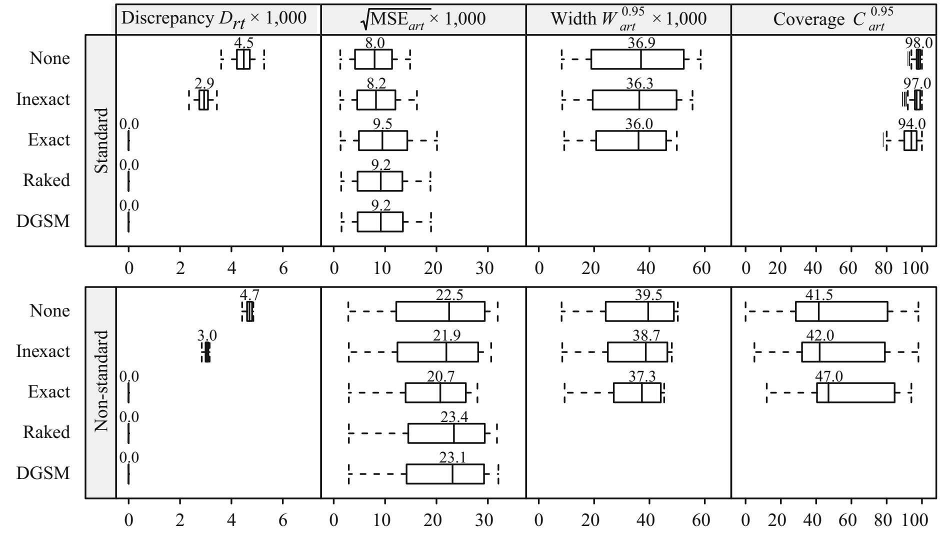


Fig. 3. Performance of model of fertility rates when applied to 'change in distribution' data sets, distinguishing between standard and nonstandard areas. The results are based on K = 100 replicates.

At each replicate of the simulation, we construct a simulated survey data set by drawing a sample of 60,000 records from the file. The sample is stratified by region, with simple random sampling within each region. There are 16 regions in total, with populations ranging from 23,000 to just over 1 million. Regional sample sizes are proportional to the square root of regional population size.

Let N_{aslr} be the number of people in age-group $a \in \{15\text{–}24, 25\text{–}34, \ldots, 55\text{–}64, 65+\}$, sex s, income level $l \in$ {No income or loss, NZD1–NZD20,000, NZD20,001–NZD40,000, NZD40,001–NZD60,000, NZD60,001–NZD100,000, NZD100,001+}, and region $r \in \{1, \ldots, 16\}$. Let Y_{aslr} be the number of people who smoke. We treat the census file as the true finite population. Within the simulated analysis, N_{aslr} is known but Y_{aslr} is not. The aim of the simulated analysis is to estimate finite-population smoking prevalence by age, sex, and income, $p_{asl} = \sum_r Y_{aslr} / \sum_r N_{aslr}$.

Let y_{aslr} and n_{aslr} be the sample equivalents of Y_{aslr} and N_{aslr}. The model

$$y_{aslr} \overset{\text{ind}}{\sim} \text{Binomial}(n_{aslr}, \gamma_{aslr}) \tag{44}$$

$$\text{logit}(\gamma_{aslr}) \overset{\text{ind}}{\sim} \text{N}\left(\beta^0 + \beta_a^{\text{age}} + \beta_s^{\text{age}} + \beta_l^{\text{income}} + \beta_r^{\text{reg}} + \beta_{al}^{\text{age:income}}, \sigma^2\right) \tag{45}$$

is fitted to the artificial survey data. Region is included in the model to account for the stratified sample design. The age effects, income effects, region effects, and age-income interaction all have normal priors with mean 0. However, following the approach that Little (2011) suggests for statistical agencies that are reluctant to adopt informative priors, we use improper uniform priors over the set of positive real numbers for the standard deviation terms. We use an improper uniform prior for the sex effect.

The model is fitted without benchmarking, and with exact and inexact benchmarking. The benchmarks are estimated mean smoking prevalence by income level

$$m_l = \frac{\sum_{a,s,r} N_{aslr} y_{aslr} / n_{aslr}}{\sum_{a,s,r} N_{aslr}}.$$

The corresponding super-population benchmarking parameters are

$$\psi_l = \sum_{a,s,r} N_{aslr} \gamma_{aslr} / \sum_{a,s,r} N_{aslr}.$$

The finite-population equivalent of ψ_l is $\psi_l^{\text{fin}} = \sum_{a,s,r} Y_{aslr} / \sum_{a,s,r} N_{aslr}$.

Under exact benchmarking, $p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi})$ is given by Equation (5). Under inexact benchmarking, $p^{[\mathbf{m} \mid \boldsymbol{\psi}]}(\mathbf{m} \mid \boldsymbol{\psi})$ is given by Equation (8), which becomes

$$p^{[\boldsymbol{m} \mid \boldsymbol{\psi}]}(\boldsymbol{m} \mid \boldsymbol{\psi}) \propto \exp\left(-\sum_l \frac{(m_l - \psi_l)^2}{2\lambda s_l^2}\right), \tag{46}$$

where s_l is the standard error of using m_l to estimate ψ_l. We examine the cases where $\lambda = 1$ and where $\lambda = 0.5$. The $\lambda = 1$ case allows discrepancies between model-based estimates and benchmarks to vary in line with sampling variation, while the $\lambda = 0.5$ case allows smaller discrepancies.

