Journal of Official Statistics
vol. 37, 4 (December 2021)

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 791–809, http://dx.doi.org/10.2478/JOS-2021-0035

Freedom of Information and Personal Confidentiality in Spatial COVID-19 Data

Michael Beenstock[1] *and Daniel Felsenstein*[2]

We draw attention to how, in the name of protecting the confidentiality of personal data, national statistical agencies have limited public access to spatial data on COVID-19. We also draw attention to large disparities in the way that access has been limited. In doing so, we distinguish between absolute confidentiality in which the probability of detection is 1, relative confidentiality where this probability is less than 1, and collective confidentiality, which refers to the probability of detection of at least one person. In spatial data, the probability of personal detection is less than 1, and the probability of collective detection varies directly with this probability and COVID-19 morbidity. Statistical agencies have been concerned with relative and collective confidentiality, which they implement using the techniques of truncation, where spatial data are not made public for zones with small populations, and censoring, where exact data are not made public for zones where morbidity is small.

Granular spatial data are essential for epidemiological research into COVID-19. We argue that in their reluctance to make these data available to the public, data security officers (DSO) have unreasonably prioritized data protection over freedom of information. We also argue that by attaching importance to relative and collective confidentiality, they have over-indulged in data truncation and censoring. We highlight the need for legislation concerning relative and collective confidentiality, and regulation of DSO practices regarding data truncation and censoring.

Key words: Spatial COVID-19 data; relative confidentiality; collective confidentiality; data censoring; data truncation.

1. Introduction

As in so many areas, the COVID-19 pandemic has imposed a sea-change on government statistical agencies. In their quest to track, contain and forecast the spread of the virus, governments have been forced to address new data governance and privacy challenges (OECD 2020a). While many of these are related to the nature of digital data sources such as mobile phone data and biometrics (Newlands et al. 2020), demands are also being made on more traditional statistical sources such as censuses, household and income surveys, and tax data. In the case of COVID-19 data, these demands call for a far from perfect trade-off between data accessibility and freedom of information for containing the pandemic on the one hand and issues of personal confidentiality on the other (OECD 2020b).

[1] Department of Economics, Hebrew University of Jerusalem, Mt Scopus, Jerusalem 91900 Israel. Email: michael.beenstock@mail.huji.ac.il
[2] Department of Geography, Hebrew University of Jerusalem, Mt Scopus, Jerusalem 91900 Israel. Email: daniel.felsenstein@mail.huji.ac.il
Acknowledgments: Many thanks to Tal Zarsky, Guy Zomer and the reviewers for their comments.

In this article, we address this trade-off in the context of spatial COVID-19 data. Since the outbreak of the COVID-19 pandemic, national statistical agencies (NSAs) have been inundated with data requests from scientific investigators, the media, and organizations concerned with public health. Much of this demand has been for spatial data, in which information on individuals is aggregated into territorial units or zones of differing levels of resolution. As both the transmission of COVID-19 and the policy response to its spread are inherently spatial (Poom et al. 2020), government agencies are increasingly requested to supply data to track and analyze the spatiotemporal dynamics of the pandemic. Given this brief, statistical agencies find themselves caught between the hammer of freedom of information and the anvil of protecting individual confidentiality.

Because spatial data are aggregated into zones, this tension should ostensibly be mediated. Reporting, for example, the number of people infected in a zone does not reveal the identity of individuals. We argue, however, that statistical agencies have confounded the absolute confidentiality of personal information, which is the objective of existing legislation, with broader concepts of confidentiality not covered by existing legislation. These include relative confidentiality, which is concerned with the probability of identification faced by each individual, and collective confidentiality, which is concerned with the probability faced by statistical agencies that at least one individual will be identified. These concepts are developed further below.

We argue that these broader concepts of confidentiality have been applied by statistical agencies, such as ministries of health, to limit public access to spatial COVID-19 data. Since these broader concepts are not covered by existing legislation, freedom of information has been unnecessarily infringed.

We draw specific attention to spatial data for morbidity, hospitalizations and mortality in zones such as cities, towns, administrative districts, neighborhoods, census tracts and postal zip codes. These zones are particularly important for mitigation policy and research because COVID-19 is contagious, and its transmission is fundamentally spatial. They are also important more generally, because the public has the right to know for their own safety where the disease is particularly severe. Spatial data are also required for non-infectious diseases for which environmental factors matter.

In summary, the unit of observation, which we study, is not the individual, but rather the number of individuals in zones with COVID-19 related outcomes at or during a given time period. For example, the number of people ever diagnosed with COVID-19 as of January 1, 2021, or the number of new cases diagnosed during the week ending on January 1. These are time series data that are typically updated daily or weekly, and are the spatial counterparts to national data for COVID-19 outcomes, which have featured continuously in the media since the outbreak of the pandemic.

We challenge current practice of NSAs in their response to the release of spatial COVID-19 data in three respects. First, we claim that they confuse absolute and relative confidentiality when dealing with spatial data resulting in excessive data protection where it is not mandated. Second, we challenge the response of NSAs to data protection through the practices of truncation and censoring of spatial COVID-19 data. Truncation arises when data for zones with small populations are not made public. Censoring occurs when morbidity data are grouped, for example, morbidity during the last week is a number between 1 and 14. Whereas truncation conceals all the data, censoring reveals part of the

data. We claim that in the case of COVID-19 data, truncation is applied heavy handedly, while censoring is generally unjustified. Third, we suggest that in reference to spatial COVID-19 data, NSAs have confounded individual and collective confidentiality. Recall that relative confidentiality refers to individuals, and collective confidentiality refers to statistical organizations.

The article proceeds as follows. Section 2 addresses the unique nature of spatial data and emphasizes that personal confidentiality in such data is relative (probabilistic) and not absolute. The concepts of relative and collective confidentiality are explained in Section 3, and their relationships to data truncation and censoring are elucidated. We show that while truncation may be justified under certain conditions relating to collective data protection, censoring has no obvious rationale. A review of spatial COVID-19 data availability in several countries is provided in Section 4. It highlights the various data limitations applied by NSAs and underscores the very different national contexts within which data are made public. In Section 5, we question the legal justification for attaching importance to relative and collective data protection. Section 6 summarizes and concludes.

Although we are concerned with universal issues in data protection policy and we provide a review of international practice, we highlight the case of Israel to illustrate our arguments. We are naturally more familiar with the intricacies of data protection where we live. However, we believe that despite some idiosyncrasies, they are not a-typical of practice elsewhere.

2. NSAs and the Nature of Spatial COVID-19 Data

2.1. Granularity and Confidentiality

NSAs traditionally conduct censuses and surveys, such as labor force surveys and income expenditure surveys, which provide detailed demographic information about individuals. They also provide geographic information. For example, in the United States the census tract block in which there are between 600 and 3,000 inhabitants is the most granular spatial zone in public use files. In the United Kingdom, the statistical ward is the most granular unit and wards are merged if they have less than 1,000 inhabitants. In Israel, the statistical area is the most granular spatial unit and the populations in these zones range between 1,000 and 5,000.

NSAs seek to guarantee absolute confidentiality. Geographical or spatial data are key candidates for disclosure (Fienberg 1994; Fienberg and Willenborg 1998). Suppose, for example, that in a most granular zone, occupations are recorded and there happens to be only one vet. Unless the vet's occupation is concealed it will be possible to know his or her income as well as other personal data. The public at large may not know that there is only one vet, but matters are different for other residents in the vet's zone, as well as perhaps in neighboring zones. If there are two observationally (demographically) similar vets, each vet will know the other vet's income, and others will know their income range. Absolute confidentiality is more likely to be infringed the smaller the number of vets and the more observationally different they are. If there are ten demographically different vets, each vet can be identified. If they are demographically identical, each vet faces a 10% probability of identification. NSAs anonymize the data so that such individuals cannot be identified.

There are, of course, numerous examples of data censoring motivated by absolute confidentiality. This generally arises with respect to large microdata sets such as the Community Innovation Survey in the EU (Franconi and Ichim 2009) and business and household survey microdata in the US such as the BLS Current Employment Statistics or Current Population Survey (Dalton et al. 2021). Another practice is top-coding, in which data relating to extreme values in variables such as income or demographic and health attributes are censored to protect the confidentiality of atypical and identifiable populations such as millionaires or the aged.

Suppose that there are a number of vets in the zone, but their data include dates of birth. If there is public access to a national register of vets, which includes names and dates of birth, individual vets may be identified through triangulation. In such cases, NSAs censor the data to protect their confidentiality.

These censoring practices are rightly motivated by absolute confidentiality, despite the fact that the general public may have no way of revealing that there is only one vet. When the unit of observation is an aggregate such as a zip code, neighborhood or statistical area matters are different. In these zones, the probability of detection faced by individuals is 1/N, where N denotes the population in the zone. The more granular the zone, the smaller is N, hence the higher is the probability of detection. There is an obvious trade-off between granularity, or spatial resolution, and relative confidentiality as measured by the probability of detection. We make two arguments. First, although not required by law, NSAs have attached importance to relative confidentiality. Second, they have set arbitrarily severe criteria regarding the trade-off between spatial resolution and relative confidentiality.

NSAs increasingly provide geocoded data of various types. These data take the form of spatial panel data in which the unit of observation has coordinates in space and time. For example, quarterly house prices in zones (e.g., Federal Housing Finance Agency for US metropolitan statistical areas), or labor market data in zones (e.g., European Union's NUTS2 regions). These spatial zones are not too granular, so the issue of confidentiality does not arise. On the other hand, data on municipal election results are often highly granular, as they are for some countries in the case of COVID-19.

2.2. *The Nature of Spatial Units*

Using zones rather than individuals as units of observation raises questions regarding the relevance of individual confidentiality in spatial data. Zones cannot be considered as 'individuals' even if their attributes, such as topology and composition, are unique. Furthermore, zones vary by shape and size. These issues constitute the well-known modifiable areal unit problem (MAUP) in spatial analysis (Openshaw and Taylor 1979; Fotheringham and Wong 1991; Nelson and Brewer 2017; Tuson et al. 2019). MAUP highlights the arbitrary nature of spatial units and the distortions arising from the way in which space is aggregated. A related matter refers to self-selection of individuals into zones or neighborhoods (Clark 1991; Kwan 2012; Burden and Steel 2016). Individuals or firms locate in zones according to their characteristics. For example, the housing locations of individuals may reflect their physical or socio-economic amenities, including school quality, crime, and parks. Additionally, their demographic composition changes through immigration and emigration. Hence, notions relating to the protection of individual confidentially in spatial data become obtuse.

A further issue concerning aggregating spatial zones relates to information loss. Shlomo (2010) tests the empirical impacts of aggregating or merging spatial units in an effort to preserve confidentiality. She finds that this approach generates more information loss than alternative methods for preserving confidentiality such as post-randomization probability (the PRAM mechanism), in which categories of variables are changed according to a prescribed probability matrix and a stochastic selection process.

2.3. Absolute and Relative Confidentiality

Laws of confidentiality are concerned with absolute confidentiality, which involves the release of information about individuals. See for example the EU's General Data Protection Regulation (GDPR) Recital 26 "The principles of data protection should apply to any information concerning an identified or identifiable natural person." https://gdprinfo.eu/recitals/no-26/. Spindler and Schmechel (2016) discuss the way the GDPR addresses absolute versus relative confidentiality. In this event, the probability of detection is one by definition. These laws do not directly address the difference between absolute confidentiality and relative, or probabilistic, confidentiality. There is an obvious qualitative difference between absolute and relative confidentiality. Sweeney (2002) has referred to this as 'k-anonymity' in which the probability of detection is 1/k. If k equals one, absolute confidentiality is at issue; if k exceeds one, relative confidentiality is at issue.

As NSAs are mandated to protect the identity of individuals and as existing legislation seeks to guarantee absolute confidentiality, one might argue that by default only absolute confidentiality should be in the purview of NSAs. According to this view, if the probability of detection is one half because the number of individuals with COVID-19 is one and there are only two inhabitants in the zone, anonymity is preserved because it is impossible to determine which of the two inhabitants has COVID-19. If, instead, both inhabitants have COVID-19 it would be necessary to anonymize or de-identify the data to prevent infringement of absolute confidentiality. Whereas absolute confidentiality is uniquely defined, relative confidentiality is not.

We document below how NSAs have restricted public access to spatial COVID-19 data ostensibly on the grounds of confidentiality and data protection. For example, In Israel the Ministry of Health does not publish COVID-19 data for zones with less than 2,000 inhabitants, and if there are more than 2,000 inhabitants, it only provides uncensored data if the number of cases is at least 15. If the number is between 1 and 14, the precise number is concealed. This censoring is administered not only to Covid-19 data aggregated into zones but also to other aggregates such as Covid-19 data by age groups (see, for example DataGov (2021b) where ,15 truncation is also applied). In statistical terms, the latter data are 'censored', whereas the former data are 'truncated'.

NSAs in other countries apply similar rules for censoring and truncation, but with different degrees of restriction. Less liberal NSAs have larger population cut-offs (3,000 instead of 2,000) and larger thresholds for the number of cases in the data (20 instead of 15). Censoring and truncation are usually justified by NSAs on the grounds of confidentiality, but they do not distinguish between absolute and relative confidentiality.

A further example of NSAs mandating excessive data protection and imposing misdirected regulation relates to the insistence of NSAs (e.g., in Israel) on compliance

with the Declaration of Helsinki. This statement outlines the ethical principles guiding medical research involving experiments with human subjects. Since COVID-19 data have not been generated experimentally, the Declaration of Helsinki is not relevant. Nevertheless, laws of individual confidentiality may be relevant if the probability of detection is large. In this case, protecting the individual's identity is not dependent on the number of COVID-19 cases in the data because the probability of detection is 1/N for all. It is relevant, however, for collective confidentiality faced by NSAs, which obviously varies inversely with the number of cases. When confidentiality is juxtaposed with the right to freedom of information, the case for limiting public access to official statistics needs to show that the latter compromises the former. While this can be upheld for absolute confidentiality, the issue is more obscure with respect to relative confidentiality. For example, spatial data for COVID-19 are required for the epidemiological study of its spatiotemporal diffusion (Elliot et al. 2020; Krisztin et al. 2020; Tsori and Granek 2021), in which context more granularity is better than less. Research on the spatial diffusion of COVID-19 will inform the design of local lockdown, social distancing and 'traffic light' policy (Giannone et al. 2020; Narayanan et al. 2020; O'Sullivan et al. 2020). Also, the public has a right to know for their own protection where the incidence of COVID-19 is greater or less. Here too, more granularity is better than less.

While NSAs use of the Helsinki Declaration imposes an unnecessary hurdle, the directive does establish the important principle that a trade-off exists between public interest and personal privacy. Although observational data (such as spatial COVID-19 data) are not obtained through 'informed consent' including 'disclosure of personal information', nevertheless the probability of individual detection needs to be balanced against the probability of benefiting from the freedom of information. In the case of spatial data for COVID-19, the needs of science and society are very large. These include replacing national lockdown policy, for which the economic and social costs are very large, by spatial lockdown policy for which these costs are much smaller.

NSAs have enabled authorized researchers complete access to anonymous but uncensored and untruncated data in 'research rooms' using stand-alone computers and under strict supervision to prevent data leakages. More recently, 'virtual' research rooms have been developed to enable remote access to unexpurgated confidential data so that researchers do not have to be present physically (Reuter and Musuex 2010). While these simply extend the trend of increasing remote access, they raise a host of issues relating to the competencies of NSAs in establishing and monitoring such facilities (Eurostat 2009). NSAs have also made available micro data under contract (MUC) to authorized researchers, who agree to legal stipulations and limitations. MUC files are more restricted than those available in research rooms. These welcome developments are not germane here, where we are concerned with public use files (PUF), which are accessible to the public at large without having to undergo bureaucratic screening.

3. Concepts of Confidentiality and Techniques of Protection

In this section we define more rigorously the concepts of relative and collective confidentiality on the one hand, and truncation and censoring on the other.

3.1. Relative and Collective Confidentiality

Let $\theta = 1/N$ denote the probability of detection faced by individuals where N is the population in the zone. If the outcome applies to a subgroup of the population, for example, adults, then N would exclude children. Let n be the number of COVID-19 outcomes (such as morbidity, hospitalizations or deaths) in the spatial zone. The probability of d detections has a binomial distribution:

$$P(d) = \binom{n}{d} \theta^d (1 - \theta)^{n-d} \tag{1}$$

The mean number of detections is $n\theta$ with variance $n\theta(1-\theta)$. Equation (1) makes the simplifying assumption that θ is the same for all subjects. After the first subject is discovered, the probability of detection increases from 1/N to 1/(N−1), and so on. Strictly speaking, therefore, P(d) has a hypergeometric distribution. However, because in the case of COVID-19 outcomes N is large relative to n, 1/(N−d) is insensitive to d. Consequently, we use Equation (1) to illustrate our arguments even if it slightly underestimates the probabilities of individual and collective detection.

Whereas individuals are naturally concerned with their risk of personal detection as expressed by θ, the statistical authorities are concerned with the probability that anyone will be detected regardless of who it might be, as expressed by $1-P(d = 0) = P(d > 0)$. We refer to this probability as the "collective probability" of detection because it expresses the collective risk that at least someone will be detected. The collective probability of detection is obviously many times greater than the individual probability of detection because it varies directly with n.

If n is absolutely large, but continues to be small relative to N (as it typically does in COVID-19 data), the Poisson distribution, which is computationally simpler, provides a good approximation to the binomial distribution, especially when $n > 20$ and $\theta < 0.05$ and when $n > 100$ and $n\theta < 10$. In this case, Equation (1) becomes:

$$P(d) = \frac{(n\theta)^d e^{-n\theta}}{d!} \tag{2}$$

Let $\lambda = n/N$ denote the incidence of COVID-19 in the population. For example, if the outcome is cumulative morbidity, λ is the proportion of the population diagnosed with COVID-19, which over three waves of COVID-19 in Israel, averaged about 0.01 or 1%. With the passage of time λ increases as new cases are diagnosed. If the outcome refers to new cases diagnosed $\lambda = \Delta n/N$. Notice that the mean number of detections is $n\theta = \lambda$ with variance $\lambda(1\text{-}1/N)$. Hence, the variance varies directly with the morbidity rate and varies directly with population. As N tends to infinity, the mean equals the asymptotic variance, as expected.

Table 1 illustrates Equation (1) for different values of N and n (or λ). In the first row in Table 1 there are 20 cases of COVID-19 in a population of 2,000, hence the individual probability of detection is 0.0005 or 0.05% and $\lambda = 0.01$ or 1%. The probability of collective detection faced by the statistical agency, measured by the probability of at least one detection, is 0.995%. (The probability of 1 detection is 0.99%). As expected, the

Table 1. Individual versus collective risk of detection.

N	θ	n	P(d > 0)	E(d)	sd
2,000	0.0005	20	0.00995	0.01	0.1
2,000	0.0005	100	0.04878	0.05	0.224
4,000	0.00025	40	0.009905	0.01	0.1
4,000	0.00025	100	0.02469	0.025	0.158
1,000	0.001	10	0.0096	0.01	0.1
1,000	0.001	100	0.09521	0.1	0.316
200	0.005	2	0.0097	0.01	0.1
400	0.0025	4	0.0096	0.01	0.1
800	0.00125	8	0.0096	0.01	0.1

Note: Based on Equation (1).

probability of collective detection is many times greater than the probability of individual detection. In row 1 the probability of collective detection is 19.9 times larger than the probability of individual detection. The expected number of detections is 0.01 with standard deviation equal to 0.1. Row 2 is the same as row 1 except there are 100 cases of COVID-19 instead of 20, so $\lambda = 0.05$. The expected number of detections increases fivefold, and the probability of collective detection increases to 4.88%, which has increased to 97.56 times larger than the probability of individual detection.

In rows 3 and 4 the population is doubled to 4,000 and in rows 5 and 6 it is halved to 1,000. In the final three rows the population is less than a thousand, and the number of cases is assumed to be one percent of the population. The individual probabilities of detection vary between 0.5% and 0.125%, while the collective rates of detection are 0.97%.