Benchmarks m_l and standard errors s_l are calculated using function `svymean` from *R* package `survey` (Lumley 2004). Calculating standard errors that properly account for the stratified sample design is complicated; function `svymean` uses replicate weights (Lumley 2011, 32).

Performance measures D_l, MSE_{as1}, W^q_{asl}, and C^q_{asl} are calculated for finite-population smoking prevalence p_{asl}. Here D_l is defined as

$$D_l = \mathrm{E}\left(\frac{|\,\bar{\psi}_l^{\text{fin}} - m_l\,|}{s_l}\right), \tag{47}$$

where $\bar{\psi}_l^{\text{fin}}$ is the posterior mean of ψ_l^{fin}. This measures discrepancies in units of standard errors. To estimate p_{asl} and ψ_l^{fin}, it is necessary to estimate $Y_{asl} = \sum_r (y_{aslr} + y_{aslr}^{\text{non}})$, where y_{aslr}^{non} is the number of non-sampled people in area *aslr* who smoke. Draws from the posterior distribution of y_{aslr}^{non} can be generated using $y_{aslr}^{\text{non}(t)} \sim \text{Binomial}(N_{aslr} - n_{aslr}, \gamma_{aslr}^{(t)})$, where $\gamma_{aslr}^{(t)}$ is the *t*th draw from the posterior sample for γ_{aslr}.

As with the fertility simulation, we use $K = 100$ replicates. The Gibbs sampler is run with six independent chains, each with 100,000 iterations. Every 250th draw from the final 50,000 iterations of each chain is recorded, yielding a combined total of 1,200 draws from the posterior distribution.

The results from the simulation are summarized in Figure 4, with $q = 0.95$. Benchmarking improves agreement between the model-based estimates and the benchmarks, with exact benchmarking giving the largest improvement, and inexact benchmarking with $\lambda = 1$ the smallest. Exact benchmarking does not achieve complete agreement, since the benchmarks are applied to super-population prevalences, rather than finite population ones. However, the median absolute difference between model-based estimates and benchmarks is only 0.01 standard errors.

Benchmarking degrades overall accuracy and coverage. However, the most striking feature of the distributions of MSE_{asl}, $W^{0.95}_{asl}$, and $C^{0.95}_{asl}$ in Figure 4 is the long tails. These long tails result from a small number of outliers, notably people aged 15–24 with incomes of NZD100,000 or higher. This group has high smoking prevalence, even though youth and high incomes are, in general, associated with low prevalence. With these particular data, rather than providing robustness to outliers, benchmarking decreases robustness.

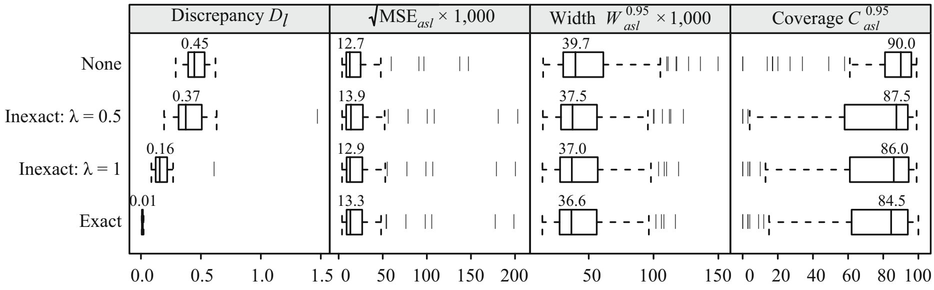


Fig. 4. Performance of model of smoking prevalence, by type of benchmarking. The results are based on K = 100 replicates.

5. Application

We now apply fully Bayesian benchmarking to a real data set. We estimate age-sex-specific mortality rates for local authority districts in England and Wales, using as benchmarks sex-specific life expectancies for regions. All code and data for the application can be obtained from github.com/bayesiandemography/britmort.

Our data are counts of deaths and populations at risk in 2014, disaggregated into 20 age groups, 2 sexes, and 348 local authority districts. The total number of deaths is 500,314, and the total population at risk is 57,408,654. The median number of deaths per area is 8, and 16% of areas have 0 deaths. Direct estimates of mortality rates (i.e., counts of deaths for each area, divided by the corresponding population at risk) for five randomly-selected districts are shown in Figure 5. Because the graphs are drawn on a log scale, estimates for which the count of deaths and direct estimate are 0 are omitted. As is apparent from the graphs, direct estimates of age-sex-specific mortality rates at the district level are unstable below age 60.