In summary, collective rates of detection are many times larger than individual rates of detection for given rates of incidence (λ). Hence, the risk of detection faced by statistical agencies, where at least one individual is detected, is much greater than the risk of detection faced by individuals. Perhaps this phenomenon motivates statistical authorities to truncate the data. If so, for given rates of incidence, Table 1 shows that the probability of collective detection is virtually independent of population size; the exposure of statistical authorities to collective detection is the same if the population is 1,000 (row 5) as it is when it is 4,000 (row 3). We therefore conclude that individual probabilities of detection remain small for populations less than 1,000, while collective probabilities of detection are insensitive to population size.

3.2. Truncation

What would an NSA achieve if it decided to truncate the data at 2,000 instead of 1,000? For these purposes we may compare rows 1 and 5 in Table 1, which share common assumptions for $\lambda = 0.01$. First, relative confidentiality faced by individuals is much greater because the probability of personal detection is 0.1% when the population is 1,000 and it is 0.05% when the population is 2,000. However, collective confidentiality is hardly different; it is 0.96% when the population is 1,000 and it is 0.995% when the population is 2,000. Hence, a more liberal NSA, which makes public data for less populated zones, decreases relative confidentiality faced by individuals, but increases collective

confidentiality faced by NSAs to a much smaller extent. This difference stems from the fact that, conditional on λ, there are fewer cases of COVID-19 in less populated zones.

3.3. Censoring

In this section we now illustrate why, contrary to NSA claims, censoring is unrelated to data protection. In contrast to the foregoing, our statistical critique now draws on a real-world example. The Ministry of Health (MoH) in Israel censors the number of cases between 1 and 14. If the population is 4,000 the individual probability of detection is 0.025%. If the number of cases is 1, the collective probability of detection equals the individual probability of detection. If the number of cases is 14 the individual probability of detection remains unchanged, but the collective probability of detection increases to 0.0349%. The true probability of collective detection is bounded by these limits. Censoring makes no difference to the individual probability of detection, but why should the MoH wish to conceal the collective probability of detection?

In any event, the data cease to be censored when the number of cases exceeds 14. For example, if the number of cases is 15, it becomes public knowledge that the collective probability of detection is 0.0374% whereas the individual probability of detection remains unchanged at 0.025%. So, what is the purpose of censoring the data when sooner or later the collective probability of detection is going to become public information? There is no rational reason.

Indeed, this issue is even more puzzling because MoH applies the same rules of censoring to the cumulative number of cases (n) as well as the number of new diagnoses (Δn). Initially the number of cases is zero, so the zone is 'clean'. MoH publishes this information because it believes correctly that issues of confidentiality do not arise. Suppose at some point in time t_1 $\Delta n = 3$ so $n = 3$. The zone ceases to be clean, but the data for n and Δn are censored because they are less than the threshold (14). At t_1 somewhere between 1–14 cases were diagnosed. Suppose later at t_2 that $\Delta n = 7$ so that $n = 10$. The data continue to be censored. Nevertheless, we at least know at t_2 that n in t_1 could not have been greater than 13, therefore somewhere between 1–13 cases were diagnosed and in t_2 the range of n is 2–14. Suppose at t_3 $\Delta n = 6$ so that $n = 16$. The latter ceases to be censored because it exceeds 14, but the former continues to be censored. At t_3 we know that there were between 2–14 new diagnoses. Finally, suppose at t_4 n increases to 18 so that $\Delta n = 2$. Since the latter is less than 14 it remains censored. However, this censoring no longer matters because Δn may be calculated directly using the uncensored data for n. Despite this MoH continues to censor Δn regardless of the fact than n has ceased to be censored.

In summary, whereas truncation may, in principle, be justified in terms of relative data protection, censoring has no rationale. It creates an artificial smoke-screen, which has nothing to do with data protection either individual or collective, and which may create the impression that NSAs have something to hide. It may also create the impression that they are irrational. Re-identification is not an issue here, as zone-based COVID-19 morbidity data released by NSAs provide no other identifying characteristics of the individuals in the zone. Finally, collective confidentiality faced by NSAs varies inversely with truncation simply because there are more cases of COVID-19 in more populated zones.

4. Spatial COVID-19 Data Availability

4.1. Comparing Countries

NSAs release COVID-19 data at different levels of spatial granularity. Even within the EU there is no uniform spatial unit that serves all member states (ECDC 2021). The choice of spatial resolution has implications for confidentiality. The constraints on data availability allow us to compare across a selection of countries for which local-level data are available (Table 2). To afford comparison we standardize the different spatial units of availability to

Table 2. Availability of spatial COVID-19 data by country and subnational spatial units.

Country	Availability of spatial COVID-19 data	Spatial unit of availability*	NSA response
Canada	Low-level granularity	Sub-provincial area health authorities (NUTS 3)	-
Australia	Low-level granularity	Zip code (LAU)	-
New Zealand	Low-level granularity	District Health Board (NUTS 3)	-
Japan	Low-level granularity	Prefectures (NUTS 2/3)	-
S. Korea	Low-level granularity	Counties (LAU)	-
Sweden	Low-level granularity	Municipalities (LAU)	-
Germany	Low-level granularity	Landkreisen (NUTS 3)	-
Italy	Low-level granularity	Provinces (NUTS 3)	-
United Kingdom	Restricted	Middle Layer Super Output Areas (LAU)	Censoring <3 cases Truncation $<$ 4,500 pop
Belgium	Restricted	Municipalities (LAU)	Censoring <5 cases
Israel	Restricted	Statistical Areas (LAU)	Censoring <15 cases Truncation $<2{,}000$ pop
United States	Incidental	Counties (LAU)	Small zones unrestricted
France	Incidental	Communes (LAU)	Small zones unrestricted
Spain	Incidental	Municipalities (LAU)	Small zones unrestricted
Netherlands	Incidental	Municipalities (LAU)	Small zones unrestricted

*Corresponding EU NUTS spatial units in parentheses: NUTS 2 regions have roughly 0.8–3.0 million inhabitants; NUTS3 regions have populations ranging from 150–800 Th; LAUs have populations ranging from double digits to over 100,000 inhabitants.

EU NUTS units. We distinguish between three types of data restrictions depending on the level of spatial resolution (Table 2).

1. For administrative reasons there happen to be no data that are sufficiently granular for issues of confidentiality to arise. The majority of countries fall into this category, for example Canada, Australia, Sweden, Germany and Italy. On the other hand, we cannot rule out that NSAs in these countries might have decided to avoid developing more granular data on the grounds of confidentiality,
2. Granular data happen to be available, but the statistical authorities restrict their availability on the grounds of confidentiality, as in Israel, Belgium and the United Kingdom, and
3. Granular data happen to be 'incidentally' available rather than by design. This occurs in countries such as the United States, France, Spain and the Netherlands. For example, COVID-19 data are available for US counties, which typically have large populations. However, a handful of counties have zones with populations less than a thousand. Although these incidentally available data may not be useful for research into the spatial diffusion of COVID-19, they establish the principle that issues of confidentiality do not arise in small zones.

Confidentiality does not overtly arise for the first category. It is always possible, however, that it arises invisibly; the data are available to government agencies but they do not acknowledge their existence. In principle, NSAs can compile such data from individual administrative records to which they have access. However, they might not have carried out this exercise, or they might not have had the necessary geocoded data to do so. Confidentiality arises overtly for the second category. As for the third category, the statistical authorities act as if issues of confidentiality do not arise.

Spatial COVID-19 data are available for almost all countries, (see, for example ECDC 2021; Naqvi 2021). However, in most cases their degree of granularity is low; even the smallest spatial unit has several thousand inhabitants, if not more (Table 2). For example, in Canada the spatial units are sub-provincial area health authorities, the smallest of which have populations exceeding 10,000. In Italy the data are by province, the smallest of which (Isernia) had a population of 84,379 in 2019. In Sweden and Germany too, the data for municipalities and Landkreisen and Kreisfreien Städte are for large spatial units. The same applies to data available for 20 District Health Boards in New Zealand, and zip codes in Australia, where even in rural areas and the outback zip code populations exceed 10,000. Data are available for 47 prefectures in Japan and 154 cities and counties in South Korea, all of which have populations that run into the 10,000s and more.

However, for some countries the spatial units for which COVID-19 data are reported have populations less than 1,000. While the vast majority of US counties have populations exceeding 10,000 and at the extreme, Los Angeles county has a population of over 10 million, some counties have small populations. For example, COVID-19 data are available for Grant County in Nebraska with a population of 660 in 2018. France comprises 36,552 communes many of which have populations less than 1,000 for which COVID-19 data are available. The Netherlands comprises 355 municipalities for which COVID-19 data are available, most of which have large populations. However, some such as Schiermonnikoog have small populations (947). There are 581 Belgian municipalities, of which five

have populations between 1,000–2,000 for which COVID-19 data are available (not truncated). However, the data are censored if the number of cases is between 1–5. The same applies to Spanish municipalities, not all of which have data for COVID-19, but some municipalities such as Priego-Cuenca (896 inhabitants) and Camaleno-Cantabria (938 inhabitants) have small populations (Table 2).

We have already mentioned that in its public use file, the Ministry of Health in Israel truncates COVID-19 outcomes for statistical areas with populations less than 2,000, and it censors outcomes for with 1–14 cases otherwise. The Office for National Statistics (UK) reports COVID-19 morbidity in England and Wales during the previous seven days for 'middle layer super output areas' (MLSOA), which are sub 'lower tier local authorities'. Although MLSOAs are the most granular data available, the smallest MLSOA has 4,500 inhabitants and many have more than 10,000. However, ONS suppresses these data 'in the interest of confidentiality' if the number of diagnoses is less than 3. Hence, ONS truncates data by ensuring that MLSOAs have at least 4,500 inhabitants and it censors them if morbidity is less than 3. In Scotland the spatial unit equivalent to the MLSOA is the Intermediate Zone (IZ) with a minimum population of 2,500. Public Health Scotland censors data if the number of Covid-19 cases in these zones is between 1 and 5.

In summary, for the vast majority of countries, spatial COVID-19 data are neither truncated nor censored because issues of confidentiality do not arise since zones have large populations. In some countries, such as Belgium, Latvia and Estonia, the data are censored or converted into ranges (see Naqvi 2021) but not truncated. In others they are truncated but not censored, and in Israel and the United Kingdom they are both censored and truncated. Finally, in countries such as the United States the data are neither censored nor truncated; they are 'incidentally' unrestricted.

4.2. Availability of Other (non-COVID-19) Spatial Data

The restrictions imposed on spatial COVID-19 data do not seem to be applied to other spatial data. Election outcome data are available spatially almost universally. For example, they are available for US counties, some of which have populations less than 1,000, as noted. In the United Kingdom, they are available for all electoral wards regardless of size. The publication of election results in a ward in which as many as 90% voted for the Labour Party is not regarded as violating privacy, even where the electoral turn-out was very high. Election results are available for locations in Israel provided the electorate exceeds 1,000. In almost all countries, election results are available to a high degree of granularity. Although in principle there is no difference between the privacy of political preferences and individual health status, in practice statistical authorities in United Kingdom, Belgium and Israel apply stricter criteria to morbidity data than they do to electoral data. On the other hand, election results are made public for reasons of democratic transparency, even where electorates are small.

In Israel, the Central Bureau of Statistics (CBS) has recently started to publish data for socio-economic clusters by statistical areas. These clusters range upwards from 1 to 10 based on a variety of social and economic outcomes in these areas. However, for reasons of confidentiality, CBS truncates the data for statistical areas with less than 120 inhabitants (of which there are very few). Whereas the Ministry of Health (MoH) truncates COVID-19

data at 2,000, and the Interior Ministry truncated election results at 1,000, the Central Bureau of Statistics truncates socio-economic data at 120. Since the socioeconomic status of individuals is just as confidential as their COVID-19 status, either MoH arbitrarily attaches more importance than CBS to confidentiality, or the inconsistency results from administrative incompetence.

Another example relates to housing transactions. In Israel, these require the payment of Acquisition Tax according to the price contracted. Following a successful legal challenge based on the Freedom of Information Act, the Tax Authority provides a public use file for the universe of individual house price transactions (dating back to 1989) to a very high degree of spatial granularity. Indeed, one of the purposes of this data transparency is to increase the efficiency of housing markets so that the buyers and sellers can inform themselves of recent transaction prices in neighborhoods of interest (Ben Shahar and Golan 2019). Since there are typically about 1,200 apartments in these zones, the probability of detection is much greater than it is for COVID-19 data. Moreover, the PUF contains data on housing characteristics, which increase identifiability. For houses bearing 'for sale' posters, identifiability is even greater. More recently, the Tax Authority has mapped the exact locations of housing units so that it is possible to know how much buyers paid for their housing and how much sellers received. Although individuals are not identified in these data, their neighbors know how much money they received. Similar house price data are available in the Netherlands via Kadaster–the Dutch land registry (Kadaster 2020), in the United Kingdom from HM Land Registry (GOVUK 2020) and in the United States through the Zillow's Assessor and Real Estate Database (ZTRAX 2020).

In summary, criteria for confidentiality in spatial data vary between countries for the same outcomes, and they vary within countries for different outcomes. Also, confidentiality criteria for COVID-19 outcomes vary between countries, and they vary within countries with respect to other outcomes. They even vary within countries for other medical data. For example, the Israeli Ministry of Health publishes spatial data on cancer incidence through the National Cancer Registry and censors the data in those zones with less than 50 cases annually. Considering that the rate of common cancers is about 100 per 100,000, this effectively means truncating the release of data to statistical areas with 50,000 residents. It thus seems that each statistical authority sets its own criteria. There is no coordination.

5. Stretching the Law of Data Protection

When a new phenomenon arises, such as COVID-19, providers of national statistics invent new criteria, which are supposed to protect individual confidentiality. These criteria have nothing to do with the protection of absolute confidentiality. Nor do they have much to do with relative confidentiality, because in practice probabilities of individual detection are very small. At most, they may have something to do with collective confidentiality faced by NSAs because the probability of collective detection is inevitably much greater than the probability of individual detection. Perhaps this lies behind the conservatism of NSAs in making public spatial data for COVID-19, which are sufficiently granular. By reducing the granularity of the data that they make public, NSAs directly reduce the individual probability of detection, which they believe will indirectly reduce the probability of collective detection. The comparison made above between rows 1 and 5 in Table 1 shows

that this belief is false. Merging two zones with 1,000 people into one zone with 2,000 people halves the individual probability of detection but increases the relative risk of collective detection by 3.64% in this numerical example. If NSAs are motivated by collective confidentiality, they should make the data more granular, not less. This means less truncation, not more.

Since the laws of confidentiality apply to persons and do not refer to collective identification, there does not seem to be a legal basis to the practice of data truncation by NSAs. The same applies *a fortiori* to the practice of data censoring by NSAs.

Laws of privacy refer to absolute confidentiality; they do not refer to relative or collective confidentiality. For example, the Law for the Protection of Privacy in Israel (1981) includes a list of offenses such as phone tapping, which clearly concern individuals. Issues of relative or collective confidentiality do not arise for phone tapping and other offenses listed. Item 9 on the list refers to the "use of information concerning individuals or its transmission to others" unless they have granted permission. This item too is concerned with absolute confidentiality. Nor has case law been concerned with infringements of relative confidentiality either in practice or in principle (Zarsky and Bar-Ziv 2019).

In 1996, the law was updated with respect to databanks containing personal data. Proprietors of data banks were required to appoint Data Security Officers (DSOs) to ensure that the law of 1981 is not infringed. Also, individuals should be given access to their own data. The law of 1996 did not introduce new concepts of confidentiality, such as relative or collective confidentiality. These concepts were introduced by the DSOs. The widespread heterogeneity to which we have drawn attention in public access to spatial COVID-19 data and other spatial data, stems from the way different DSOs interpret their mandate. It also explains how even within the same NSA different criteria are applied by different DSOs; COVID-19 data are truncated at 2,000, whereas data for cancer are truncated at approximately 50,000. In summary, legislation for data protection regarding data banks has created a vacuum filled by DSOs who have invented new concepts of confidentiality, which are interpreted arbitrarily.

As noted, NSAs have concentrated entirely on relative confidentiality and have attached less importance to the social and scientific benefits flowing from freedom of information. Historically and legally, the trade-off between freedom of information to achieve societal goals and the protection of privacy of individuals has been implemented through data de-identification or anonymization. The practical mechanism for ensuring this privacy is invariably a variation of the classic k-anonymity algorithm (Sweeney 2002). This provides a framework for quantifying the likelihood of re-identification for anonymized data. A key strategy adopted by NSAs in this process is that of limited release, whereby data are transformed by limiting their granularity both temporal and spatial using censoring and trancation. The limited effectiveness of these constraints on re-identification becomes ever-more pronounced in a data environment fed by geo-located mobile data. In this context recent research shows that absolute (individual) confidentiality can be compromised by a limited set of data points. For example, just 4 spatio-temporal points are enough to detect 90% of observations in a credit card data base of 1 million and 95% in a cellular phone database of 1.5 million (De Montjoye et al. 2018). European NSAs ascribe to the EU's General Data Protection Framework (GDPR), which offers a legal basis for issues of data privacy and data security, and restricting data through limited release would seem to be consistent with that

goal. See GDPR, Article 89, Recitals 162–3 https://gdpr-info.eu/recitals/no-162/. However, while upholding GDPR practice, data confidentiality should not be confused with the data privacy and data security mandated by the GDPR (Prewitt 2011). Data confidentiality deals with data disclosure and informed consent ("don't tell"). Data privacy addresses data collection ("don't ask") and data security deals with safeguards imposed on information that has already been collected. Confounding these issues may explain why NSAs have confounded individual and collective data confidentiality.

Different legal traditions exist with respect to protecting data confidentiality. Frameworks such as the GDPR in the EU opt for a centralized approach whereas individual states in the United States set their own rules. In general, confidentiality in the United States is restricted to financial, genetic and medical data that are personal, whereas GDPR applies to all data including political data as well as innocuous data such as hair color.

In terms of actual legislation, the traditions range from using a global approach grounded in primary legislation (Israel) to ad hoc regulation governing individual sectors such as health, communications and so on, as exists in the United States. We agree with Zarsky and Bar-Ziv (2019) that although Israel ostensibly has a centralized, global approach to the protection of confidentiality, in practice there is extensive heterogeneity in the way the law is applied by different statistical agencies. Indeed, as we have seen, even the same statistical agency applies different criteria to different data.

When the case for data confidentiality is confronted with 'the public interest' as in the case of COVID-19, the legal tradition in Israel is rooted in individual confidentiality. Thus, Zarsky and Bar-Ziv (2019) note that anonymized personal medical data (protected under the Law of the Rights of the Patient 1996) can be released if the goal is to protect public health. However, structural tension exists in the law with respect to collective confidentiality. Here legal reading tends to an overly-constraining interpretation that results in the protection of individuals who are part of collective entities such as geographic zones or neighborhoods. According to this interpretation, if statistical inference about individuals is based on group characteristics (the ecological fallacy issue not withstanding), then data restrictions may be justified. This group or 'attribute' disclosure (Fienberg and Willenborg 1998) arises, for example, if public data on average neighborhood earnings is expected to impact negatively on residents with earnings that are significantly different to the neighborhood average. Zarsky and Bar-Ziv (2019) note that such statistical stereotyping may challenge the laws governing individual privacy. However, laws of privacy do not stipulate that statistical stereotyping is illegal.

Further tension exists between the competing legal demands for protection of confidentiality and the societal benefits resulting from its release, such as improved medical research and enhanced quality of life. In the context of COVID-19, releasing spatial data may be construed as stigmatizing zones with high rates of contagion. This has to be juxtaposed with the need for authorities to provide accurate spatial data in order to increase trust, legitimacy and public compliance. Furthermore, the public has the right to know where COVID-19 is particularly prevalent for their personal protection. Faced with new COVID-19 data demands from cell phone tracking, geo-located purchasing and vehicle movements, some commentators see further data release without sufficient safeguards as the thin end of the wedge and a slide towards socially negative directions such as growing economic inequality and social unrest (Dwork et al. 2020).

6. Conclusions

The growing demand for spatial COVID-19 data highlights some of the inconsistencies in NSA attempts to balance the competing claims for freedom of information on the one hand with protecting personal confidentiality on the other. As we show, while NSA response has varied greatly across countries, it has been consistent in confounding absolute and relative confidentiality and in failing to distinguish between individual and collective confidentiality. The result is heavy-handed NSA activity in the area of data protection. This is expressed via overly-severe data truncation and data censoring that is unrelated to data protection.

By definition, national legislation in the area of personal confidentiality relates to individual and not collective confidentiality. NSA and government ministries have appointed DSOs with the express aim of instituting de-identification and anonymization practices to preserve personal confidentiality. With the increasing demands on NSAs to provide spatial data, DSOs have taken to filling the void unaddressed by individual confidentiality legislation, and have invented new ground-rules for collective and relative confidentiality. There is a need to regulate DSOs and to set guidelines governing their mandate.

With respect to absolute and relative confidentiality, matters are similar. In the absence of explicit legislation, which only addresses absolute anonymity (i.e., by default k anonymity = 1), DSOs have again stepped into the void and determined arbitrary probabilities of detection. Whether k is 5 or 15 is not an issue that should be left to the individual discretion of DSOs. While legislation obviously cannot dictate the 'right' level for k, this is an area for which one size does not fit all. Empirical research can go a long way in providing guidelines for the formulation of consistent criteria in this field.

The above issues are pertinent to all spatial data protection whether economic, genetic or medical and not just spatial COVID-19 data. However, COVID-19 is contagious and has serious externalities and spatial spillovers that do not apply to other diseases, although they may apply to diseases subject to environmental influences such as certain forms of cancer. The public 'right to know' is particularly acute in the case of COVID-19. A freedom of information issue exists with spatial COVID-19 data that does not exist with other similar spatial data. This heightens the concern over arbitrary DSO data protection practices.

In summary, we make the following recommendations regarding the public availability of spatial COVID-19 data:

1. Data censoring should be abandoned; it serves no purpose.
2. Data truncation should be greatly curtailed. Probabilities of detection should be increased from 1 per million to no more than 1%.
3. National statistical offices should regulate the ad hoc practices of DSOs.
4. Ministries of Justice should review the case for relative confidentiality.

7. References

Ben Shahar, D., and R. Golan. 2019. "Information shock and price dispersion: A natural experiment in the housing market." *Journal of Urban Economics* 112: 70–84. DOI: https://doi.org/10.1016/j.jue.2019.05.008.

Burden, S., and D. Steel. 2016. "Empirical zoning distributions for small area data." *Geographical Analysis* 48(4): 373–390. DOI: https://doi.org/10.1111/gean.12104.

Clark, W.A.V. 1991. "Residential Preferences and Neighborhood Racial Segregation: A Test of the Schelling Segregation Model." *Demography* 28: 1–19. DOI: https://doi.org/10.2307/2061333.

Dalton, M., J.A. Groen, M.A. Loewenstein, D.S. Piccone, and A.E. Polivka. 2021. "The K-Shaped recovery: Examining the Diverging Fortunes of Workers in the Recovery from the Covid-19 Pandemic using Business and Household Survey Microdata." *Covid Economics* 71: 19–58. Available at: file:///C:/Users/Owner/Downloads/CovidEconomics71%20(1).pdf (accessed November 2021).

DataGov. 2021a. *Covid-19 Data by Statistical Areas.* Available at: https://data.gov.il/dataset/covid-19/resource/d07c0771-01a8-43b2-96cc-c6154e7fa9bd (accessed November 2021).

DataGov. 2021b. *Covid-19 Data by Sex and Age Categories.* Available at: https://data.gov.il/dataset/covid-19/resource/89f61e3a-4866-4bbf-bcc1-9734e5fee58e (accessed November 2021).

De Montjoye, Y.-A., S. Gambs, V. Blondel. et al. 2018. "On the privacy-conscientious use of mobile phone data." *Scientific Data* 5: 180286. DOI: https://doi.org/10.1038/sdata.2018.286.

Dwork, C., A. Karr, K. Nissim, and L. Vilhuber. 2020. "On Privacy in the Age of COVID-19." *Journal of Privacy and Confidentiality* 10(2). DOI: https://doi.org/10.29012/jpc.749.

Elliot, R.J.R., I. Schumacher, and C. Withagen. 2020. "Suggestions for a Covid-19 Post Pandemic Research Agenda in Environmental Economics." *Environmental and Resource Economics* 76(4): 1187–1213. DOI: https://doi.org/10.1007/s10640-020-00478-1.

ECDC. 2020. *EU/EEA and UK Regional Data on Covid-19.* Available at: https://www.ecdc.europa.eu/en/publications-data/sources-eueea-regional-data-covid-19 (accessed November 2021).

Eurostat. 2009. *Working Session on Statistical Data Confidentiality.* Office for Official Publications of the European Communities, Luxembourg. Available at: https://ec.europa.eu/eurostat/documents/3888793/%205844781/KS-78-09-723-EN.PDF/f977ff33-bc9b-4d07-aec6-7dfd9ccc5d59?version=1.0 (accessed November 2021).

Fienberg, S.E. 1994. "Conflicts between the Needs of access to Statistical, Information and the Demands for Confidentiality." *Journal of Official* 10(2): 115–132. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/conflicts-between-the-needs-for-access-to-statistical-information-and-demands-for-confidentiality.pdf (accessed September 2021).

Fienberg, S.E., and L.C.R.J. Willenborg. 1998. "Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data." *Journal of Official Statistics* 14(4): 337–345. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/introduction-to-the-special-issue-disclosure-limitation-methods-for-protecting-the-confidentiality-of-statistical-data.pdf (accessed September 2021).

Fotheringham, A.S., and D.W.S. Wong. 1991. "The modifiable areal unit problem in multivariate statistical analysis." *Environment and Planning A* 23: 1025–1044. DOI: https://doi.org/10.1068/a231025.

Franconi, N., and D. Ichim. 2009. "Community Innovation Survey: Comparable Dissemination": 11–23 in *Working Session on Statistical Data Confidentiality*. Office for Official Publications of the European Communities, Luxembourg. 17–19 December 2007, Manchester, UK.

Giannone, E., N. Paixão, and X. Pang. 2020. "The Geography of Pandemic Containment." *Covid Economics* 52: 68–95. Available at: file:///C:/Users/Owner/Downloads/CovidEconomics52%20(4).pdf (accessed November 2021).

GOVUK. 2020. HM Land Registry: *Price Paid Data.* Available at: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

Kadaster. 2020. Available at: https://kadasterservice.nl/situaties/kadastrale-woning-gegevens

Krisztin, T., P. Piribauer, and M. Wögerer. 2020. "The spatial econometrics of the coronavirus pandemic." *Letters in Spatial and Resource Sciences* 13: 209–218. DOI: https://doi.org/10.1007/s12076-020-00254-1.

Kwan, M.p. 2012. "The Uncertain Geographic Context Problem." *Annals of the Association of American Geographers* 102(5): 958–968. DOI: https://doi.org/10.1080/00045608.2012.687349.

Naqvi, A. 2021. "*Covid-19 European Regional Tracker*." *Nature: Scientific Data* 8: 181. DOI: https://doi.org/10.1038/s41597-021-00950-7.

Narayanan, R.P., J. Nordlund, P.K. Pace, and D. Ratnadiwakara. 2020. "Demographic, jurisdictional, and spatial effects on social distancing in the United States during the COVID 19 pandemic." *PLoS ONE* 15(9). DOI: https://doi.org/10.1371/journal.pone.023957.

Nelson, J.K., and C.A. Brewer. 2017. "Evaluating Data Stability in Aggregation Structures Across Spatial Scales: revisiting the Modifiable Areal Unit Problem." *Cartography and Geographic Information Science* 44(1): 35–50. DOI: https://doi.org/10.1080/15230406.2015.1093431.

Newlands, G., C. Lutz, A. Tamo-Larrieux, E.F. Villaronga, R. Harasgama, and G. Scheit. 2020. "Innovation under pressure: Implications for data privacy during the Covid-19 pandemic." *Big Data and Society* 7(2). DOI: https://doi.org/10.1177/2053951720976680.

OECD. 2020a. *Tracking and tracing COVID: Protecting privacy and data while using apps and biometrics (COVID-19)*, OECD Policy Responses to Coronavirus (Covid-19), April 2020 OECD, Paris. Available at: https://www.oecd.org/coronavirus/policy-responses/tracking-and-tracing-covid-protecting-privacy-and-data-while-using-apps-and-biometrics-8f394636/ (accessed November 2021).

OECD. 2020b. *Ensuring data privacy as we battle COVID-19*, OECD Policy Responses to Coronavirus (Covid-19), April 2020, OECD, Paris. Available at: https://www.oecd.org/coronavirus/policy-responses/ensuring-data-privacy-as-we-battle-covid-19-36c2f31e/ (accessed November 2021).

Openshaw, S., and P.J. Taylor. 1979. "A million or so correlation coefficients: three experiment on the modifiable areal unit problem." In *Statistical Applications in the Spatial Sciences*, edited by N Wrigley: 127–144. London: Pion.

O'Sullivan, D., M. Gahegan, D.J. Exeter, and B. Adams. 2020. "Spatially-explicit models for exploring COVID-19 lockdown strategies." *Transactions in GIS*. DOI: https://doi.org/10.1111/tgis.12660.

Prewitt, K. 2011. "Why It Matters to Distinguish Between Privacy and Confidentiality." *Journal of Privacy and Confidentiality* 3(2): 41–47. DOI: https://doi.org/10.29012/jpc.v3i2.600.

Poom, A., O. Jarv, M. Zook, and T. Toivonen. 2020. "COVID-19 is spatial: Ensuring that mobile Big Data is used for social good." *Big Data and Society* 7(2). DOI: https://doi.org/10.1177/2053951720952088.

Reuter, W.H., and J.M. Museux. 2010. "Establishing an Infrastructure for Remote Access to Microdata at Eurostat." In *Privacy in Statistical Databases. PSD 2010. Lecture Notes in Computer Science, 6344*, edited by J. Domingo-Ferrer and E. Magkos. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-15838-4_22.

Shlomo, N. 2010. "Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility." *Journal of Privacy and Confidentiality* 2(1): 73–91. DOI: https://doi.org/10.29012/jpc.v2i1.584.

Spindler, G., and P. Schmechel. 2016. "Personal Data and Encryption in the European General Data Protection Regulation." *7 JIPITEC- Journal of Intellectual Property, Information Technology and E-Commerce Law* 163. DOI: https://www.jipitec.eu/issues/jipitec-7-2-2016/4440.

Sweeney, L. 2002. "k-Anonymity: a model for protecting privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5): 557–570. Available at: https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_Article.pdf (accessed November 2021).

Tsori, Y., and R. Granek. 2021. "Epidemiological model for the inhomogeneous spatial spreading of COVID-19 and other diseases." *PLoS ONE* 16(2). DOI: https://doi.org/10.1371/journal.pone.0246056.

Tuson, M., M. Yap, M.R. Kok, K. Murray, and B. Turlach. 2019. "Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem." *International Journal of Health Geographics* 18: 6. DOI: https://doi.org/10.1186/s12942-019-0170-3.

Zarsky, T., and S. Bar-Ziv. 2019. "Privacy's 'Identity Crisis': Regulatory Strategies in the Age of De-Identification." *Law, Society and Culture* 2: 125–166. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350266 (accessed November 2021).

ZTRAX. 2020. Zillow's Assessor and Real Estate Database (ZTRAX). Available at: https://www.zillow.com/research/ztrax/ (accessed November 2021).

Received January 2021
Revised April 2021
Accepted September 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 811–836, http://dx.doi.org/10.2478/JOS-2021-0036

Response Burden and Data Quality in Business Surveys

Marco Bottone[1], *Lucia Modugno*[1], *and Andrea Neri*[1]

Response burden has long been a concern for data producers. In this article, we investigate the relationship between some measures of actual and perceived burden and we provide empirical evidence of their association with data quality. We draw on two business surveys conducted by Banca d'Italia since 1970, which provide a very rich and unique source of information. We find evidence that the perceived burden is affected by actual burden but the latter is not the only driver. Our results also show a clear link between a respondent's perceived effort and the probability of not answering some important questions (such as those relating to expectations of future investments and turnover) or of dropping out of the survey. On the contrary, we do not find significant effects on the quality of answers to quantitative questions such as business turnover and investments. Overall, these findings have implications for data producers that should target the perceived burden, besides the actual burden, to increase data quality.

Key words: Response burden; data quality; business surveys.

1. Introduction

Policymakers need high quality and detailed information on firms' decisions and performances to monitor the state of the economy and to assess the effectiveness of their policies. On the other hand, participating in a survey is a cost for businesses. It does not lead to any obvious financial return and it takes time away from profitable activities. If they receive frequent inquiries or if the information to be provided is burdensome, businesses are likely to give a low priority to these requests, especially if they are not mandatory. This means that they may refuse to collaborate or, even if they agree to participate, they may provide low-quality data (with low timeliness, with a high number of missing items or with measurement errors). It is therefore crucial to measure and monitor business response burden. This principle is clearly stated in the Quality Assurance Frameworks produced by several international organizations. The European Statistical Associations states that "response burden must be proportionate to the needs of the users and must not be excessive for respondents. The statistical authorities should monitor the response burden and set targets for its reduction over time." Along the same line, the United Nations require statistical agencies to "choose data sources taking into account accuracy and reliability, timeliness, cost and the burden on respondents." The literature generally distinguishes between actual and perceived burden (see Willeboordse 1997 and Haraldsen et al. 2013 for a comprehensive overview and discussion of this topic). While actual burden refers to objective measures of the complexity of the survey, perceived

[1] Bank of Italy, Via Nazionale, 91, Rome, 00184, Italy. Emails: marco.bottone@bancaditalia.it, lucia.modugno@bancaditalia.it and andrea.neri@bancaditalia.it

Acknowledgments: The views expressed are not necessarily shared by the Bank of Italy.

burden is essentially a subjective measure provided by the respondents. Another relevant distinction made by the literature is between gross and net burden. While the former only considers the costs of survey participation, the latter takes account the benefits enjoyed by respondents for their contribution (such as the feedback of survey results). To the best of our knowledge, in all the empirical studies actual burden is considered in gross terms, probably because the benefits are difficult to measure. There are a few different approaches for measuring such burden. Ideally, it should be calculated as the product of three factors: the number of people involved in the survey, the time they spent and their average hourly salary. Yet, this measure is rarely used, first because it depends upon a very noisy estimate of the cost/hour for the respondent, second because it is burdensome to collect. Generally, it is calculated by multiplying the number of completed questionnaires by an estimate of the average time required for completing and submitting the response, multiplied by the number of survey cycles during the year. Sometimes the burden is computed using the total sample size rather than the number of completed responses. In a minority of cases, it is measured in financial terms by multiplying the hours spent by the average hourly cost of respondent's time (Snijkers et al. 2013; Bavdaž et al. 2015). The main limit of actual burden is that the positive effects of survey participation can hardly be measured in objective terms. Perceived burden refers to the respondents' assessment of how burdensome they find it to comply with the data request. The importance of perceptions was initially stressed by Bradburn (1978). Respondents' assessment is likely to also include other aspects that may affect the burden that merely time measurement does not take into account. For example, Fisher and Kydoniefs (2001) suggest that respondents' perceptions can be affected by factors such as their motivation and belief in the utility of surveys or the method of data collection, or the item sensitivity. Yan et al. (2019) find that low motivation, difficult recall tasks, challenging survey effort, and negative perception of the survey all directly contribute to respondents' perception of burden. Dale et al. (2007) suggest to use two qualitative and simple questions to measure perceptions. The first relates to the perception of time taken while the second relates to the perception of the overall burden. They also recommend asking questions about which conditions mainly make the survey burdensome. Summing up, the measurement of perceptions is important for two main reasons. First, it enables us to capture more factors relating to the burden other than those relating to time and money. Second, it is an easy way to also consider the positive effects of survey participation. The final rate provided by the respondents is the balance between burdens and gratifications (Haraldsen 2004). Measurement of burden can be undertaken using different approaches. It can be collected directly from business survey participants including a small set of questions at the end of the questionnaire, or subsequently by follow-up contact with a sub-sample of the surveyed population. It can also be estimated indirectly, as a product of recordkeeping studies, tests or experts' valuations. Practices relating to the response burden measurement are likely to vary a lot across institutions. Dale et al. (2007) provide an overview of the questions used by several statistical agencies to collect information on the response burden. Bavdaž et al. (2015) conducted a survey of 41 national statistical institutes (NSIs) from 39 countries. They find large heterogeneity in measurement practices not only between institutes but also within them. Most institutes do some kind of measurement of actual burden, while perceived burden is less frequently assessed. This heterogeneity raises several issues. Probably, the

main one is which measures of response burden should be collected for monitoring data quality.

The current study contributes to the literature by addressing two research questions. The first one is to what extent actual burden is associated with perceived burden. There are not many studies dealing with both issues simultaneously, probably because of data limitations. Yet, for data producers it is useful to know to what extent such measures are associated: in the extreme case in which actual and perceived burden are perfectly correlated, it would not be necessary to collect both, reducing the overall burden on respondents. The second and main research question is whether perceived burden has a direct impact on data quality once many other potential associates are controlled for (including actual burden). Even if respondent burden is widely recognized to have an impact on quality, to the best of our knowledge, only a few articles attempt to empirically validate this statement and the evidence provided is often indirect. For example, Haraldsen and Jones (2007) use the number of values corrected during editing; whereas Giesen (2012) takes the timeliness of the response to evaluate the effects of perceived burden on response behavior. For data producers, it is important to know which factors affect data quality in order to identify strategies to improve it.

We draw on a unique and rich dataset made by two main business surveys conducted by the Bank of Italy, which collect information on both actual and perceived burden. Actual burden is measured by five indicators: number of pages of the questionnaire, number of questions, number of people involved in the survey, use of external consultancies and the average completion time. Perceived burden is measured by a qualitative question asking respondents to rate their effort to complete the questionnaire. It is also worth mentioning that these surveys are not mandatory. Because of that, respondents' burden may have very negative effects on data quality. Our research plan consists of first studying the relationship between perceived and actual burden and their dynamics over time. We then provide evidence of the relationship between perceived burden and data quality while controlling for actual burden and other factors. The structure of the article is as follows. Section 2 describes the conceptual framework we use in our analysis, while Section 3 presents our data. In Section 4, we provide an empirical analysis of the association between our measures of actual and perceived burden. In Section 5 we investigate the relationship between perceived burden and several measures of data quality. Section 6 concludes.

2. A Conceptual Framework for Response Burden and Data Quality

A first model of survey burden was developed in 1978 by Bradburn. He suggests a definition of response burden consisting of four elements: interview length, required respondent effort, frequency of being interviewed and the stress of psychologically disturbing questions, which may be asked (Bradburn 1978). He also pointed out that it is necessary to carefully focus on the respondent's subjective perceptions of the time and effort required to fill in the questionnaire. Fisher and Kydoniefs (2001) propose a model in which response burden is a combination of 32 elements that can be grouped in three categories: respondent burden (personality traits and attitudes of the respondents), design burden (characteristics of the survey such as duration of the interview and the wording of

the questions) and interaction burden (what happens when respondents with certain characteristics are confronted with a survey that has certain properties). Haraldsen (2004) and Haraldsen et al. (2013) further elaborate on the previous models by highlighting that it is crucial to assess whether the perception of the burden outweighs the positive aspects of the survey. The key variable to monitor for data quality is the final perceived balance between negative and positive aspects. In this article, we adopt the Total Business Survey Burden Model (TBSB) discussed by Jones et al. (2005) and Dale et al. (2007). Within this framework, perceived burden is described as an intermediate variable between data quality and three elements: survey design components, respondent characteristics and contextual business factors. Survey design refers to aspects such as the data collection method, the communication strategy, the length and content of the questionnaire. A respondent's characteristics mainly relate (her/his) cognitive ability, motivation and capacity to collect the necessary information. Contextual factors are linked to the organization of the businesses (which may facilitate or prevent the collection of information) and to its strategies. Two important points can be drawn from the total business survey burden model. First, perceived burden is the only factor that directly affects data quality. All the other factors relating to the survey design, respondent characteristics and the general context only contribute to shaping this burden. Second, perceived burden is a more general concept than actual burden since it may originate from many sources other than those related to the time and/or money it takes to comply with a survey request.