Let y_{asd} be the count of deaths for age group a, sex s and district d, and let γ_{asd} and w_{asd} be the corresponding mortality rate and population at risk. We apply the model

$$y_{asd} \overset{\text{ind}}{\sim} \text{Poisson}(w_{asd}\gamma_{asd}) \tag{48}$$

$$\log\ \gamma_{asd} \overset{\text{ind}}{\sim} \text{N}(\beta^0 + \beta_a^{\text{age}} + \beta_s^{\text{sex}} + \beta_d^{\text{dis}} + \beta_{as}^{\text{age:sex}}, \sigma^2). \tag{49}$$

Age effects are assumed to follow a random walk with drift,

$$\beta_a^{\text{age}} \sim t_4(\eta_a^{\text{age}}, \tau_{\text{age}}^2) \tag{50}$$

$$\eta_0^{\text{age}} \sim \text{N}(0, 10^2) \tag{51}$$

$$\eta_a^{\text{age}} \sim \text{N}(\eta_{a-1}^{\text{age}} + \delta_{a-1}^{\text{age}}, \omega^2), \quad a > 0 \tag{52}$$

$$\delta_0^{\text{age}} \sim \text{N}(0, 1) \tag{53}$$

$$\delta_a^{\text{age}} \sim \text{N}(\delta_{a-1}^{\text{age}}, \varphi^2), \quad a > 0. \tag{54}$$

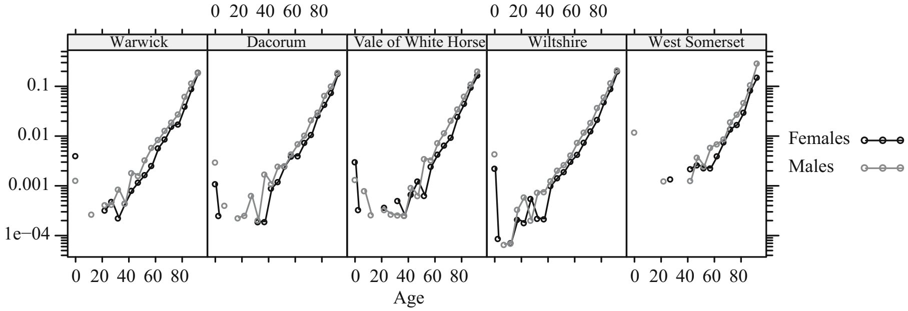


Fig. 5. Direct estimates of mortality rates, by age and sex, in five randomly-selected local authority districts in England and Wales, 2014. For some combinations of age, sex, and district, the counts of deaths, and hence direct estimates, are 0. Estimates for these combinations are not shown, since the graph is on a log scale.

The drift term accounts for the fact that log-mortality rates rise linearly over much of the age range. The use of a t_4 distribution in Equation (50) allows for occasional large departures from trend, as occurs at age 0. The sex effect has a normal prior $\beta_s^{\text{sex}} \sim \text{N}(0, 1)$. The district effect has a normal prior $\beta_d^{\text{dis}} \sim \text{N}(0, \tau_{\text{dis}}^2)$, with a weakly informative half-*t* prior on the standard deviation, $\tau_{\text{dis}} \sim t_7^+(0, 1)$. The interaction has a normal prior $\beta_{as}^{\text{age:sex}} \sim \text{N}(0, \tau_{\text{age:sex}}^2)$, with a weakly informative half-*t* prior on the standard deviation, $\tau_{\text{age:sex}} \sim t_7^+(0, 0.5^2)$. We use a smaller scale for the interaction than for the main effect on the principle that interactions are typically smaller in size than main effects (Gelman et al. 2008). Standard deviations terms σ, τ_{age}, ω and φ all have $t_7^+(0, 1)$ priors.

We benchmark the estimates to sex-specific life expectancies for regions. The region is an administrative unit further up the English geographical hierarchy than the local authority district. Counting Wales as one region, there were ten regions in England and Wales in 2014. Life expectancy is the mean number of years a newborn baby would live if prevailing mortality rates were to continue indefinitely. The procedure for calculating life expectancy is given in Preston et al. (2001, chap. 3), but for our purposes, the key point is that life expectancy is a nonlinear deterministic function of age-specific mortality rates.

Let

$$\zeta_{asr} = \sum_{d \in \Delta_r} w_{asd} \gamma_{asd} \Big/ \sum_{d \in \Delta_r} w_{asd} \tag{55}$$

be the mortality rate in age group a, sex s and region r, where Δ_r is the set of d such that district d belongs to region r. Life expectancy for sex s in region r is

$$\psi_{sr} = f_{\text{life}}(\zeta_{1sr}, \ldots, \zeta_{Asr}), \tag{56}$$

where f_{life} is the nonlinear function for calculating life expectancy from age-specific mortality rates, and $A = 20$ is the number of age groups. Similarly, let

$$z_{asr} = \sum_{d \in \Delta_r} y_{asd} \Big/ \sum_{d \in \Delta_r} w_{asd} \tag{57}$$

be the direct estimate of the mortality rate in age group a, sex s, and region r. The benchmark for sex s and region r is then

$$m_{sr} = f_{\text{life}}(z_{1sr}, \ldots, z_{Asr}). \tag{58}$$

Since life expectancies are ordinarily reported to at most two decimal places, most users can tolerate discrepancy of size 0.01. We specify agreement with the benchmarks as

$$m_{sr} \overset{\text{ind}}{\sim} \text{N}(\psi_{sr}, 0.005^2). \tag{59}$$

We fit our model with and without benchmarks, using four independent chains, each with 80,000 iterations. Every 100th draw from the final 40,000 iterations of each chain is recorded, yielding a combined total of 1,600 draws from the posterior distribution.

Benchmarking improves agreement between the modelled life expectancies by sex and region and the benchmarks. Figure 6 compares benchmarks with point estimates

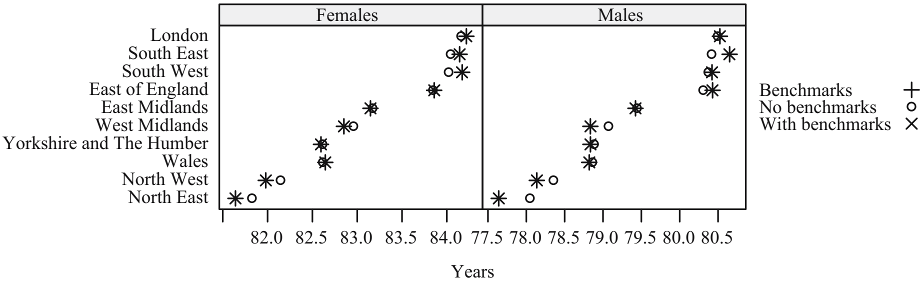
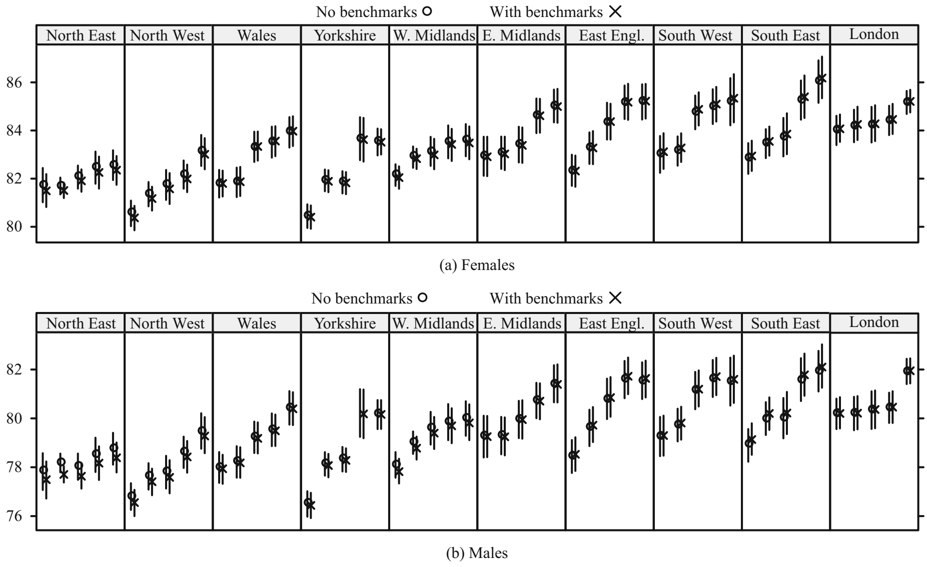


Fig. 6. Point estimates of life expectancy by sex and region: benchmarks versus posterior medians from models.

(posterior medians) from models with and without benchmarking. Without benchmarking, the model-based estimates are noticeably different from the benchmarks, especially for males. With benchmarking, the model-based estimates and benchmarks are indistinguishable.

Figure 7 shows life expectancies by sex and district, with and without benchmarking. Benchmarking shifts most posterior medians. The shifts are larger in some regions, such as the North East, than in others, such as London. Benchmarking changes the width of credible intervals, but only very slightly. The mean width of credible intervals for all age-sex-district-specific log-mortality rates increases from 0.33 to 0.34, and the mean width of credible intervals for sex-district-specific life expectancies decreases from 1.33 to 1.32 (results not shown).

Figure 8 illustrates how benchmarking affects age-sex-specific mortality rates at the district level. The percent differences between posterior medians of mortality rates from benchmarked models and those from non-benchmarked models are all below 4%.



(a) Females

(b) Males

Fig. 7. Life expectancies for 50 local authority districts, with five randomly-selected districts fromeach region. The vertical lines are 95% credible intervals.