This framework can be applied in order to gain a better understanding of the effects of perceived burden in two important phases of the survey: the recruitment of respondents and data collection.

In the initial phase of recruitment, the decision of businesses to participate depends on their perceptions of anticipated burden. These perceptions are based on the information they are provided. In web surveys, for instance, the advertised interview length is negatively associated with response rates (Galesic and Bosnjak 2009; Yan et al. 2010a) and it is positively associated with break-offs (Galesic 2006). A similar result is found for mail surveys (Edwards et al. 2002). In the case of longitudinal surveys, the expected burden is also determined by respondents' prior experience of the survey. For instance, Bergman and Brage (2008) find that respondents with a negative experience are less willing to accept new survey requests. Researche has shown that other factors may mediate or worsen this initial perception. For instance, interest in the survey topic is found to be an important contributor to the decision to participate (Groves et al. 2006). Moreover, a positive opinion the sponsor also generally acts as a mediator of the estimated burden and increases the probability of participation. Surveys sponsored by government agencies have higher response rates than surveys sponsored by non-government agencies (Presser et al. 1992). Tomaskovic-Devey et al. (1994) suggest that the expected burden is also related to the overall context. For example, the organizational practices and divisions of labor and information may facilitate or inhibit the assembly of relevant knowledge to reply adequately to survey requests. Establishments with greater performances and financial resources are more likely to have the organizational slack to complete a survey.

When respondents start filling in the questionnaire, their initial perceptions of burden may change because the questions prove to be more (or less) complex than expected or because of changes in other factors relating to the overall context. Even if a business has

decided to participate in the survey, an increase in perceived burden may result in decisions to break off, to skip a question or to provide poor quality data (Yan et al. 2014). One factor that may contribute to this situation is the complexity of the questionnaire. Long questions or a high number of response options taking longer to answer are more prone to response order effects (Holbrook et al. 2007; Galesic and Bosnjak 2009) and are more likely to induce break-offs (Peytchev 2011) or item nonresponse (Yan et al. 2010b) and less reliable responses (Tourangeau et al. 2019). Other articles focusing on web surveys find that higher dropout rates are associated both with the formal characteristics of questions, for example, the position of questions, the number of questions on a page and poor visual design, and with the characteristics of the respondents (Crawford et al. 2001; Heerwegh and Loosveldt 2002; Galesic 2006). Bavdaž (2010) suggests a model (later expanded by Haraldsen 2013) that links the difficulty to retrieve the requested information to the quality of the responses. The likely outcomes range from exact information to item nonresponse. The alternatives in between are approximations, solid or rough estimates and blunders. Contextual aspects may play a role too. Organizations insulated from their environment and in unregulated environments may have little interest in disclosing information (Tomaskovic-Devey et al. 1994). On the contrary, businesses such as publicly traded firms that are dependent on their environment for resources generally have a higher motive to respond. The current study tests two hypotheses, drawn from this conceptual framework. The first one is that the measurement of perceived difficulty is necessary for data producers since it enables the capture of different information from the one contained in the measures of actual burden (relating time and money). The second assumption is that perceived burden has a direct effect on several dimensions of data quality (such as the propensity to participate in the survey, to provide all the information requested and to give accurate answers) even after controlling for actual burden.

3. Data

The Bank of Italy has a long-standing tradition of conducting business surveys aimed to monitor the economic outlook, to study firms' behavior and to assess the effectiveness of economic policy measures.

The main one is the Survey of Industrial and Service Firms (INVIND hereafter), carried out between the end of January and mid-May, which gathers information on investments, gross sales, the workforce, expectations and other economic variables relating to Italian industrial and service firms with at least 20 employees. It began in 1972 (although the microdata available are those from the wave conducted in 1985) and only covered industrial processing firms with at least 50 workers. From 2002 onward, the sample has consisted of about 4,000 firms, of which around 3,000 belong to the industrial sector and the remaining firms belong to the service sector.

The questionnaire is usually composed of two parts: a core part that collects quantitative information on actual and expected structural characteristics (such as turnover, investments and number of employees) and a monographic section dealing with special topics aimed at a more conjunctural analysis. The core questions are compulsory, in the sense that without a response the whole questionnaire is not considered complete and therefore corresponding data are treated as a unit nonresponse.

It is worth noting that we have access to firm's fiscal identifiers and therefore we have been able to link survey data with other administrative records and in particular with register data on firms' balance sheets. This data linkage is very useful both in order to enrich the survey with some variables not directly measured and to perform ex-post comparison between the same quantities measured from the survey and available from the administrative sources. We use this data linkage in the analysis of measurement error.

A second survey that we use is the Business Outlook Survey of Industrial and Service Firms (BOS), carried out between September and October. It has been conducted since 1993 to respond to short-term economic analysis needs. It mainly collects qualitative information on firms' performance and their future expectations. The target sample size is 4,000 units.

Both surveys are conducted by the local branches of the Bank of Italy. The surveys use a mixed mode approach of self-enumeration (web and interactive pdfs), telephone interviews and face-to-face interviews. In the 2019 wave, the INVIND and BOS surveys achieved the following percentages of questionnaires by mode (INVIND response rate is shown first): web collection (6% and 11%); interactive pdfs (74% and 45%); telephone interviews (13% and 45%); and face-to-face (7% and 1%). Moreover, the participation is voluntary in both surveys.

The two surveys are conducted on approximately the same sample of firms. Those who have participated in past waves (of either survey) and are still in the target population are always contacted for a new data collection wave. A firm can drop out of the sample either because of its choice, because of bankruptcy or because it has fallen below the surveyed size threshold. Larger businesses (with more than 5,000 employees) are always contacted in the next wave. The response rates are around 70% for INVIND and 75% for BOS. A refreshment sample is selected to compensate for attrition to reach the minimum target of 4,000 firms in each wave. As a consequence of this design, the samples used for the two surveys are made of almost the same units: in the period 2015–2019, most firms participated in both surveys in the same year (Table 1).

Our analyses are mainly based on the INVIND survey. We exploit the lower and more stable complexity of the BOS survey for two main purposes. First, we use it for a preliminary investigation on the existence of a 'hardening survey climate' phenomenon in our data. This issue refers to the fact that it is now more difficult to conduct a survey involving people as respondents than it used to be (either in business or household surveys). If this is the case, many results that we observed in our study could be driven by such a phenomenon, rather than by the complexity of the survey. Second, we use the BOS survey to study whether the association between perceived and actual burden across time is different from the one resulting from the INVIND survey.

Table 1. Number of firms participated in INVIND, BOS and in both in the period 2015–2019.

	Year of data collection					
Survey	2015	2016	2017	2018	2019	Total
INVIND only	490	597	491	502	404	2,484
BOS only	552	473	673	504	647	2,849
Both	3,770	3,798	3,717	3,889	3,807	18,981
Total	4,812	4,868	4,881	4,895	4,858	24,314

4. The Measurement of Actual and Perceived Response Burden

Our data enables us to construct five measures of actual burden. The first one is the number of pages of the questionnaire. This measure reflects not only the number of questions and their length but also the need to provide instructions or clarifications to help respondents. The second one is the number of fields to fill in (the number of variables in each questionnaire). In case of multiple choice questions, each response option is considered as a different field, since respondents have to read it, think about it and possibly retrieve the necessary information. Both measures have been reconstructed for both surveys since 1985.

Figure 1 shows the evolution of these two measures over time. The complexity of the INVIND questionnaire has increased overall from 1985 to 2019. In particular, the number of variables more than tripled during the period, with greater growth occurring between the late 1990s and the early 2000s. Since then, the total number of variables in each questionnaire has remained stable, while the number of pages has increased considerably: it has grown more than two times since 2009 and seven times since the beginning of the survey. The marked increase in 2010 was due to the positioning of the instructions for the respondents below each question instead of in a separate document as previously done. This change increased the length of the questionnaire by about five pages in that year alone. The BOS questionnaire has also undergone similar changes, albeit more gradually, particularly in terms of the number of variables (Figure 1).

Starting from the 2017 wave, the INVIND survey also collects information on the number of people who have contributed to the completion of the survey, whether or not external consultancy was required and the time spent filling in the questionnaire. Unlike the previous indicators, these are individual-specific and capture the difficulty personally faced by each respondent.

Our measure of perceived burden is based on a simple question that can be translated as follows: "How would you rate the level of complexity of the survey?". Response options range from 1 ("low") to 4 ("excessive"). This question has been present since 2004 for the INVIND survey and since 2010 for the BOS survey. Starting from the 2017 wave, respondents are also asked to rate, with a score from one to ten, the contribution of five

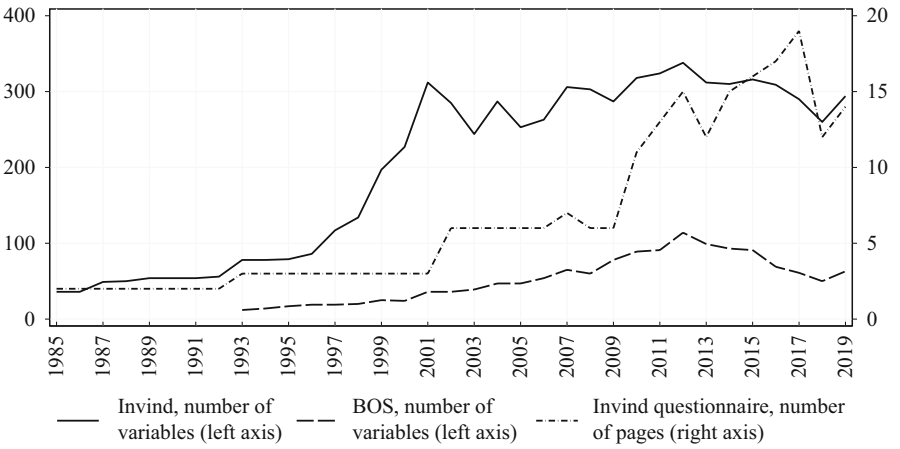

Fig. 1. The complexity of INVIND and BOS questionnaires over time.

factors to response burden, including (1) too many questions, (2) several people involved in answering the questions, (3) the use of unclear terms, (4) not exhaustive response options, and (5) difficulties in choosing the correct answer. Table 2 shows the 2017 to 2019 responses to these questions. Across the years, firms that perceived an excessive level of difficulty indicated higher ratings for the contributing factors of questionnaire length and the need to involve more people to obtain the required data.

Figure 2 shows the dynamics of the perceived response burden in the two surveys. Over the period, the percentage of firms reporting a "high" burden in the INVIND survey is always greater than 35%, but in a cyclical pattern. In the BOS survey, such a percentage decreases from about 15% in 2012 to around 5% in 2019. This descriptive evidence suggests two main results. First, the perceived burden moves quite a lot with the actual burden represented here by the number of variables in the questionnaire. The association between the measures of perceived and actual burden is further confirmed when considering the individual-specific measures of actual burden (the number of people involved, the percentage of external consultants and the time spent). The higher the perceived burden, the higher the actual burden (Table 3).

Second, we don't find evidence to support the phenomenon of "hardening survey climate" according to which the complexity in conducting surveys increases over time, regardless of the survey's characteristics. Figure 2 suggests that this phenomenon is not particularly relevant in the case of Italian firms. In the BOS survey, the perception of effort does not increase over time, even if the businesses participating in the survey are almost the same as those participating in the INVIND survey. This suggests that perceived burden is mainly affected by specific idiosyncratic survey factors. In any case, the econometric analyses performed below will take this phenomenon into account, whenever possible, by using a full set of yearly dummies that allows us to control for any possible time fixed effect.

It is worth stressing that since our measures of actual and perceived burden have been collected starting in different years, we will be using different datasets depending on the

Table 2. Average score for each factor disaggregated by the perceived response burden (INVIND 17–19).

Response burden	Too many questions	More people involved	Use of unclear terms	Not exhaustive response options	Difficulties in choosing the answer
			2017		
Low	3.2	2.3	2.1	2.1	2.1
Average	4.8	4.2	2.9	3.0	3.2
High	7.0	6.1	4.0	3.4	4.0
Excessive	8.3	7.2	4.9	4.0	4.7
			2018		
Low	2.6	2.1	1.8	1.9	1.9
Average	4.4	4.0	3.2	3.0	3.2
High	6.6	6.0	4.3	3.4	4.3
Excessive	8.5	7.1	5.3	3.9	5.3
			2019		
Low	3.0	2.4	2.1	2.2	2.1
Average	4.7	4.2	3.2	3.3	3.5
High	6.9	6.1	4.3	3.8	4.5
Excessive	8.6	6.8	5.2	4.3	4.9

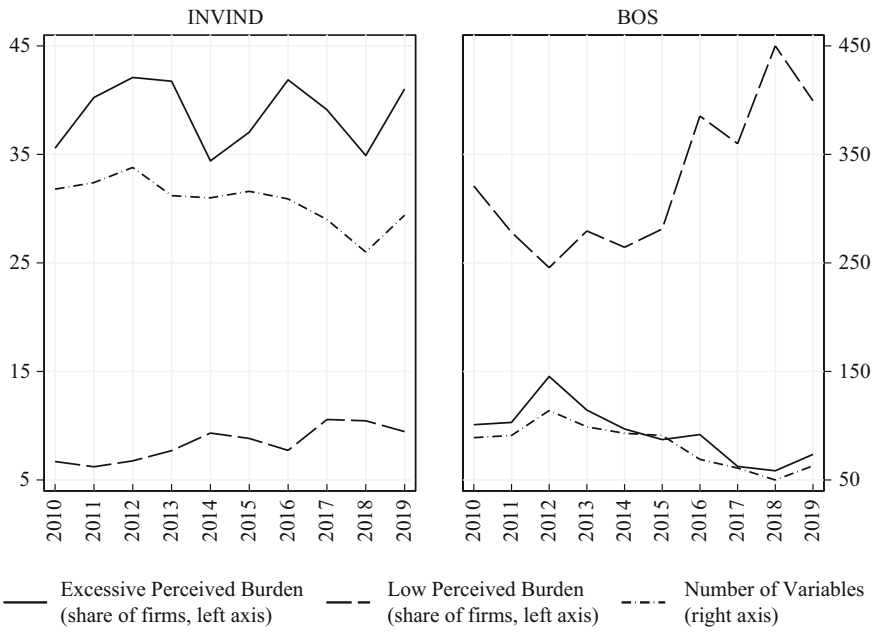

Fig. 2. Firms' perceived burden and number of variables over time.

Table 3. Average people, external consultant and time per level of perceived response burden.

	Average number of people involved			
Perceived response burden	2016	2017	2018	Total
Low	1.5	1.4	1.4	1.5
Average	2.4	2.3	2.4	2.3
High	3.4	3.0	3.1	3.1
Excessive	3.6	3.2	3.1	3.3
	Firms using external consultant (%)			
Perceived response burden	2017	2018	2019	Total
Low	16	14	14	15
Average	23	22	22	22
High	29	34	31	32
Excessive	45	41	41	42
	Average completion time (hours)			
Perceived response burden	2017	2018	2019	Total
Low	1.8	2.1	2.2	2.0
Average	3.7	3.6	3.6	3.6
High	6.4	5.3	5.3	5.7
Excessive	8.2	7.5	5.9	7.1

purpose of the analysis. Yet the two main datasets refer to the period 2004–2019, when only actual burden is studied, and to the period 2017–2019 when also the perceived burden is considered.

To better evaluate the relationship between the perceived response burden and the various factors that can affect it, we create a dummy variable equal to 1 if the perceived response burden is "high" or "excessive" and 0 otherwise. We then estimate two different logistic regressions (as a robustness check). The first one is run on waves from 2004 to 2019 for which only the measures of actual burden we have reconstructed are available. We use the following covariates (Table 4):

Table 4. Probability of reporting a high ('elevated' or 'excessive') perceived response burden (logit model).

	(1)	(2)	(3)
N. of variables	1.004**	1.004***	1.004***
	(0.001)	(0.001)	(0.001)
N. of pages	1.039***	1.040***	1.047***
	(0.007)	(0.008)	(0.008)
Internal instruct.	0.786*	0.796*	0.776*
	(0.084)	(0.087)	(0.086)
N. of quantitatives	0.999	0.999	0.999
	(0.001)	(0.001)	(0.002)
N. of waves	0.936***	0.929***	0.933***
	(0.007)	(0.007)	(0.007)
N. of waves^2	1.002***	1.001***	1.001***
	(0.000)	(0.000)	(0.000)
log(empl)		1.198***	1.059
		(0.025)	(0.061)
Δturnover$_t$		0.947	0.939
		(0.046)	(0.048)
Δemployment$_t$		0.954	0.999
		(0.112)	(0.118)
$\frac{\text{investment}_{t-1}}{\text{turnover}_{t-1}}$		0.862	0.890
		(0.079)	(0.073)
$\frac{\text{investment}_t}{\text{turnover}_t}$		1.063	1.077
		(0.099)	(0.097)
50–199 empl. x N. of waves		1.029***	1.000
		(0.004)	(0.006)
200–499 empl. x N. of waves		1.022***	1.014*
		(0.006)	(0.006)
$\geq$500 empl. x N. of waves		1.014*	1.018**
		(0.007)	(0.007)
Δper capita Val. Add. Sett.		1.830*	1.433
		(0.500)	(0.401)
Constant	0.172***	0.0905***	0.120***
	(0.065)	(0.035)	(0.052)
Observations	5905	59028	59028
Pseudo R^2	0.005	0.013	0.022

Odds ratios. Standard errors in parentheses. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Based on INVIND 2004–2019 waves. Column 3 includes dummies for industry, area and size class. Missing cases to the question on perceived response burden are excluded.

- indicators of the questionnaire's complexity: the total number of variables and pages, the number of questions requiring quantitative information and whether or not the instructions were placed inside the questionnaire;
- the total number of waves (for both the BOS and INVIND) in which firms have participated;
- firm characteristics such as firm size, the sector of activity, indicators of performance such as the variation in turnover and employment, the ratio between investments and turnover and the growth rate of the (per capita) annual sectorial value added.

The first set of variables refers to decisions about the survey design that are under the Bank of Italy's full control, while firm characteristics are mainly used as control variables that should account for the overall context. It is worth stressing that the indicators of the questionnaire's complexity do not vary among respondents in a given wave, they only change over time. To account for the possible confounding effects of other time-varying factors, we include the growth rate of the (per capita) annual sectorial value added as a proxy of the general economic outlook. This variable should capture possible effects of the economic cycle on the perceived response burden.

The second logistic regression model is based on waves from 2017 to 2019, for which we have information on actual burden directly collected from respondents. We model perceived burden as a function of Table 5:

- three individual-specific measures of actual burden: the number of people involved in the survey, completion time, and whether or not the firm has used external consultancy. In particular, for the first two variables we have created a dummy variable equal to 1 if the value is greater than the 75th percentile of the corresponding variable, 0 otherwise;
- the total number of waves (for both the BOS and INVIND) in which firms have participated;
- the same firm's characteristics used in the previous model;
- time dummy variables and their interaction with all the above mentioned variables.

It is worth noting that even if questions on perceived and actual burden are placed in the same section at the end of the questionnaire, the item nonresponse for the former is around 10%, while it is about 18% for the latter. Since the decision to skip the questions on actual burden is probably an indicator of excessive burden endured by respondents, we include in all the regressions three dummy variables taking the value 1 if the business has not answered to such questions. Our results may be summarized as follows.

An increase in the total number of pages in the questionnaire is significantly linked to an increase in the probability of a large or excessive perceived response burden, with the former playing a stronger role than the latter. This result can be explained by the fact that firms may download the template of the questionnaire, scroll through it and look at the number of total pages before starting to fill it in. The number of pages is probably the most immediate measure that firms use to anticipate their burden. As expected, the number of variables collected in the survey is also positively associated with a higher perceived burden.

Table 5. Probability of reporting a high ('elevated' or 'excessive') perceived response burden (logit model).

	(1)	(2)	(3)
People inv.(> 75th)	2.056***	2.023***	1.945***
	(0.170)	(0.176)	(0.170)
People inv. miss	1.129	1.125	1.037
	(0.265)	(0.263)	(0.250)
External Cons: Y	1.642***	1.657***	1.711***
	(0.132)	(0.137)	(0.146)
External Cons miss	1.698*	1.688*	1.668*
	(0.411)	(0.410)	(0.415)
Completion time (> 75th)	2.367***	2.345***	2.360***
	(0.209)	(0.208)	(0.212)
Completion time miss	1.547**	1.552**	1.666***
	(0.229)	(0.231)	(0.249)
N. of waves	0.945***	0.941***	0.944***
	(0.013)	(0.013)	(0.014)
N. of waves^2	1.001*	1.001*	1.001*
	(0.000)	(0.001)	(0.001)
log(empl)		1.015	0.789
		(0.055)	(0.109)
Δturnover$_t$		0.891	0.891
		(0.091)	(0.088)
Δemployment$_t$		1.108	1.161
		(0.280)	(0.294)
$\frac{\text{investment}_{t-1}}{\text{turnover}_{t-1}}$		0.954	0.962
		(0.190)	(0.147)
$\frac{\text{investment}_t}{\text{turnover}_t}$		0.897	0.920
		(0.250)	(0.255)
50–199 empl. x N. of waves		1.011	0.991
		(0.008)	(0.011)
200–499 empl. x N. of waves		1.004	0.994
		(0.011)	(0.012)
$\geq$500 empl. x N. of waves		0.997	0.973*
		(0.014)	(0.013)
Constant	0.423***	0.407***	1.131
	(0.038)	(0.096)	(0.575)
Observations	11846	11842	11842
Pseudo R^2	0.081	0.082	0.094

Odds ratios. Standard errors in parentheses. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Column 2 includes dummies for industry, area and size class; Column 3 adds time dummies, also interacted with all the dummies in column 2. Based on INVIND 2017–2019 waves. Missing cases to the question on perceived response burden are excluded.

Furthermore, we find that an excessive number of people involved and an excessive amount of time (greater than the respective 75th percentiles) as well as the use of external consultants increase the probability of reporting a high response burden. Moreover, all the three dummy variables indicating a nonresponse to the questions about actual burden are positively linked with the probability of declaring a higher perceived burden. This result may be explained by the fact that the respondent feels so stressed by the survey, that

she(he) prefers not to waste additional time responding the questions on actual burden (that are placed at the end of the questionnaire).

Interestingly, we do not find evidence that the number of quantitative variables is associated with perceived burden. This is probably because quantitative questions in the INVIND questionnaire mainly concern balance sheet data that are probably not so difficult for the respondent to obtain. On the other hand, the inclusion of instructions within the questionnaire is negatively correlated with the score assigned to the perceived burden: respondents may not feel like consulting a separate sheet of paper and therefore a question without explanation can be considered more burdensome, and, even if they do, this may take longer and more effort than having questions accompanied by corresponding instructions.

Perceived response burden is also associated with other factors. For instance, larger firms tend to report a high burden, especially when the number of surveys they have participated in increases. This result may seem to be counterintuitive since large-sized businesses are supposed to have good documentation systems and people who respond to surveys as part of their job. One possible explanation is that in the INVIND survey there is a big effort to enroll larger businesses; firms with more than 5,000 employees are always eligible to be included in the target population, even if they have refused to do so in the past. Moreover, large businesses are always selected in the business surveys conducted by national statistical offices or by other statistical agencies. Since they receive frequent enquiries and since they cannot easily refuse participation, they are likely to manifest their dissatisfaction by declaring a high burden.

The predicted probability of observing a high response burden, obtained by the first estimated logistic regression (Table 4), has a "U-shaped" relationship with the number of waves in which firms have participated (Figure 3): it initially decreases as the total number of

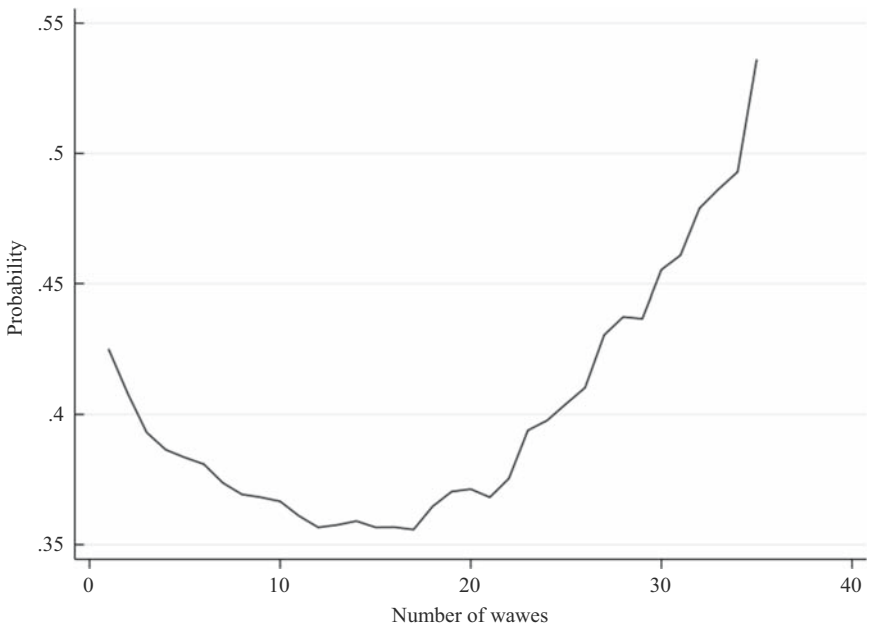

Fig. 3. Predicted probability of reporting a higher response burden obtained by the estimated logistic regression in Table 4.

waves increases but subsequently after around ten waves, it starts to rise. The initial decrease may be the result of two things. First, it could be due to a self-selection process: firms who find the questionnaire burdensome may decide to drop out the following year. As a result, the sample consists of collaborative firms that don't think that participation requires an excessive effort. Second, the decrease could reflect the presence of a learning process over multiple waves that makes it easier for firms to complete the questionnaire. However, the reduction in the response burden is less intense as the frequency of participation increases, meaning that, at some point, a certain level of stress could take over, thereby raising the burden.

A final point is worth mentioning. Even if we find a positive association between actual and perceived burden, the measures of actual burden explain only a small fraction of the total variability of the perceived burden as it is indicated by the low values of the Pseudo R-square indicator. The main reason is that, though we are using a very simple and coarse measure of perceived difficulty, this can capture many unobserved positive and negative effects relating survey participation, such as the respondent's interest in the topic, their ability to answer the questions or their opinion in the utility of the survey. Also, some unobserved factors relating to business activity, such as the internal organization or the documentation system may affect respondents' perceptions.

5. Perceived Response Burden and Data Quality

Data quality can be defined as "fitness for use" of statistical information. The European statistical system which provides guidelines for all European national statistical institutes, defines it as the result of eight dimensions: relevance, accuracy, timeliness, punctuality, accessibility, clarity, comparability and coherence (European Commission. Statistical Office of the European Union 2014).

In this article, we focus on data accuracy, that is, the degree to which the information correctly describes the phenomena it was designed to measure.

One of the main risks of a high response burden concerns the firms' decision not to respond. They may decide not to participate in the survey at all (unit nonresponse), or they may employ some response strategies that allow them to reduce the effort that they have to expand without leaving the survey altogether. Some of these strategies include skipping some questions (item nonresponse), using the "don't know" or "no opinion" response options, or choosing the first reasonable response. Other behaviors include speeding through the survey by giving low-effort responses or not fully answering open-text questions. In the following subsections, we provide some empirical evidence of how these aspects affect data accuracy using the INVIND survey.

5.1. Unit Nonresponse

Since BOS and INVIND are two longitudinal surveys, we can study the association between perceived response burden and attrition.

Descriptive analysis shows that the propensity to leave the panel significantly increases for firms that perceived a high response burden in previous surveys. In particular, from 2017 to 2019, 9% of businesses stating a low level of burden for INVIND do not participate in the BOS conducted in the same year; compared to 14% that stated an excessive difficulty rating. Similarly, around 16% of the firms declaring a low burden in

BOS refuse to participate in the following the INVIND survey (conducted a year later). This percentage rises to 21% for businesses that complained about excessive effort.

To further investigate the role of response burden on attrition, we run a logistic regression using as a dependent variable a dummy indicating whether firms that participated in waves 2016–2018, then decided to drop out in the subsequent surveys (2017–2019 respectively). We regress this variable on the perceived response burden declared in the last survey before dropping out, indicators of actual burden (the number of people involved, the percentage of external consultants and the time spent), a dummy variable for nonresponse at the question on perceived burden, the total number of waves in which firms have participated, as well as their characteristics. (Table 6).

The results show a significant increase in the probability of attrition for firms that report an excessively high response burden and for those that do not answer the question on response burden. The number of waves they have participated in plays a significant role in reducing the probability of attrition, confirming the existence of possible learning effects (as already shown in Figure 3).

Moreover, the variation in the number of employees that we use as a proxy of economic performance has a significant effect in reducing the probability of attrition. This may in part be due to the fact that firms with better performance have a lower probability of dropping out of the survey (a similar result is shown in D'Aurizio and Papadia 2019).

Finally, we find that including the three individual-specific indicators of actual burden in the regression doesn't affect the estimate of parameters: columns 3–5 of Table 6 show similar estimates as the remaining ones (columns 1–2). Moreover, the actual burden faced by each respondent is not significantly correlated with the probability of leaving the panel once controlled for the perceived burden and the other observables. One possible explanation is that when the respondent has been contacted again, he or she remembers the stress induced by taking part in the previous year's survey rather than its actual difficulty, and this affects the decision whether or not to participate in the new round of the survey. Hence, the actual burden would affect the probability of attrition only through the perceived burden.

5.2. Item Nonresponse

The response behavior of survey participants depends on many factors. Some respondents may decide to answer hastily and carelessly, since they perceive the survey as too time-consuming. In other cases, they may limit their attention to questions that are mandatory for completing the questionnaire (for which an alert signals the invalidity of the entire questionnaire if they are left blank). These behaviors may become more likely as the complexity of the questionnaire grows.

Figure 4 shows the cumulative distributions of firms according to their share of missing items in the 1999 and 2017 INVIND waves. In these two years, the questionnaires present very different levels of complexity: the number of variables increased from about 200 to about 300, the number of pages from about 5 to about 20. To make the two distributions comparable, we selected only industrial processing firms with more than 50 employees and excluded the compulsory variables. In the 1999 survey, 90% of businesses had a share of item nonresponse lower than 35%. In the 2017 survey this share rises to 60%. Moreover,

Table 6. Probability of attrition (logit model).

	Attrition in 2016–2018		Attrition in 2017–2018		
	(1)	(2)	(3)	(4)	(5)
High Perc. Burd.	1.425***	1.286**	1.432**	1.488***	1.389**
	(0.133)	(0.123)	(0.165)	(0.173)	(0.164)
Perc. Burd. miss.	2.308***	1.922***	2.083***	1.679*	1.421
	(0.319)	(0.274)	(0.381)	(0.374)	(0.329)
People inv.(> 75th)				0.842	0.837
				(0.123)	(0.127)
People inv. miss				1.335	1.370
				(0.408)	(0.434)
External Cons: Y				1.136	0.970
				(0.143)	(0.127)
External Cons miss				1.436	1.344
				(0.530)	(0.507)
Completion time (> 75th)				0.809	0.842
				(0.122)	(0.128)
Completion time miss				0.641	0.667
				(0.207)	(0.213)
N. of waves		0.884***			0.895***
		(0.016)			(0.020)
N. of waves^2		1.003***			1.002**
		(0.001)			(0.001)
log(empl)		0.961			1.193
		(0.179)			(0.282)
$\Delta\text{turnover}_t$		0.813			0.584+
		(0.194)			(0.172)
$\Delta\text{employment}_t$		0.434+			0.504
		(0.192)			(0.284)
$\frac{\text{investment}_{t-1}}{\text{turnover}_{t-1}}$		1.132			1.545
		(0.211)			(0.588)
$\frac{\text{investment}_t}{\text{turnover}_t}$		0.942			0.681
		(0.230)			(0.258)
Constant	0.152***	0.281+	0.151***	0.155***	0.112**
	(0.009)	(0.187)	(0.011)	(0.015)	(0.092)
Observations	12758	12756	8445	8445	8443
Pseudo R^2	0.011	0.044	0.009	0.013	0.045

Odds ratios; Standard errors in parentheses. $+p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Columns 2 and 5 include dummies for time, industry, area and size class.

firms that answer at least half of the non-compulsory questions decreases by about 20 percentage points over the same period.

Item nonresponse rates vary across different sections of the questionnaire. By informally inspecting the questionnaire content, we find that sections requiring qualitative information on easy-to-understand topics (such as the section about funding) get relatively low nonresponse rates. These rates increase when the complexity of the formulation of the question and of the terminology used seems to grows.

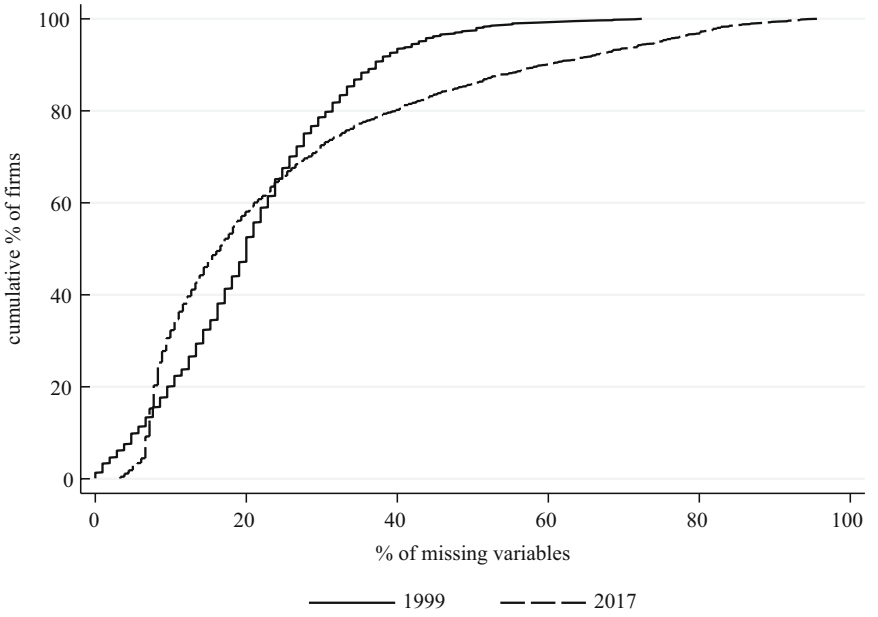

Fig. 4. Cumulative distributions of firms for shares of missing variables in INVIND 1999 and 2017.

The increase in the share of missing variables also involves questions that are of great importance for economic analysis. Figure 5 shows the time series of the percentage of missing data on questions about expectations on investment and turnover for the next year. Although their complexity has remained constant over the years in terms of formulation, the share of missing answers about investment plans reaches 15% in 2019, compared with less than one-third of that figure in the 1980s; the share related to expectations on turnover (collected since 1997) rises from about 3% to almost 10%. This implies that the response burden can involve the potential loss of information on historical and important variables, as well as on new ones.

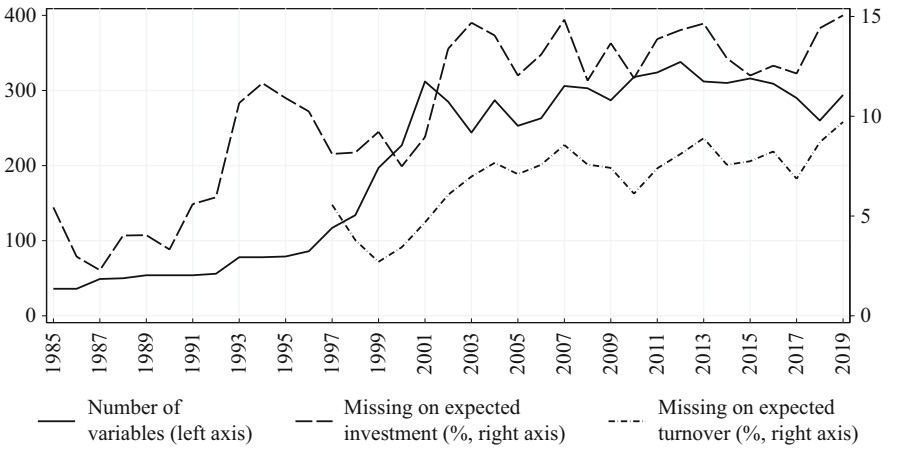

Fig. 5. Share of missing data on questions about expected investment and expected turnover.

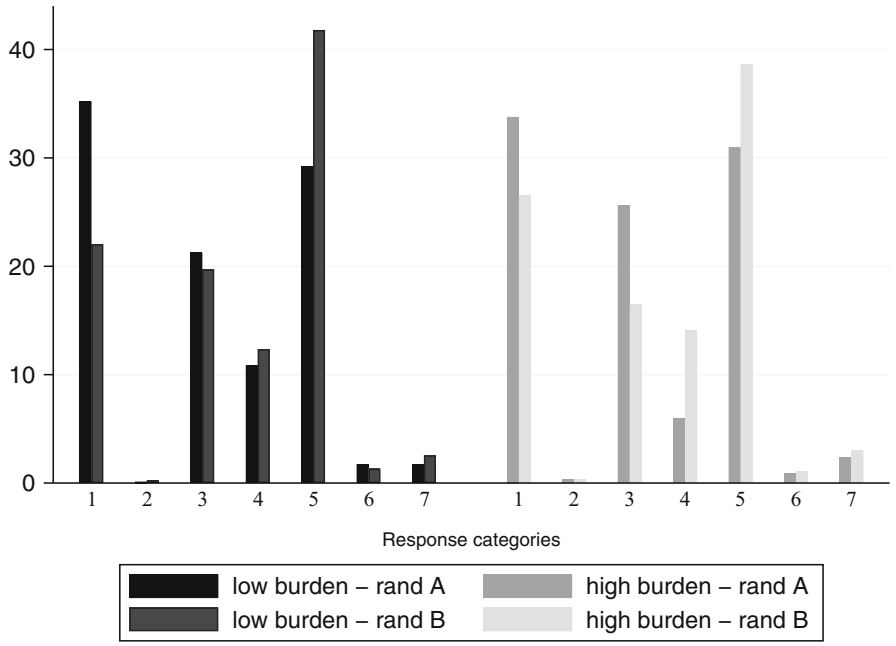

Fig. 6. Distribution of factors affecting the expected price dynamics of firms. Randomized experiment with reversed response categories (percentages).
Note: Response categories are labeled as: 1 = total demand; 2 = change in the financial burdens borne by the firm; 3 = competitors' prices; 4 = labor costs; 5 = raw materials prices; 6 = expectations for exchange rates; 7 = inflation expectations. For group A the response options are ordered as above. For group B the order is reversed. INVIND 2017 waves.

As further analysis, we estimate two regression models for the probability of not answering the questions about expected turnover and investment as a function of the perceived response burden, our measures of actual burden, some firm characteristics, indicators of firm's performance and time dummies (Table 7). The probability of nonresponse is significantly greater for firms that perceive an excessive response burden and is even greater when the respondent does not even provide answers to questions about the evaluation of the questionnaire. A missing value to such questions (which are located at the end of the questionnaire) may be an indication that the respondent thinks that the survey is too demanding. We also find that the longer the completion time, the higher the probability of item nonresponse.

We also find that perceived burden plays a much more important role in explaining nonresponse than actual burden. The more significant variables are those indicating that the respondent has not answered the questions about the time needed to complete the survey and the number of people involved.

5.3. Measurement Error

Our data also enable us to provide some evidence of the association between perceived burden and accuracy.

We use the CEBIL/CERVED company database that provides information on the balance sheet of all the joint stock companies operating in Italy. We link this database to

Table 7. Probability of item nonresponse (logit model).