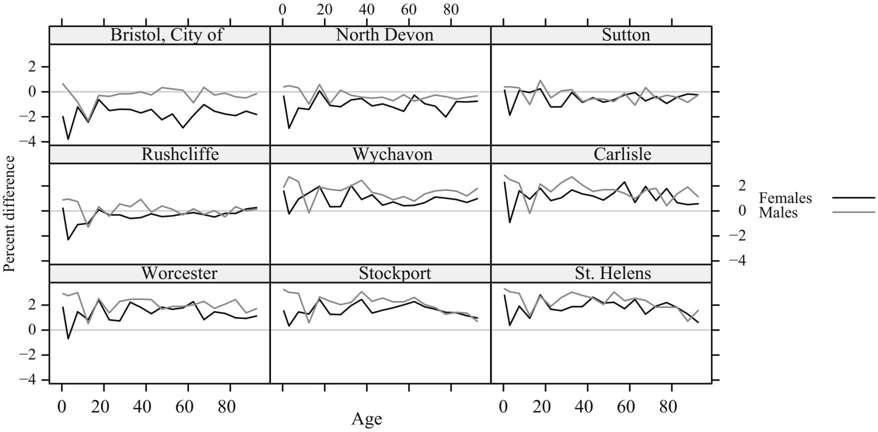


Fig. 8. Percent differences between posterior medians from benchmarked models and posterior medians from non-benchmarked models, for age-sex-specic mortality rates in nine randomly-selected local authority districts.

6. Discussion

We conclude by summarizing the advantages of the fully Bayesian benchmarking methods described in this article.

Our benchmarking methods allow full posterior distributions to be generated for a wide range of models. With full posterior distributions, uncertainty measures can be automatically produced for all unknown quantities in the small area models, as well as derived quantities, such as the finite-population smoking prevalence (ψ_l^{fin}) in the smoking simulation in Subsection 4.2, or life expectancy (ψ_{sr}) in the mortality application in Section 5.

With some applications, it is not necessary to obtain complete agreement between benchmarks and model-based estimates. Policy analysts, for instance, may have weaker requirements for agreement among estimates than administrators. In such cases, inexact benchmarking, using the methods described in this article, gives statisticians the ability to control the level of agreement with the benchmarks. As illustrated by the fertility and smoking simulations, using inexact benchmarking can achieve smaller mean squared errors than using exact benchmarking.

Finally, in some applications, the most natural benchmarks are quantities that have a non-linear relationship with the small area parameters, which can be accommodated under our approach. In this article, we consider the case of life expectancies, but other non-linear benchmarks such as growth rates and ratios can also be implemented using our methods.

7. References

Albert, I., S. Donnet, C. Guihenneuc-Jouyaux, S. Low-Choy, K. Mengersen, and J. Rousseau. 2012. “Combining expert opinions in prior elicitation.” *Bayesian Analysis* 7(3): 503–532. DOI: https://doi.org/10.1214/12-BA717.

Bell, W.R., G.S. Datta, and M. Ghosh. 2013. “Benchmarking small area estimators.” *Biometrika* 100(1): 189–202. DOI: https://doi.org/10.1093/biomet/ass063.

Berg, E. and W. Fuller. 2009. "A SPREE Small Area Procedure for Estimating Population Counts". In *Proceedings of the Survey Methods Section, Statistical Society of Canada*. Section on Survey Methods, Statistical Society of Canada. Available at: http://www.ssc.ca/survey/documents/SSC2009_EBerg.pdf (accessed August 2019).

Berg, E.J., W.A. Fuller, and A.L. Erciulescu. 2012. "Benchmarked small area prediction." In *Proceedings of the Section on Research Methods, Joint Statistical Meeting*. Section on Research Methods, Joint Statistical Meeting. Available at: http://www.asasrms.org/Proceedings/y2012/Files/305110_74288.pdf (accessed August 2019).

Datta, G.S., M. Ghosh, R. Steorts, and J. Maples. 2011. "Bayesian benchmarking with applications to small area estimation." *TEST* 20(3): 574–588. DOI: https://doi.org/10.1007/s11749-010-0218-y.

De Waal, T. 2016. "Obtaining numerically consistent estimates from a mix of administrative data and surveys." *Statistical Journal of the IAOS* 32(2): 231–243. DOI: https://doi.org/10.3233/SJI-150950.

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro-level estimation of poverty and inequality." *Econometrica* 71(1): 355–364. DOI: https://doi.org/10.1111/1468-0262.00399.

Fabrizi, E., C. Giusti, N. Salvati, and N. Tzavidis. 2014. "Mapping average equivalized income using robust small area methods." *Papers in Regional Science* 93: 685–701. DOI: https://doi.org/10.1111/pirs.12015.

Fabrizi, E., N. Salvati, and M. Pratesi. 2012. "Constrained small area estimators based on M-quantile methods." *Journal of Official Statistics* 28(1): 89–106. Available at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/constrained-small-area-estimators-based-on-m-quantile-methods.pdf (accessed August 2019).

Fay, R.E. and R.A. Herriot. 1979. "Estimates of income from small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association* 74: 269–277. DOI: https://doi.org/10.1080/01621459.1979.10482505.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. 2014. *Bayesian Data Analysis*, Third Edition. New York: Chapman and Hall.

Gelman, A., A. Jakulin, M.G. Pittau, and Y.-S. Su. 2008. "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics* 2: 1360–1383. DOI: https://doi.org/10.1214/08-AOAS191.