	Nonresponse on expected investment		Nonresponse on expected turnover	
	(1)	(2)	(3)	(4)
High Perc. Burd.	1.230***	1.328***	1.330***	1.503***
	(0.033)	(0.082)	(0.047)	(0.121)
Perc. Burd. miss.	3.077***	1.677***	3.949***	2.908***
	(0.097)	(0.166)	(0.151)	(0.343)
People inv.(> 75th)		0.976		0.950
		(0.077)		(0.100)
People inv. miss		1.779***		1.554*
		(0.284)		(0.321)
External Cons: Y		0.838*		0.802
		(0.070)		(0.095)
External Cons miss		0.902		0.823
		(0.147)		(0.170)
Completion time (> 75th)		0.792**		0.703**
		(0.067)		(0.080)
Completion time miss		2.119***		2.346***
		(0.226)		(0.314)
log(empl)	1.213***	1.478***	1.251***	1.453***
	(0.042)	(0.114)	(0.050)	(0.131)
$\Delta\text{turnover}_t$	0.995	0.927	0.804**	1.008
	(0.010)	(0.123)	(0.062)	(0.153)
$\Delta\text{employment}_t$	1.022	0.640	0.623**	0.459*
	(0.077)	(0.167)	(0.092)	(0.168)
$\frac{\text{investment}_{t-1}}{\text{turnover}_{t-1}}$	1.014	1.090	1.034	1.170
	(0.020)	(0.096)	(0.026)	(0.285)
$\frac{\text{investment}_t}{\text{turnover}_t}$	1.143**	0.933	1.019	0.638
	(0.057)	(0.197)	(0.043)	(0.250)
Constant	0.0687***	0.0197***	0.0315***	0.0108***
	(0.009)	(0.006)	(0.005)	(0.004)
Observations	66861	12806	66861	12806
Pseudo R^2	0.041	0.095	0.075	0.154

Odds ratios; Standard errors in parentheses. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Column 2 includes dummies for time, industry, area and size class. Cols (1) and (3) are based on INVIND 2004–2019 waves; Cols (2) and (4) are based on INVIND 2017–2019 waves.

the INVIND survey and compute two proxies of response error for each business. The proxies are the difference (in absolute terms) between the value of turnover and investments reported in the survey and the corresponding values resulting from the administrative records (with the same reference year).

We then run a regression using these proxies as dependent variables and include perceived burden, actual burden and other controls (firm size, measures of performance, sector of activity and location) as covariates (Tables 8 and 9).

We find that the response error does not seem to be affected by perceived response burden. A high perceived burden is positively associated with the response error on

Table 8. Response error in turnover (linear model).

	(1)	(2)	(3)	(4)	(5)
High Perc. Burd.	272.9***	187.4	110.6	122.1	128.9
	(72.583)	(167.842)	(149.348)	(144.143)	(155.375)
Perc. Burd. miss.	223.9*	-62.47	-290.3	-394.5	-399.2
	(113.277)	(114.804)	(266.805)	(308.197)	(309.690)
People inv.(> 75th)			352.8*	399.5*	423.7*
			(161.699)	(159.246)	(185.002)
People inv. miss			347.0	-100.9	-218.7
			(286.148)	(209.664)	(214.320)
External Cons: Y				-170.4	-154.9
				(148.217)	(163.175)
External Cons miss				519.5	372.9
				(310.343)	(337.895)
Completion time (> 75th)					-88.87
					(217.269)
Completion time miss					275.0
					(231.697)
Part of a group	783.1***	508.3**	475.7**	462.6**	467.9**
	(87.598)	(165.805)	(155.791)	(160.547)	(164.001)
Δturnover$_t$	12.33	-4.616	-5.278	-5.492	-5.587
	(13.671)	(22.346)	(22.317)	(22.412)	(22.265)
Δemployment$_t$	-237.3	111.1	138.3	151.8	155.9
	(353.268)	(828.462)	(827.458)	(833.752)	(828.957)
log(empl)	1128.9***	900.1***	870.2**	866.0**	876.2**
	(121.970)	(265.727)	(274.881)	(274.458)	(268.983)
$\frac{\text{investment}_{t-1}}{\text{turnover}_{t-1}}$	-53.81	36.99	41.39	28.69	31.83
	(34.949)	(79.125)	(80.716)	(78.115)	(76.291)
$\frac{\text{investment}_t}{\text{turnover}_t}$	307.5	-40.29	-56.10	-50.34	-57.17
	(196.377)	(161.974)	(159.216)	(166.366)	(162.606)
Constant	-2893.7***	-2551.7**	-2560.6**	-2509.2**	-2552.2**
	(460.174)	(928.094)	(934.283)	(923.869)	(912.223)
Observations	59598	11462	11462	11462	11462
R^2	0.091	0.121	0.121	0.122	0.122

The dependent variable is computed as the absolute value of the difference between the value declared in the survey and the one from administrative records. Amounts in thousands of euros. Standard errors in parentheses. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. All regressions include dummies for time, industry, area and size class. Col (1) based on INVIND 2004–2017 waves, cols (2–5) INVIND 2017–2019 waves.

turnover and investments only if we consider past waves for which measures of actual burden are not available (column 1). Its effect loses significance once the analysis is limited to the last waves (column 2). Moreover, controlling also for the three proxies of actual response burden shows that the average response error is associated with some indicators of actual burden (such as the number of people involved or the time required for completion). Therefore, it seems that it is the actual arduousness of the survey that affects the measurement errors rather than the one subjectively perceived. For example, involving more people means fragmentizing the questionnaire in multiple parts and probably sending only that part to the person. This can have two implications: first, different

Table 9. Response error in investment (linear model).

	(1)	(2)	(3)	(4)	(5)
High Perc. Burd.	33.86	49.63	31.20	24.25	41.02
	(26.335)	(53.001)	(55.327)	(55.946)	(56.636)
Perc. Burd. miss.	111.3**	29.92	-53.58	-90.45	-79.65
	(42.006)	(94.303)	(118.245)	(121.099)	(121.294)
People inv.(> 75th)			78.27	69.53	102.9
			(59.647)	(64.942)	(67.525)
People inv. miss			115.5	-37.75	-17.36
			(87.413)	(173.223)	(178.286)
External Cons: Y				37.90	53.96
				(67.706)	(68.004)
External Cons miss				197.6	207.8
				(177.897)	(182.913)
Completion time (> 75th)					-133.6*
					(60.703)
Completion time miss					-51.32
					(101.090)
Part of a group	113.2***	94.22	87.38	89.47	93.66
	(28.403)	(58.214)	(58.396)	(58.963)	(58.947)
Δturnover$_t$	18.54***	21.41**	21.24**	21.29**	21.09**
	(3.020)	(7.087)	(7.079)	(7.118)	(7.201)
Δemployment$_t$	-738.9***	-842.6**	-835.6**	-838.8**	-832.3**
	(107.365)	(266.941)	(266.586)	(268.185)	(271.272)
log(empl)	50.95	-144.7	-151.0	-150.5	-143.1
	(48.359)	(107.358)	(107.385)	(106.954)	(106.991)
$\frac{\text{investment}_{t-1}}{\text{turnover}_{t-1}}$	-53.28	-49.65	-47.88	-52.99	-58.60
	(37.132)	(128.049)	(130.444)	(126.079)	(124.371)
$\frac{\text{investment}_t}{\text{turnover}_t}$	820.6***	1465.0***	1460.7***	1470.8***	1478.5***
	(113.049)	(313.774)	(315.787)	(308.416)	(307.242)
Constant	808.3***	961.2*	951.5*	936.9*	924.7*
	(181.330)	(391.829)	(392.301)	(386.261)	(386.490)
Observations	59598	11462	11462	11462	11462
R^2	0.069	0.066	0.067	0.067	0.068

The dependent variable is computed as the absolute value of the difference between the value declared in the survey and the one from administrative records. Amounts in thousands of euros. Standard errors in parentheses. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. All regressions include dummies for time, industry, area and size class. Col (1) based on INVIND 2004–2017 waves, cols (2–5) INVIND 2017–2019 waves.

respondents will answer questions probably without having the corresponding instructions and the awareness of the complete task; second, the main respondent, that is, the one answering the evaluation final part, probably feels less stressed since they have accomplished only a residual part of the questionnaire.

We also find that response error is positively correlated with some characteristics of the business. For instance, as far as turnover is concerned, the error is greater for firms that are part of a group and for large businesses. In these cases, the presence of multiple establishments or multiple companies that are strictly related may lead to ambiguity concerning which unit to respond for.

6. Concluding Remarks

In this article, we have provided some empirical evidence of the relationship between perceived (the difficulty rating) and actual burden and of the association between perceived burden and data quality. Drawing from a unique and rich dataset, we have been able to conduct our analysis also controlling for other contextual factors relating to firms' characteristics.

Our key findings may be summarized as follows.

- The perceived difficulty in completing the survey is associated with measures of actual burden such as the complexity of the questionnaire. We find that it is not simply the number of questions that increases perceived burden, but also the number of pages of the questionnaire, which is probably used by businesses to anticipate their effort (regardless of the effective difficulty of the questions). We also find that the higher the number of people involved in the survey, the higher the perceived burden;
- perceptions are also driven by other firm-specific characteristics such as the number of employees. Large businesses tend to report a high burden. This is also probably because many efforts are devoted to preventing them from dropping out of the survey. Given their difficulties in terms of refusing to participate, they are likely to complain and complain that the questionnaire is burdensome;
- we find empirical evidence that supports our first assumption that the measure of difficulty rating captures different information from the one contained in the measures of actual burden. The latter and other contextual factors capture only a small fraction of the overall variability of perceived burden. The unexplained variability is probably associated with unobservable characteristics such as the respondent's interest in the topic, their ability to answer the questions or their opinion on the utility of the survey. Besides, unobserved factors relating to business activity, such as the organization of the business, may play a role;
- we also find that data quality is directly associated with perceived burden, even after controlling for actual burden and other characteristics (our second assumption). The probability of attrition increases with a higher perceived burden. Moreover, even if the firm participates in the survey, an excessive perceived burden is associated with a high probability that the respondent will not complete the whole questionnaire. The questions that are more likely to be skipped are those that are not compulsory, in the sense that without a response to those questions, the whole questionnaire is considered incomplete (and therefore corresponding data are treated as a unit nonresponse), as well as questions that require more effort to answer (such as the questions about firms' expectations on future investment and turnover). On the other hand we do not find evidence that an excessive difficulty rating is associated with more inaccurate answers, at least as far as turnover or investments are concerned. One possible explanation is that since firms know that their responses can be linked to register data, the best response behavior is to retrieve such information directly from their balance sheets in order to provide consistent information to the outside world.

In summary, our analysis shows that even a simple and coarse measure of difficulty rating, like the one we use in this article, is a good instrument for monitoring data quality.

It is easy to collect and captures many unobserved factors, which play a role in determining the final quality of survey data.

7. Appendix

7.1. Questions on Response Burden

English translation of the questions on response burden:

- How would you rate the level of complexity of the survey?
  1 = modest; 2 = average; 3 = large; 4 = excessive.
- To what extent do you think the following factors made it difficult to fill in the questionnaire?
  (For each factor please assign a score ranging from 1 to 10, where 1 indicates that the factor played a very limited part in making the questionnaire difficult to fill in while 10 indicates that it played a very large part)

	Rate
A Too many questions	☐
B It was necessary to seek the help of several people to answer the questions	☐
C It was not always easy to understand the questions because some of the terms were not clear	☐
D The possible answers did not include my situation	☐
E For some questions, it was difficult to choose the correct answer	☐

- How many people from your firm, including yourself, were involved in completing the survey?
- Was it necessary to involve external consultants (*e.g., accountant, labor consultant, and so on.*)? (Yes/No)
- Could you please indicate how much time approximately it took your firm to collect the necessary information and fill in the questionnaire? (*please indicate the number of hours*)

8. References

Bavdaž, M. 2010. "Sources of Measurement Errors in Business Surveys." *Journal of Official Statistics* 26 (1): 24–42. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/sources-of-measurement-errors-in-business-surveys.pdf (accessed September 2021).

Bavdaž, M., D. Giesen, S.K. Černe, T. Löfgren, and V. Raymond-Blaess. 2015. "Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes." *Journal of Official Statistics* 31 (4): 559–588. DOI: https://doi.org/10.1515/jos-2015-0035. .

Bergman, L.R., and R. Brage. 2008. "Survey Experience and Later Survey Attitudes, Intention and Behavior." *Journal of Official Statistics* 24 (1): 99–113. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/survey-experiences-and-later-survey-attitudesintentions-and-behaviour.pdf (accessed September 2021).

Bradburn, N.M. 1978. "Respondent burden." *In Proceedings of the Survey Research Methods Section of the American Statistical Association*, 35–40. Available at: http://www. asasrms.org/Proceedings/papers/1978_007.pdf (accessed March 2021)

Crawford, S.D., M.P. Couper, and M.J. Lamias. 2001. "Web surveys: Perceptions of burden." *Social Science Computer Review* 19 (2): 146–162. DOI: https://doi.org/10.1177/089443930101900202.

Dale, T., J. Erikson, J. Fosen, G. Haraldsen, J. Jones, and Ø. Kleven. 2007. *Handbook for Monitoring and Evaluating Business Survey Response Burdens*. European Commission. Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/12-HANDBOOK-FORMONITORING-AND-EVALUATING-BUSINESS-SURVEY-RESONSE-BURDEN.pdf/600e3c6d-8e8d-44f7-a8f5-0931c71d9920 (accessed March 2021)

D'Aurizio, L., and G. Papadia. 2019. "Using Administrative Data to Evaluate Sampling Bias in a Business Panel Survey." *Journal of Official Statistics* 35 (1): 67–92. DOI: https://doi.org/10.2478/jos-2019-0004.

Edwards, P., I. Roberts, M. Clarke, C. DiGuiseppi, S. Pratap, R. Wentz, and I. Kwan. 2002. "Increasing response rates to postal questionnaires: systematic review." *BMJ* 324 (7347): 1183. DOI: https://doi.org/10.1136/bmj.324.7347.1183.

European Commission. Statistical Office of the European Union. 2014. *Manuals and guidelines. ESS Handbook for Quality Reports*. DOI: https://doi.org/10.2785/983454.

Fisher, S., and L. Kydoniefs. 2001. "Using a theoretical model of response burden to identify sources of burden in surveys." Paper presented at the 12th International Workshop on Household Survey Nonresponse, September 12–14, Oslo, Norway.

Galesic, M. 2006. "Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics* 22: 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf (accessed September 2021).

Galesic, M., and M. Bosnjak. 2009. "Effects of questionnaire length on participation and indicators of response quality in a web survey." *Public Opinion Quarterly* 73 (2): 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Giesen, D. 2012. "Exploring causes and effects of perceived response burden." In Proceedings of the Fourth International Conference on Establishment Surveys: American Statistical Association, June 11–14, Montréal, Canada. Available at: https://ww2.amstat.org/meetings/ices/2012/papers/302171.pdf (accessed March 2021)

Groves, R.M., M.P. Couper, S. Presser, E. Singer, R. Tourangeau, G.P. Acosta, and L. Nelson. 2006. "Experiments in producing nonresponse bias." *Public Opinion Quarterly* 70 (5): 720–736. DOI: https://doi.org/10.1093/poq/nfl036.

Haraldsen, G. 2004. "Identifying and Reducing Response Burden in Internet Business Surveys." *Journal of Official Statistics* 20 (2): 393–410. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/identifying-and-reducing-response-burdens-in-internet-business-surveys.pdf (accessed September 2021).

Haraldsen, G. 2013. "Quality issues in business surveys." *Designing and Conducting Business Surveys*, 83–125. John Wiley & Sons. DOI: https://doi.org/10.1002/9781118447895.ch03.

Haraldsen, G., and J. Jones. 2007. "Web and paper questionnaires seen from the business respondent's perspective." In Proceedings of the Third International Conference for Establishments Surveys, JUNE 18–21, Montréal, Canada. Available at: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000259.PDF (accessed March 2021)

Haraldsen, G., J. Jones, D. Giesen, and L.-C. Zhang. 2013. "Understanding and coping with response burden." In *Designing and Conducting Business Surveys*, 219–252. John Wiley & Sons. DOI: https://doi.org/10.1002/9781118447895.ch06.

Heerwegh, D., and G. Loosveldt. 2002. "An evaluation of the effect of response formats on data quality in web surveys." *Social Science Computer Review* 20 (4): 471–484. DOI: https://doi.org/10.1177/089443902237323.

Holbrook, A.L., J.A. Krosnick, D. Moore, and R. Tourangeau. 2007. "Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes." *Public Opinion Quarterly* 71 (3): 325–348. DOI: https://doi.org/10.1093/poq/nfm024.

Jones, J., J. Rushbrooke, G. Haraldsen, T. Dale, and D. Hedlin. 2005. "Conceptualising total business survey burden." In *Survey Methodology Bulletin* 55. UK Office for National Statistics, 1–10. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/smb-55/index.html (accessed March 2021)

Peytchev, A. 2011. "Breakoff and Unit Nonresponse Across Web Surveys." *Journal of Official Statistics* 27 (1): 33. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/breakoff-and-unit-nonresponse-across-web-surveys.pdf (accessed September 2021).

Presser, S., J. Blair, and T. Triplett. 1992. "Survey sponsorship, response rates, and response effects." *Social Science Quarterly* 73 (5): 699–702. Available at: http://www.jstor.org/stable/42863089.

Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack. 2013. *Designing and conducting business surveys*. New York: Wiley.

Tomaskovic-Devey, D., J. Leiter, and S. Thompson. 1994. "Organizational survey nonresponse." *Administrative Science Quarterly* 39 (3): 439–457. DOI: https://doi.org/10.2307/2393298.

Tourangeau, R., T. Yan, and H. Sun. 2019. "Who Can You Count On? Understanding The Determinants of Reliability." *Journal of Survey Statistics and Methodology* 8 (5): 903–931. DOI: https://doi.org/10.1093/jssam/smz034.

Willeboordse, A. 1997. "Minimizing response burden." In *Handbook on Design and Implementation of Business Surveys*, edited by Willeboordse. 111–118. Available at: https://ec.europa.eu/eurostat/documents/3859598/5825949/CA-09-97-818-EN.PDF/f5ee3198-6fc0-4672-96a1-8fdb4a81ca93?version = 1.0 (accessed March 2021).

Yan, T., F.G. Conrad, R. Tourangeau, and M.P. Couper. 2010a. "Should I stay or should I go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys." *International Journal of Public Opinion Research* 23 (2): 131–147. DOI: https://doi.org/10.1093/ijpor/edq046.

Yan, T., R. Curtin, and M. Jans. 2010b. "Trends in Income Nonresponse Over Two Decades." *Journal of Official Statistics* 26 (1): 145–164. Available at:

https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/trends-in-income-nonresponse-over-two-decades.pdf (accessed September 2021).

Yan, T., S. Fricker, and S. Tsai. 2014. "The impact of response burden on data quality in a longitudinal survey." In Proceedings of the International Total Survey Error Workshop, October 1–3, Washington, D.C. Available at: https://www.niss.org/events/2014-international-total-survey-error-workshop-itsew-2014 (accessed March 2021)

Yan, T., S. Fricker, and S. Tsai. 2019. "Response burden: What is it and what predicts it?" In *Advances in Questionnaire Design, Development, Evaluation and Testing*, 193–212. John Wiley & Sons. DOI: https://doi.org/10.1002/9781119263685.ch8.

Received August 2019
Revised August 2020
Accepted March 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 837–864, http://dx.doi.org/10.2478/JOS-2021-0037

Evaluating the Utility of Linked Administrative Data for Nonresponse Bias Adjustment in a Piggyback Longitudinal Survey

Tobias J.M. Büttner[1], *Joseph W. Sakshaug*[1], *and Basha Vicari*[1]

Nearly all panel surveys suffer from unit nonresponse and the risk of nonresponse bias. Just as the analytic value of panel surveys increase with their length, so does cumulative attrition, which can adversely affect the representativeness of the resulting survey estimates. Auxiliary data can be useful for monitoring and adjusting for attrition bias, but traditional auxiliary sources have known limitations. We investigate the utility of linked-administrative data to adjust for attrition bias in a standard piggyback longitudinal design, where respondents from a preceding general population cross-sectional survey, which included a data linkage request, were recruited for a subsequent longitudinal survey. Using the linked-administrative data from the preceding survey, we estimate attrition biases for the first eight study waves of the longitudinal survey and investigate whether an augmented weighting scheme that incorporates the linked-administrative data reduces attrition biases. We find that adding the administrative information to the weighting scheme generally leads to a modest reduction in attrition bias compared to a standard weighting procedure and, in some cases, reduces variation in the point estimates. We conclude with a discussion of these results and remark on the practical implications of incorporating linked-administrative data in piggyback longitudinal designs.

Key words: Attrition; auxiliary data; between-wave events; panel survey; weighting.

1. Introduction

Unit nonresponse is an important component of the Total Survey Error framework (e.g., Groves et al. 2009). If respondents are a non-random subgroup of the sample, then population estimates can be biased. For panel surveys, the risk of nonresponse bias increases with every subsequent wave as attrition occurs. To better understand mechanisms of attrition, monitor, and possibly mitigate the effects of bias, survey methodologists are reliant on auxiliary data. Yet, rich individual-level auxiliary data are rarely available for

[1] Federal Employment Agency/Institute for Employment Research, 104 Regensburger Straße, Nuremberg 90478, Germany. Emails: tobias.buettner@arbeitsagentur.de, joe.sakshaug@iab.de and basha.vicari@iab.de

Acknowledgments: This study uses data from the National Educational Panel Study (NEPS): Starting Cohort 6 – Adults, DOI:10.5157/NEPS:SC6:9.0.1. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. This study also uses the weakly anonymous ALWA survey data linked to administrative data of the IAB (ALWA-ADIAB 7509 v1). Instead of the original administrative data of ALWA-ADIAB, we have used updated data from the Integrated Employment Biographies (IEB v13.1). Data access was provided by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB). We thank FDZ and the department for Data and IT-Management (DIM) at IAB for their help.

both respondents and nonrespondents in telephone and address-based samples (Smith 2011). In panel surveys, potentially relevant information can be drawn from previous study waves to describe both groups. However, more up-to-date data covering the period in-between study waves might be more informative of the underlying response mechanisms. For example, losing one's job in-between waves might have a detrimental effect on subsequent wave response, resulting in the underestimation of labor market transitions. A promising area of research is to investigate the effects of between-wave events on attrition.

In this article, we investigate the utility of federal administrative records as an auxiliary data source for assessing and adjusting for panel nonresponse bias. As discussed later, administrative data have several advantages over other auxiliary data sources. However, rarely are such data available for nonrespondents in general population surveys. Rather, such data are typically available only for respondents who give consent and can be successfully linked to the target administrative database, as is routinely done in large-scale surveys for substantive purposes (Calderwood and Lessof 2009). To overcome this limitation, we propose a methodological framework aimed at a very specific type of longitudinal design that exploits the use of existing data linkages to study attrition bias. Specifically, we focus on so-called "piggyback" longitudinal designs, where the longitudinal sample is recruited from a stand-alone, cross-sectional survey (Cohen 2005; Edwards et al. 2011). Such designs are used in several longitudinal surveys, including the US Medical Expenditure Panel Survey-Household Component (MEPS-HC), which is subsampled from participants of the US National Health Interview Survey (Ezzati-Rice et al. 2008); the English Longitudinal Study of Ageing (ELSA), which is sampled from the Health Surveys for England (Taylor et al. 2007); and the GESIS Panel, which recruits refreshment samples from the German General Social Survey (Schaurer and Weyandt 2018).

Next to cost saving opportunities, the piggyback design can be useful for informing nonresponse adjustments in subsequent longitudinal surveys, as any information collected from the piggyback survey – including any performed data linkages – can describe the sample more comprehensively than many variable-poor sampling frames (Cohen 2005). Of course, this is predicated on the representativeness of the piggyback survey, which is typically assumed by applying nonresponse adjustment weights, which are carried over and further adjusted in the subsequent longitudinal survey. We examine whether administrative data linked to respondents in a piggyback survey are useful for assessing and adjusting for nonresponse bias over eight waves of a subsequent longitudinal survey of the general population. By doing so, we contribute to the relatively sparse literature on administrative auxiliary data and current- and between-wave events affecting nonresponse in longitudinal surveys.

2. Background

2.1. Attrition in Longitudinal Surveys

Experience shows that many longitudinal surveys suffer from a significant drop in participation rates over time, particularly in the initial waves (Watson and Wooden 2009; Sakshaug and Huber 2016). For example, the German Socio-Economic Panel (GSOEP) and the British Household Panel Survey (BHPS) experienced a loss of about one third of

their initial samples after eight annual waves (Spieß and Kroh 2004; Taylor et al. 2010). Such loss threatens the precision and accuracy of survey estimates, especially if the attrition is selective. Lepkowski and Couper (2002) suggest a theoretical framework distinguishing three conditional outcomes of obtaining a response in a longitudinal survey: location, contact, and cooperation of a sample unit. All three are assumed to be influenced by survey design features (e.g., mode, time between waves, number of contact attempts) and characteristics of the sample unit (e.g., propensity to move, at-home-patterns). Couper and Ofstedal (2009) focus on moving as an important determinant of location propensity, which may be influenced by societal-level factors (e.g., urbanization) and person-level factors (e.g., housing situation). Further, researchers discuss timing issues, emphasizing the relationship of a subject's *current* circumstances and their propensity of response. The likelihoods of moving, being located, and cooperating with the survey request tend to correlate with current job status, educational paths, and family or health circumstances, among others (Couper and Ofstedal 2009; Lemay 2009). Lemay (2009) distinguishes two ways in which life events might influence attrition: a *sociodemographic* explanation, which suggests that individual characteristics are inherently associated with attrition, and a *psychosocial* explanation that focuses on the "shock" caused by a disruptive event (e.g., moving after having lost a job, a change in household size) that affects later participation. To investigate these relationships and adjust for selective attrition, it is necessary to have auxiliary data that go beyond survey data.

2.2. *Auxiliary Data Sources for Nonresponse and Attrition Adjustment*

Auxiliary data are most effective for nonresponse bias adjustment when they are associated with both the propensity to respond and the substantive survey variables. Little and Vartivarian (2005) suggest that the association with the survey variables is more important, as it can also reduce the variances of (weighted) survey estimates. One often-used auxiliary data source is paradata – data about the survey process (Kreuter 2013), such as call record data or interviewer observations of the household/neighborhood. While these data are moderately associated with response propensities, their associations with survey variables are rather weak (Lin and Schaeffer 1995; Kreuter et al. 2010b; Kreuter and Kohler 2009; Sakshaug and Antoni 2019; West et al. 2014). Another possibility is to link commercial data to the sampling frame. For example, West et al. (2015) evaluate the utility of two commercial databases for nonresponse adjustment in the US National Survey of Family Growth and report only minor improvements compared to a paradata-only adjustment scheme (see also, Sinibaldi et al. 2014). Smith and Kim (2013) and Smith (2011) note several limitations of commercial databases, including outdated or inconsistent information and scarce documentation about the curation and quality of the data.

In the context of panel attrition, previous-wave survey data, in addition to paradata, are commonly used for adjustment (Kroh 2010; Taylor et al. 2010), as they are generally available for both respondents and nonrespondents at the current wave. A major advantage of these data is their strong correlation with the survey variables in the attrition-affected wave. However, a limitation is that they do not measure between-wave events or individuals' current circumstances that affect response propensity nor are they expected to

be strongly correlated with variables measuring change. Attrition models underlying these weighting schemes then assume that nonresponse is independent of these (unobserved) events and circumstances (Hoonhout and Ridder 2019).

2.3. *Linked Administrative Data as an Auxiliary Data Source*

Individual-level administrative data exhibit some promising features that might overcome some of the aforementioned limitations of other auxiliary sources. Although administrative data are not designed for research purposes, researchers use them extensively because they often contain detailed substantive information (e.g., welfare receipt, employment status, program participation, healthcare utilization). Typically, administrative data are generated longitudinally, which makes them a viable source for studying biographical changes, program evaluation, or simply as a complement to surveys to lower the burden of data collection (Olson 1999; Scholz et al. 2006; Antoni and Bethmann 2019). Besides substantive research, linked-administrative data are also used for methodological purposes, such as the assessment of nonresponse and measurement errors (Kreuter et al. 2010a; Meyer and Mittag 2019) and for improving survey data that are affected by these errors (Davern et al. 2019).

For administrative data to be useful, however, researchers must establish a case-by-case link with the survey sample. This step is technically straightforward when a unique identifier (e.g., Social Security number) is available, such as in countries that use population registers as sampling frames (UNECE 2007). If no unique identifier is available, then indirect linkage techniques (e.g., probabilistic linkage) are an alternative (Christen 2012; Sakshaug et al. 2017). The utility of the linked-data additionally depends on at least two further criteria. First, legal and ethical regulations may require that respondents provide informed consent to link their responses to the administrative data (Calderwood and Lessof 2009). However, without a 100% consent rate linked-data estimates are at risk of bias. Consent biases have been identified in several studies (e.g., Young et al. 2001; Knies et al. 2012); though, Sakshaug and Kreuter (2012) show that non-consent biases are generally small compared to nonresponse and measurement biases.

The second criterion is that all sample units are present in the target administrative database. If this is not the case, then the linked sample cases are only representative of the subpopulation that overlaps with the administrative population. For example, Meyer and Mittag (2019) report an imperfect overlap in a linkage of the New York State sample of the 2008–2013 Current Population Survey Annual Social and Economic Supplement (CPS-ASEC) to administrative records from the Office of Temporary and Disability Assistance and the Department of Housing and Urban Development. The authors correct for this shortcoming using an inverse probability weighting procedure. Sakshaug et al. (2017) investigate various linkage procedures between the German "Labor Market and Social Security" (PASS) survey and an administrative employment database of the German Federal Employment Agency (BA). They report varying linkage rates depending on self-reported employment status and age or sex, which is plausible since some subgroups (e.g., self-employed, civil servants, retired persons) are beyond the responsibility of the BA. In short, both non-consent and non-overlapping populations can lead to linkage bias if differences in linkage rates correlate with the linked-outcomes of interest.

2.4. Linked Administrative Data for Attrition Adjustment in a Longitudinal Setting

Given the increasing trend of linking surveys with administrative data (Couper 2017), there is a potential to leverage these data for nonresponse adjustment. Exploration of this topic, however, is mainly limited to cross-sectional studies. Sakshaug and Antoni (2019) found promising correlation patterns between linked administrative variables and the response and survey variables, but adding these auxiliary variables into the nonresponse weighting scheme had only minor impact on survey estimates. Bee et al. (2015) report mixed evidence for nonresponse bias adjustment depending on the variable of interest using CPS data enriched by tax records. The present study builds on these previous studies by evaluating the utility of linked-administrative data to assess and correct for nonresponse bias in a longitudinal setting. As previously discussed, it is plausible that events occurring between study waves or situational factors at the time of the current wave relate to attrition. This information is generally unavailable from standard paradata or prior-wave survey data, but utilizing linked administrative data that follow the life course of all (or most) sample members, including attritors, might be a viable source of this information.

Only few studies have considered auxiliary administrative data to evaluate the effects of between-wave events and current-wave status on attrition (Neukirch 2002; Trappmann et al. 2015). For instance, Trappmann et al. (2015) study the effect of changes in employment status, basic income support, moving, and household composition on dropping out within the first three waves of the PASS survey. They report positive effects of between-wave moving and ending benefit receipt on attrition propensities. Further, they showed that the existing weighting scheme eliminated biases in estimates of change in the linked variables and reduced attrition biases reasonably well.

2.5. Research Questions

Previous studies using linked-administrative data to investigate attrition are mainly based on register samples often representing special populations (e.g., benefit recipients). In contrast, non-register samples of the general population are rarely linked to administrative data for the purpose of studying nonresponse bias in surveys (for exceptions, see Sakshaug et al. 2017; Bee et al. 2015) as they usually lack unique identifiers to facilitate direct linkage. Nevertheless, many general population surveys, including cross-sectional surveys used for piggybacking longitudinal surveys, attempt to (indirectly) link administrative data to respondents for substantive research purposes (Antoni and Seth 2011; Freedman et al. 2014; Knies and Burton 2014; Korbmacher and Czaplicki 2013). We propose and evaluate the potential for exploiting these existing linkages specifically for the purpose of monitoring and adjusting for attrition in piggyback longitudinal designs.

We illustrate this new framework using the first eight waves of the National Educational Panel Study (NEPS) – Adult Cohort, a preceding cross-sectional forerunner survey, and federal administrative data in Germany. Since this particular piggyback design involves a selection step between the preceding cross-sectional survey and the subsequent longitudinal survey in the form of panel willingness consent, we consider the magnitude of this bias source and the utility of administrative data for its adjustment. In addition, we investigate whether incorporating the administrative data into the standard NEPS

weighting scheme improves estimation and reduces attrition bias. Specifically, we address the following research questions (RQ1–4):

1. To what extent do linked administrative variables describing a sample member's current status and/or changes in one's status before the attempted interview correlate with the response outcome and substantive survey variables? How do these correlations compare to standard weighting variables used for attrition adjustment, that is, paradata or previous-wave survey information?,
2. Does the inclusion of these administrative variables improve model fit in the panel willingness model and the selection models used for attrition adjustment?,
3. Does the inclusion of these administrative variables in a standard weighting scheme reduce panel willingness and attrition biases compared to the conventional weighting variables?, and
4. To what extent, if any, do weighted survey estimates differ depending on whether administrative variables are used in the weighting scheme? Is there evidence of a reduction in attrition bias for substantive estimates?

3. Data and Methods

3.1. The National Educational Panel Study (NEPS) – Adult Cohort

The NEPS is a multicohort longitudinal study that follows six cohorts, each representing different stages of educational and professional pathways in Germany (Blossfeld et al. 2011). The NEPS collects detailed information on education, competence development and contextual factors (e.g., family and peers, educational institutes attended, the workplace). We use data from the NEPS – Adult Cohort (SC6). SC6 consists of three subsamples: a cross-sectional sample from the preceding forerunner survey "Working and Learning in a Changing World" (ALWA), an augmentation sample, and a refreshment sample (Hammon et al. 2016). We use the ALWA sample only. This sample comprises individuals living in private households in Germany born between 1956 and 1986. The ALWA study was designed as an independent cross-sectional study to be used as a piggyback for the NEPS. The ALWA shows some substantive overlap with the NEPS since the investigators anticipated that it would lead to a longitudinal survey before funding was secured. Only ALWA participants who explicitly provided "panel willingness" consent were later recruited for the NEPS (Antoni et al. 2010). For the exact wording of the consent question, see Online supplemental data. The ALWA sample was drawn in 2005 in a two-stage process, with 250 municipalities drawn proportional to their size and a systematic sampling of 152 persons drawn from selected municipality records. Telephone data collection took place between 2007 and 2008. Among 9,649 eligible ALWA respondents, 8,997 (93.2%) expressed willingness to participate in a follow-up interview. This piggybacking step forms the entire eligible sample of the ALWA-portion of the NEPS.

For our case study, the NEPS piggyback sample diminishes for two reasons. First, due to technical reasons, 187 panel-willing individuals who requested Turkish or Russian language interviews could not be linked. For this reason, all 227 (187 plus 40 panel refusers) individuals with Turkish or Russian CATI were discarded in advance resulting in

9,422 eligible ALWA respondents. Since studies have shown that proficiency in the host country's language (Burkam and Lee 1998) and being a member of an ethnic minority (Lepkowski and Couper 2002; Hammon et al. 2016) have a negative effect on the propensity to respond, this might be a shortcoming. However, we assume its potential effect on the results is small as the loss amounts to only 2% of the sample. Second, to address the research questions and fully exploit the piggyback design, only survey units that can be successfully linked to administrative records in the ALWA are used. Among the 9,422 eligible respondents, 8,635 (91.7%) consented to record linkage, which is slightly higher than other consent rates reported in Germany (Sakshaug and Kreuter 2012). Among the panel-willing eligible respondents – the subgroup that is relevant for piggybacking – the rate is a little higher (8,201 out of 8,810, or 93.1%). For 7,460 units (86.4% of all consenters) linkage was achieved using a combination of deterministic and probabilistic methods (for details, see Antoni and Seth 2011). This second drop can be partially attributed to the issue of nonoverlapping populations as described earlier. After these exclusions, 7,085 of the 8,997 panel-willing respondents remained. Figure 1 summarizes all selection steps in the order they are modeled later.

By using the ALWA for piggybacking, there are eight NEPS waves available for attrition analysis. These waves stem from annual mixed-mode (telephone and face-to-face) interviewing conducted between 2009–2017. The response rates (RR1 following AAPOR 2016) drop from 73.3% in wave 1 to 63.4% in wave 2 and a decreasing decline to 43.1% in wave 8 (see Figure 2). The larger dropout in the initial waves is consistent with other longitudinal surveys, which reflects a "pruning out of the uncooperative" cases (Olsen 2018, 513). In a longitudinal piggyback design, this pruning might however take place in earlier steps such as initial nonresponse and screening for panel willingness in the preceding survey. Rather high response rates in other piggyback studies (e.g., Cheshire et al. 2012) seem to support this notion.

To evaluate the utility of auxiliary administrative data for attrition adjustment, we compare the NEPS weighting approach to the same approach augmented by linked

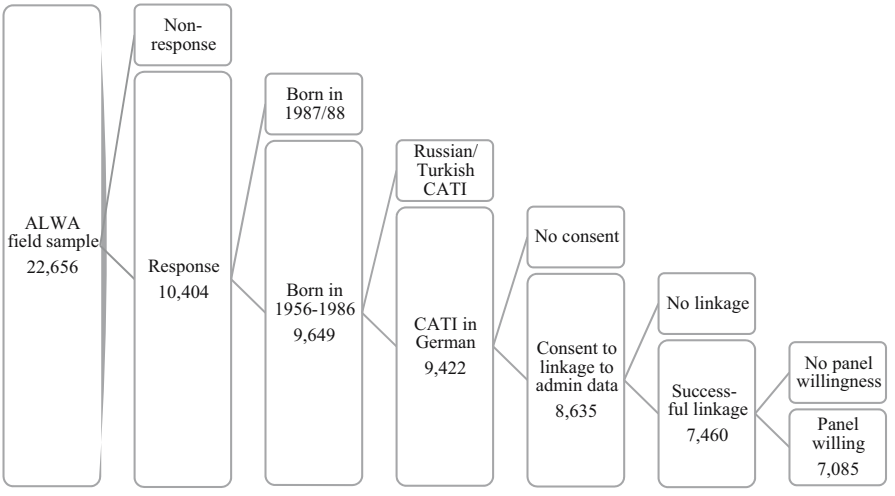

Fig. 1. Selection steps and numbers of cases from the ALWA sample to the NEPS piggyback sample.

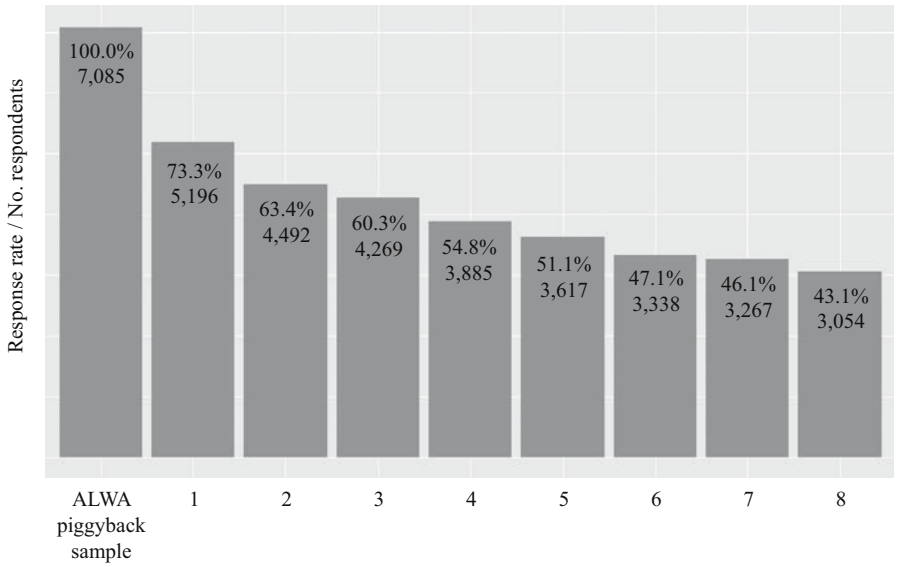

Fig. 2. Response rates by wave.

administrative variables. Therefore, the range of weighting variables are mostly predetermined. Hammon et al. (2016) and Hammon (2018) used time-constant sampling frame information of individuals' residence (federal state and BIK-10 municipality codes, which describe municipalities in Germany with regard to their size and regional structure (e.g., core, periphery, see Behrens and Wiese 2019) and survey variables (birth year, sex, marital status, mother tongue, household size, income and educational attainment). In addition, time-varying survey variables (marital status, household size and income) were used from the previous survey waves. The only paradata used for weighting is the number of contact attempts. See the Online supplemental data (Table S1) for descriptive statistics of all weighting variables. All missing values are replaced by single imputations using CART models (Burgette and Reiter 2010).