Ghosh, M., T. Kubokawa, and Y. Kawakubo. 2015. "Benchmarked empirical Bayes methods in multiplicative area-level models with risk evaluation." *Biometrika* 102(3): 647–659. DOI: https://doi.org/10.1093/biomet/asv010.

Lindley, D.V. 1983. "Reconciliation of Probability Distributions." *Operations Research* 31: 866–880. DOI: https://doi.org/10.1287/opre.31.5.866.

Lindley, D.V., A. Tversky, and R.V. Brown. 1979. "On the Reconciliation of Probability Assessments (with discussion)." *Journal of the Royal Statistical Society, Series A* 142: 146–180. DOI: https://doi.org/10.2307/2345078.

Little, R.J. 2012. "Calibrated Bayes, an alternative inferential paradigm for official statistics." *Journal of Official Statistics* 28(3): 309. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/calibrated-bayes-an-alternative-inferential-paradigm-for-official-statistics.pdf (accessed August 2019).

Lumley, T. 2004. “Analysis of Complex Survey Samples.” *Journal of Statistical Software* 9(1): 1–19. DOI: https://doi.org/10.18637/jss.v009.i08.

Lumley, T. 2011. *Complex surveys: A guide to analysis using R,* Volume 565. John Wiley & Sons.

Morris, P.A. 1974. “Decision Analysis Expert Use.” *Management Science* 20: 1233–1241. DOI: https://doi.org/10.1287/mnsc.20.9.1233.

Morris, P.A. 1977. “Combining Expert Judgements: A Bayesian Approach.” *Management Science* 23: 679–693. DOI: https://doi.org/10.1287/mnsc.23.7.679.

Nandram, B. and H. Sayit. 2011. “A Bayesian analysis of small area probabilities under a constraint.” *Survey Methodology* 37: 137–152. Available at: www.150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11603-eng.pdf (accessed August 2019).

Nandram, B., M.C.S. Toto, and J.W. Choi. 2011. “A Bayesian benchmarking of the Scott-Smith model for small areas.” *Journal of Statistical Computation and Simulation* 81(11): 1593–1608. DOI: https://doi.org/10.1080/00949655.2010.496726.

O’Hagan, A., C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkenson, J.E. Oakley, and T. Rakow. 2006. *Eliciting Experts’ Probabilities*. John Wiley and Sons, Ltd.

Pfeffermann, D. 2013. “New important developments in small area estimation.” *Statistical Science* 28(1): 40–68. DOI: https://doi.org/10.1214/12-STS395.

Pfeffermann, D. and C.H. Barnard. 1991. “Some new estimators for small-area means with application to the assessment of farmland values.” *Journal of Business & Economic Statistics* 9(1): 73–84. DOI: https://doi.org/10.1080/07350015.1991.10509828.

Pfeffermann, D., A. Sikov, and R. Tiller. 2014. “Single-and two-stage cross-sectional and time series benchmarking procedures for small area estimation.” *TEST* 23(4): 631–666. DOI: https://doi.org/10.1007/s11749-014-0400-8.

Pfeffermann, D. and R. Tiller. 2006. “Small-area estimation with state-space models subject to benchmark constraints.” *Journal of the American Statistical Association* 101(476): 1387–1397. DOI: https://doi.org/10.1198/016214506000000591.

Poole, D. and A.E. Raftery. 2000. “Inference for deterministic simulation models: The Bayesian melding approach.” *Journal of the American Statistical Association* 95: 1244–1255. DOI: https://doi.org/10.1080/01621459.2000.10474324.

Prado, R. and M. West. 2010. *Time series: modeling, computation, and inference*. CRC Press.

Preston, S., P. Heuveline, and M. Guillot. 2001. *Demography: Modelling and Measuring Population Processes*. Oxford: Blackwell.

Ranalli, M.G., G.E. Montanari, and C. Vicarelli. 2018. “Estimation of small area counts with the benchmarking property.” *Metron* 76(3): 349–378. DOI: https://doi.org/10.1007/s40300-018-0146-2.

Rao, J.N.K. and I. Molina. 2015. *Small area estimation*, Second edition. John Wiley & Sons.

Roback, P.J. and G.H. Givens. 2001. “Supra-Bayesian pooling of priors linked by a deterministic simulation model.” *Communications in Statistics-Simulation and Computation* 30(3): 447–476. DOI: https://doi.org/10.1081/SAC-100105073.

Steorts, R.C. and M. Ghosh. 2013. "On estimation of mean squared errors of benchmarked empirical Bayes estimators." *Statistica Sinica* 23(2): 749–767. DOI: https://doi.org/10.5705/ss.2012.053.

Toto, M.C.S. and B. Nandram. 2010. "A Bayesian predictive inference for small area means incorporating covariates and sampling weights." *Journal of Statistical Planning and Inference* 140(11): 2963–2979. DOI: https://doi.org/10.1016/j.jspi.2010.03.043.

U.S. Census Bureau. 2014. "Model-based Small Area Income and Poverty Estimates (SAIPE) for School Districts, Counties, and States" (accessed July 2014).

Vesper, A.J. 2013. *Three Essays of Applied Bayesian Modeling: Financial Return Contagion, Benchmarking Small Area Estimates, and Time-Varying Dependence*. PhD thesis, Harvard University. Available at: https://dash.harvard.edu/handle/1/11124829 (accessed August 2019).

Wang, J., W.A. Fuller, and Y. Qu. 2008. "Small area estimation under a restriction." *Survey Methodology* 34(1): 29. Available at: www150.statcan.gc.ca/n1/pub/12-001-x/2008001/article/10619-eng.pdf (accessed August 2019).

You, Y. and J. Rao. 2002. "A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights." *Canadian Journal of Statistics* 30(3): 431–439. DOI: https://doi.org/10.2307/3316146.

You, Y. and J. Rao. 2003. "Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights." *Journal of Statistical Planning and Inference* 111: 197–208. DOI: https://doi.org/10.1016/S0378-3758(02)00301-4.

You, Y., J. Rao, and P. Dick. 2004. "Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation." *Statistics in Transition* 6: 631–640. Available at: https://pts.stat.gov.pl/en/journals/statistics-in-transition/ (accessed February 2020).

You, Y., J. Rao, and M. Hidiroglou. 2013. "On the performance of self benchmarked small area estimators under the Fay-Herriot area level model." *Survey Methodology* 39: 217–230. Available at: www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11830-eng.htm (accessed August 2019).

Received November 2018
Revised May 2019
Accepted September 2019

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 225–228, http://dx.doi.org/10.2478/JOS-2020-0011

Book Review

Debra R. Miller[1]

Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop, and Brita Dorer, eds. *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts*. Hoboken, NJ: Wiley, 2019. ISBN 978-1-118-88498-0, 1104pps.

This book grew out of a 2016 multinational, multiregional, and multicultural contexts (3MC) conference held in Chicago. The book follows in the steps of Harkness et al. (2003) as well as Harkness et al. (2010). It is divided into 12 sections and 48 chapters. In addition to an introduction, the sections include sampling approaches; cross-cultural questionnaire design and testing; languages, translation, and adaptation; mixed mode and mixed methods; response styles; data collection challenges and approaches; quality control and monitoring; nonresponse; multi-group analysis; harmonization, data documentation, and dissemination; and looking forward.

The book addresses an audience of survey researchers who collect and compare data across nations, regions, and cultures. The book describes what 3MC researchers are currently doing and guides others who are doing similar work. For 11 of the 12 sections, the focus is on designing and implementing various stages of 3MC surveys. The final data harmonization section is on recycling data from existing studies.

The introductory section presents the promise of 3MC research as the ability to treat surveys as a unit of analysis based on the growing number of national-level surveys. Importantly, the section expands the total survey error paradigm to include comparison error for each model component. The section also promotes mixed methods research as a validity framework.

The sampling approaches section includes chapters on GIS technology (for studies where census data is not current) and within-household respondent selection (with an eye toward future methods). The section discusses cluster and household selection (e.g., via satellite photos) for readers interested in innovatively using GIS technology though the discussion could serve monocultural researchers as well. The section also suggests that cross-national surveys could serve to implement research on selection techniques for face-to-face surveys.

The cross-cultural questionnaire design and testing section discusses questionnaire design for comparative researchers in 3MC survey contexts. The section particularly discusses population-specific sensitive questions, anchoring vignettes to avoid culturally-based differential item functioning, differences in communication styles, bilingual cognitive testing, and the comparability of behavior coding.

[1] Integrated Methodology and Mentoring, 3220 Apple Street Apt 13, Lincoln, NE 68503 U.S.A. Email: millerdebra35@gmail.com

The section on languages, translation, and adaptation initially discusses how to optimize interview language choice across preparatory (researcher and field manager) and field (interviewer and respondent) stages relative to group and individual level multilingualism. The section suggests that administration language influences respondents' answers, especially for bilingual respondents. One chapter addresses readers who "collect, organize, and make good use of" documentation (Behr et al. 2019, 341) and provides information on the input and output of translation documentation. The section ends by comparing translation measurement properties through a coding scheme based on the Survey Quality Predictor software.

The mixed modes and mixed methods section distinguishes across- and within-country mixed mode designs for reducing coverage and non-response error or to reduce cost. The section continues by discussing mixed mode adjustment methods including logistic regression, propensity score matching, calibration to predetermined values, and a multiple imputation method. The section also discusses technological opportunities for expanding mixed methods by distinguishing designed from organic data.

The section on response styles starts by questioning the cross-cultural comparability of subjective probability response patterns. It then discusses attitudinal rating response styles across cultures, particularly acquiescent and extreme response styles. One recommendation is testing for measurement invariance. The section ends by discussing translation of balanced or unbalanced response scales with reference to five surveys in four countries and by cautioning against naïvely comparing data from different countries and languages.

The section on challenges and approaches to data collection emphasizes dimensions of social and cultural context, political context, economic conditions and infrastructure, physical environment, and research traditions and experience. The chapter on sub-Saharan Africa discusses challenges with not knowing what proportion of a population is nomadic, language multiplicity, insecurity, and limited resources. It suggests that the increase in communication hardware and software could mitigate data deprivation. The chapter on the Middle East and Arab Gulf focuses primarily on Qatar and recommends anchoring vignettes to address the limitation of directly-posed questions. The chapter on Latin America and the Caribbean discusses five comparative survey projects and cites challenges of increasing violence and lack of current census information. The chapter on India and China suggests options for moving beyond the dominant face-to-face mode.

The data collection section suggests study branding, collecting contact information, between-wave contact, and customized approaches to respondents. The chapter on multinational event history calendar interviewing promotes a standardized approach, providing feedback to interviewers as soon as possible, adjusting the order of domains and adapting interviewer training according to cultural differences, providing an additional device on which respondents can see the calendar, and using paradata as a means to make behavior coding more efficient. The chapter on the collection of biological samples in a survey discusses but does not address differing cultural acceptability of obtaining blood samples. A theoretical chapter discusses quality relative to principles of respect for persons, beneficence, and justice. The section ends by discussing the General Data Protection Regulation's promise to harmonize legislation and practice.

The quality control and monitoring section pays particular attention to survey quality as pertaining to all survey life cycle phases. Two of the section's five case studies are set in India, one in the Kingdom of Saudi Arabia, and two in Europe. The section emphasizes that firms as well as interviewers have fabricated data and recommends sharing information concerning fabrication as a way of deterring firms from the act.

The chapters in the nonresponse section provide a discussion of the sources of nonresponse bias and how this may differ across cultures. The section lays out several strategies for minimizing nonresponse across countries such as attending to characteristics of sampling frames and gathering paradata. A nonresponse model for within-household cooperation in the California Health Interview Survey desirably provides a cultural ecosystems framework but is unfortunately based on phone numbers as a proxy of community. The section ends by explaining how nonresponse arises when non-national language speakers lack the option of answering in their native language.

For analysts or data users, the multi-group confirmatory factor analysis section addresses exact and partial measurement invariance as well as approximate measurement invariance with Bayesian priors in Mplus.

The section on harmonization, data documentation, and dissemination introduces a database of "22 large international survey projects encompassing 1721 individual surveys from 142 countries" with nearly 2.3 million individuals (Granda 2019, 933). The section defines survey data recycling as "a framework for integrating information from extant survey and nonsurvey sources to create multicountry multiyear data sets" (Slomczynski and Tomescu-Dubrow 2019, 937), drawing on total survey error, total survey quality, and total quality management. One chapter identifies types of data processing errors including illegitimate variables values, misleading variable values, contradictory variable values, variable values discrepancy, and lack of variable value labels. The section proposes classifying item metadata as new variables relevant for assessing "intersurvey reliability and validity of variables created via *ex post* harmonization of survey data" (Kołczyńska and Slomczynski 2019, 1027). Ultimately, the section discusses the quality of existing design and post-stratification weights, as well as the advantages of recalculating weights.

The final section on looking forward presents a concern that covering too many populations or countries fails to allow properly handling processes. The section summarizes prevailing problems of different research camps or traditions; the cost of 3MC surveys; need for strong central leadership; need to strengthen user roles for policy impact; and issues related to data quality such as overly specialized methodologists.

The book's balance of qualitative and quantitative chapters accentuate mitigating total survey error to enhance quality. Notably, "the overarching goal of 3MC surveys is to minimize comparison error" (Scott et al. 2019, 719). Most chapters describe what 3MC researchers are currently doing and offer practical suggestions for designing surveys to mitigate error components. Furthermore, several chapters offer cautions for interpreting existing data sets. Readers who conduct 3MC surveys, use 3MC survey data, or who want to expand their awareness of quality assurance and control in such contexts will benefit from reading the book.

References

Behr, D., S. Dept, and E. Krajčeva. 2019. “Documenting the survey translation and monitoring process.” In *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts*, edited by T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, and B. Dorer, 341–356. Hoboken, NJ: Wiley.

Granda, P. 2019. “Data harmonization, data documentation, and dissemination.” In *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts*, edited by T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, and B. Dorer, 933–936. Hoboken, NJ: Wiley.

Harkness, J.A., F.J.R. van de Vijver, and P.Ph. Mohler. 2003. *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

Harkness, J.A., M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.Ph. Mohler, B.-E. Pennell, and T.W. Smith. 2010. *Survey methods in multinational, multiregional, and multicultural contexts*. Hoboken, NJ: Wiley.

Kołczyńska, M. and K.M. Slomczynski. 2019. “Item metadata as controls for *ex post* harmonization of international survey projects.” In *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts*, edited by T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, and B. Dorer, 1011–1034. Hoboken, NJ: Wiley.

Scott, L., P.Ph. Mohler, and K. Cibelli Hibben. 2019. “Organizing and managing comparative surveys.” In *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts*, edited by T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, and B. Dorer, 707–730. Hoboken, NJ: Wiley.

Slomczynski, K.M. and I. Tomescu-Dubrow. 2019. “Basic principles of survey data recycling.” In *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts*, edited by T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, and B. Dorer, 937–962. Hoboken, NJ: Wiley.