The substantive survey variables used to compare estimates under the different weighting procedures (see Subsubsection 3.3.3) are chosen based on literature research (e.g., Stöckinger et al. 2018) and discussions with data users. They include indicator variables for having children younger than six years-old in the household, for regular employment and self-employment, and indicators for working in one of three main economic sectors (WZ 2008): Manufacturing, Trade, and Health/Social Sector. Additional variables include gross monthly earned income, the number of further education courses taken in the last year, and assessment scores from the Cambridge Social Interaction and Stratification (CAMSIS) status scale. Finally, a key feature of the NEPS is the periodic collection of competence measurements (Weinert et al. 2019). We include metrics of reading, mathematics and scientific competence, vocabulary comprehension, literacy in information and communication technology (ICT), as well as reasoning skills and cognition speed (see Table S2 in the Online supplemental data for descriptive statistics of all substantive survey outcome variables).

3.2. IEB Administrative Data

The administrative data linked to the ALWA respondents comes from the Integrated Employment Biographies (IEB) of the BA. The data originate from compulsory notifications by employers to the BA and from administrative process data (Jacobebbinghaus and Seth 2007). The employer notifications are processed into individual spell data describing employment status, establishment characteristics, wages, and working time. The additional process data contain information about unemployment spells and benefit receipt, participation in active labor market programs, job search, as well as personal characteristics. As mentioned before, employment not liable to social security contributions (e.g., self-employed, civil servants) or people who never utilized one of the BA's services are not covered by the IEB. To retain all linked sample units, it is necessary to restrict the administrative variables to those generally observed for everybody, or create new variables where a missing spell in the administrative data at the time of the interview attempt clearly defines a category. For example, sample units without an employment spell in the IEB can be categorized as being "not regularly employed." Unfortunately, this rules out a few interesting variables like change of address or change in household composition, as this information is only reported to the BA for a small subgroup of current clients, mainly basic income recipients. Whenever possible, we tried to construct a proxy for these variables (e.g., started commuting to work in a different district).

Nevertheless, the IEB provides comprehensive longitudinal information on most sample units allowing to investigate their correlations with response and survey variables at, or prior to, the time of an attempted interview. For the forthcoming analyses, we use the following variables measured at the time of the current wave interview: currently employed, part-time employed, marginally employed, regular unemployment ("Unemployment Benefit I"; UBI), basic income support ("Unemployment Benefit II"; UBII), average daily wage, commuting to work in a different district, total number of employment and UBI spells in the last five years, and ever received UBI/UBII. Variables measuring changes in the individual's status between the previous and current waves include: change from not receiving UBI to receiving it (and vice versa), from non-employment to employment (and vice versa), from non-commuting to commuting to a different district, from part-time to full-time employment, and change of employer for persons who are employed in the current and previous wave. We refrained from controlling for the opposite directions of "to commuting" and "to full-time employment". The directions we control for are assumed to measure work-related changes that decrease a sample unit's availability and likelihood to respond. In this line, Trappmann et al. (2015) report that the effects of switching to full-time employment between waves on response are stronger than of switching to employment in general.

3.3. Statistical Analysis

3.3.1. Implemented Weighting Scheme

The ALWA respondents used as the sample basis for the NEPS-SC6 are those who expressed willingness to participate in a follow-up survey. In order to conserve the representativeness of the ALWA sample, the original ALWA weights are adjusted for

selection at the panel willingness step. Preceding this, adjustments for linkage consent and successful linkage to the administrative data are also performed to account for selection at these steps (the regression results modelling linkage consent and successful linkage (given consent) can be found in the Online supplemental data in Table S3). The attrition weights are constructed by replicating the weighting scheme of the NEPS, that is, using the same variables, the same approach to temporary dropout and so on (for details, see Hammon et al. 2016 and Hammon 2018). As our focus is on longitudinal nonresponse, we omit the NEPS calibration steps. We also deviate from the original NEPS weighting scheme for waves 6 to 8, where response models are conditioned on being part of the active sample (i.e., the initial sample minus units that dropped out from the sample permanently) in wave 5. With an increasing number of permanent dropouts in the later waves, this approach departs from modelling the individual decision to respond for which we investigate the explanatory potential of up-to-date administrative information. We therefore condition each of the response models for waves 6–8 on being in the active sample of the respective waves. Each wave's weights are calculated by multiplying the original ALWA weight with the inverse of the estimated propensities of the aforementioned outcomes and the previous waves' response models. All models are fitted using logistic regression with positive outcomes (response, panel willing, linkage consent, successful linkage) coded as 1 (and 0 otherwise). Finally, as in the original NEPS weighting, the weights are trimmed above the 99th-percentile for each wave with trimmed probability mass distributed evenly over the non-trimmed cases (Valliant et al. 2013). To evaluate the impact of utilizing the linked-administrative variables, all selection models, beginning with the panel willingness model, are fitted twice: once using only the original NEPS weighting variables and once using these variables plus the linked-administrative variables. This produces two sets of adjustment weights, which we compare with regard to their impact on attrition bias.

3.3.2. Attrition Bias Reduction

The utility of the linked-administrative variables for attrition bias reduction is investigated in bivariate analyses by calculating absolute Pearson correlation coefficients (RQ1) and estimating multivariate logistic regression models on the response outcomes (RQ2). RQ3 is addressed by comparing mean estimates for a selection of linked-administrative variables with and without attrition adjustment (Sakshaug and Antoni 2019). Treating the administrative variables as outcome variables enables us to compare the wave t sample mean unaffected by attrition $\left(\bar{Y}_{n,t}\right)$ (i.e., based on the respective waves' linked respondents and nonrespondents) to the following three estimates: unadjusted, original NEPS adjustment, original-administrative adjustment, which are based on the linked-respondents only $\left(\bar{Y}_{r,t}\right)$. Attrition bias is then calculated as:

$$Attrition\,Bias_t = \bar{Y}_{r,t} - \bar{Y}_{n,t} \tag{1}$$

We also report the absolute attrition bias (AAB), calculated as:

$$Absolute\,Attrition\,Bias\,(AAB_t) = \left|\bar{Y}_{r,t} - \bar{Y}_{n,t}\right| \tag{2}$$

We note that the unadjusted estimates $\bar{Y}_{r,t}$ and $\bar{Y}_{n,t}$ are unadjusted only with regard to later wave attrition. They are still weighted estimates with weights correcting for selectivity stemming from the piggyback design; that is, weights that adjust for the design

of the ALWA survey, the linkage steps, and panel willingness. In order to add to the generalizability of the results, the combined original-administrative weights for $\bar{Y}_r$ are calculated without including the target administrative outcome variable in the selection models. For example, when evaluating the bias in an estimate of the administrative variable "average daily wage", this variable is removed from the selection models used to generate the combined original-administrative weights.

3.3.3. Impact on Weighted Survey Estimates

The approach depicted in Subsubsection 3.3.2 aims to remedy the problem of $\bar{Y}_n$ typically being unobserved. However, it is clear that any weighting scheme should be evaluated with regard to estimates of the actual survey variables of interest (RQ4). Despite the lack of population benchmarks, we additionally compare unweighted and both weighted estimates of the aforementioned substantive survey variables (see Subsection 3.1). For each variable and wave, we calculate the difference between the weighted and unweighted estimates:

$$\textit{Difference between weighted and unweighted estimates}_t = \bar{Y}_{r,wtd,t} - \bar{Y}_{r,unwtd,t} \tag{3}$$

and the Absolute Difference (AD):

$$\textit{Absolute Difference}\,(AD_t) = \left|\bar{Y}_{r,wtd,t} - \bar{Y}_{r,unwtd,t}\right| \tag{4}$$

To assess the impact of the weighting schemes on the variability of the weighted estimates, we additionally report coefficients of variation (CV) for all three estimates and the difference between CVs for both weighting schemes' estimates:

$$\begin{aligned}\textit{CV change using administrative variables}_t &= CV\left(\bar{Y}_{r,orig+admin,t}\right) - CV(\bar{Y}_{r,orig,t}) \\ &= \left(\frac{SE(\bar{Y}_{r,orig+admin,t})}{|\bar{Y}_{r,orig+admin,t}|} \times 100\%\right) - \left(\frac{SE(\bar{Y}_{r,orig,t})}{|\bar{Y}_{r,orig,t}|} \times 100\%\right)\end{aligned} \tag{5}$$

A negative value indicates reduced variation in the estimate due to the combined original-administrative weights relative to the estimate based on the original weights (i.e., without using the administrative data).

All analyses are conducted using the family of "svy" commands in Stata 15.1 (StataCorp 2017). Estimates of standard errors are based on Taylor-Series linearization.

4. Results

4.1. Correlation Between Linked Administrative Variables, Response, and Survey Variables

To evaluate the utility of the administrative variables, we first examine their correlation patterns with the response outcome and the NEPS substantive survey variables (RQ1). Figure 3 depicts the absolute Pearson correlation coefficients, calculated across all eight NEPS waves (for tabular versions of all correlation coefficients, see Tables S4 and S5 in the Online supplemental data). For comparison, the same correlations are shown for the original NEPS weighting variables. Starting with the response outcome, the correlations

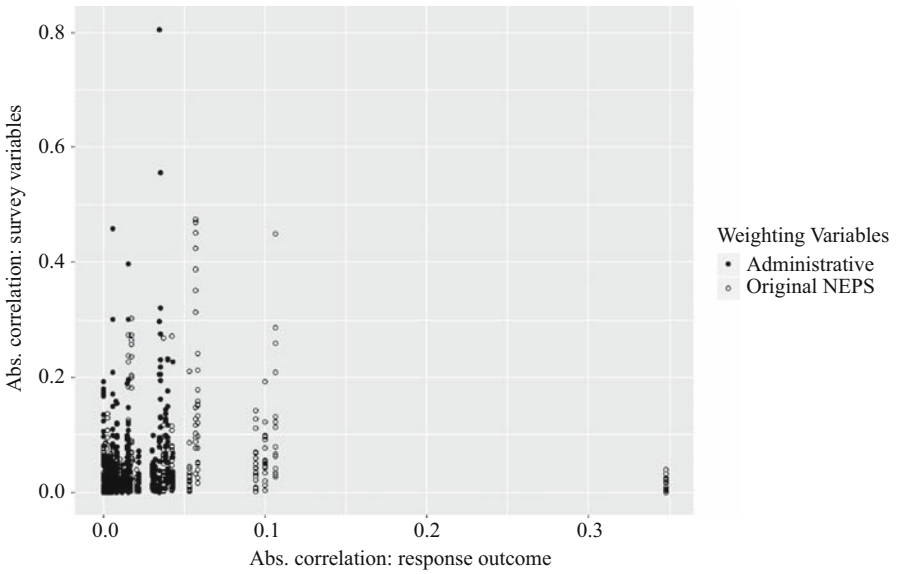

Fig. 3. Absolute Pearson correlation coefficients of the administrative and original NEPS weighting variables with the substantive survey variables and response outcome; Waves 1–8 pooled.

are generally low for both the administrative and NEPS weighting variables. The majority of correlations lie below 0.1 with all correlations involving administrative variables no larger than 0.04. The largest correlation (0.35) is observed for the paradata variable: number of contact attempts. The absolute correlations between both sets of variables and the panel willingness indicator are also very small, peaking at 0.05 for the original NEPS and 0.03 for the administrative variables (results not shown).

Around 83% (administrative variables) and 91% (original NEPS variables) of all possible correlations with the substantive survey variables are less than or equal to 0.1. Regarding the original NEPS weighting variables, the education indicators are moderately correlated (between 0.31 and 0.47) with the various competence measures and the CAMSIS score. In addition, sex, year of birth, born abroad, and mother tongue show higher correlations with the substantive survey variables. With respect to the administrative variables, there are moderate-to-high correlations between current wages (0.28) and employment ("Not employed"-category: 0.80) and their similarly-measured survey variables. The highest administrative variable correlation indicating a between-wave change is for leaving employment and the employment survey variable (0.23). The strong correlations with the substantive survey variables suggest that the administrative adjustment variables could have a decreasing effect on variances of the weighted estimates. We explore this possibility later.

4.2. Utility of Linked Administrative Variables in Panel Willingness and Response Models

Next, we compare the selection models with and without the linked-administrative variables (RQ2). The selection models are estimated separately for panel willingness and

each of the eight NEPS waves. The panel willingness model, shown in Table S6 in the Online supplemental data, shows a negative association with being born abroad and positive associations with parents' education and household income. Overall, including the administrative variables does not add much explanatory power to the already low fit of the original panel willingness model (Pseudo-R^2 of 0.03 in both models). The only statistically significant administrative variable is current UBII receipt, which is positively associated with panel willingness.

The response models for wave 1 indicate positive associations with the originally included variables age, gender (male), and higher education, whereas being born abroad, German mother tongue, and a high number of contact attempts have negative associations. These associations remain statistically significant after adding the administrative variables. Only one administrative variable, average daily wage, is statistically significant, having a positive association with response. All response models for waves 1 to 8 can be found in the Online supplemental data.

For brevity, Table 1 summarizes all significant administrative variables from the response models for waves 1–8. Some administrative variables are significant in multiple waves. Average daily wages, the number of UBI spells in the last five years, and having

Table 1. Summary of all statistically significant administrative predictors in NEPS waves 1–8 response models.

Administrative variable	NEPS-wave							
	1	2	3	4	5	6	7	8
Average daily wage (in EUR)	+				-			
Working hours								
Not employed	REF	REF	REF	REF	REF	REF	REF	REF
Part-time					+		+	
Full-time			+		+			
Receiving Unemployment Benefit II					+			
Commuting to work (five-digit municipality code)		+						
Number of employment spells last five years				+				
Number of UBI spells last five years			+	-				
Ever received UBII in lifetime					-			
Variables of change since t-1								
Became UBI recipient							+	
Became employed			-					
Left employment			-	-	-			
Stopped receiving UBII			-					
Different employer compared to last wave					+			-

Notes: +/- indicate the direction of statistically significant ($p < 0.05$) associations with survey response.

changed employers since the last interview are significant in two waves each. Their estimated coefficients, however, switch signs in their respective response models. For example, average daily wage is positively related to response in wave 1, but negatively in wave 5. Two other findings are possibly more generalizable. First, being employed (part-time or full-time) at the time of the interview attempt is positively related to response in three waves. Second, having left employment since the previous interview relates negatively to response in three waves. Note that individuals becoming UBI or UBII recipients are only a subset of all those who leave employment (about 23% of all linked sample units between waves 1 and 8). The IEB administrative data do not allow for further investigation of these changes in labor force participation.

As a sensitivity analysis, we dropped the wage variable, which is strongly correlated with the working hours indicators. This causes the indicator for working part-time in wave 5 to become insignificant, but at least one of the two working indicators is additionally significantly positive in waves 1, 3, and 6. This finding further underpins the predictive power of current employment on the response outcome. The results depicted in Subsections 4.3 and 4.4 are generally not affected from excluding the wage variable from the weighting scheme.

Table 2 summarizes two model fit statistics, Pseudo R^2 (higher is better) and Akaike's Information Criterion (AIC; lower is better), for the panel willingness model and each of the eight NEPS wave response models with and without the linked-administrative variables. Both statistics generally indicate no substantial improvements in model fit when including the linked-administrative variables.

4.3. Utility of Administrative Data for Attrition Bias Reduction

Next, we turn to the question of whether adding linked administrative variables to the original NEPS weighting procedure affects nonresponse biases over the course of the eight panel waves (RQ3). To this end, we compare estimated means and proportions of administrative variables using both sets of weights with the corresponding estimates based on the whole sample of linked sample units. The latter is therefore an estimate unaffected by initial nonresponse and attrition. Moreover, we report unweighted estimates using each waves' linked respondents and coefficients of variation of all estimates.

Before turning to attrition biases, the results of the estimated panel willingness biases and the impact of using administrative variables for their adjustment are shown. Table 3 depicts the absolute panel willingness biases for a subset of administrative and ALWA survey outcome variables (see Table S2 in the Online supplemental data for descriptive statistics for these variables). The column with the sample mean/percentage is based on all panel-willing sample units, adjusted for the design and selectivity in the consent and linkage steps. This information puts the absolute biases into perspective. They are generally very small, with the indicator variables not exceeding 0.4 percentage points. Still, both weighting schemes have a decreasing effect on the bias. Across all categorical variables, using the original NEPS weights results in an average absolute panel willingness bias of 0.2 percentage points, whereas the average bias using the original-administrative weights is slightly smaller with 0.1 percentage points. The difference between the pairs of weighted estimates is very small, which is why the original-administrative panel willingness weights are used for the remainder of the

Table 2. Model fit statistics for panel willingness models and each NEPS wave response models on original NEPS and administrative weighting variables.

	Model fit statistics	Panel willingness	NEPS-wave							
			1	2	3	4	5	6	7	8
Original NEPS	Pseudo R^2	0.03	0.07	0.14	0.12	0.12	0.11	0.14	0.12	0.10
	AIC	2974.4	7714.5	4349.9	3132.7	3282.3	3760.9	3612.3	2940.1	2541.6
Original NEPS + administrative data	Pseudo R^2	0.03	0.08	0.14	0.14	0.13	0.12	0.15	0.13	0.11
	AIC	2984.2	7718.4	4362.7	3119.3	3292.7	3756.6	3627.7	2943.4	2558.6

Table 3. Absolute panel willingness biases under both weighting schemes.

Data	Variable	Sample mean/ percentage	Absolute panel willingness bias		
			No weights	Weighted (orig)	Weighted (orig + admin)
Administrative data	Currently employed (%)	73.96	0.3	0.2	0.1
	Currently marginally employed (%)	12.24	0.1	0.1	0.1
	Receiving UB II (%)	5.75	0.0	0.1	0.0
	Ever received UB I in lifetime (%)	51.78	0.4	0.3	0.1
	Ever received UB II in lifetime (%)	9.71	0.1	0.1	0.1
	Commuting to work (%)	28.70	0.3	0.3	0.1
	Average daily wage (in EUR)	56.28	0.3	0.1	0.0
	Number of employment spells last five years	4.73	0.0	0.0	0.0
ALWA survey	Currently employed (%)	67.73	0.4	0.3	0.2
	Currently self-employed (%)	12.45	0.4	0.4	0.4
	Children younger than six years in household (%)	16.63	0.0	0.1	0.1
	Net monthly income (in EUR)	1672.26	6.2	0.6	1.2

Notes: Changes in the coefficient of variation after including administrative variables in the weighting procedure are all close to zero and not shown; biases for indicator variables in percentage points.

attrition analysis – a larger difference already at this point would otherwise question whether differences in later waves' estimates are driven by attrition bias.

Table 4 summarizes the effects of using no weights, the original and the original-administrative weights on attrition bias and variance across all waves for selected variables (see Tables S15 to S26 in the Online supplemental data for results in greater detail). We skipped the analysis for the variable "Number of UBI Spells last five Years", since already one other variable relates to the receipt of UBI in the past. Moreover, we dropped all variables with estimated proportions lower than 0.05 in the fully linked sample. The absolute biases for these were quite small and differences between both weighting schemes negligible. Averaged over all waves, estimates for most (7 out of 11) variables are less biased using the combined original-administrative adjustment. In two cases, the absolute bias increases. Estimated mean variances slightly decrease in all but one case.

Figure 4 depicts the development of absolute biases from wave 1 to wave 8 with the unweighted scenario giving an impression of attrition bias present in the panel. It highlights some considerable biases from attrition, often in the later waves. For example, estimates of average daily wage in wave 8 based on the full or the realized sample differ in absolute terms by almost EUR 5 (with a full sample mean of EUR 78.30. Other variables with larger biases are current employment and having ever received UBI or UBII. Using the additional administrative data clearly lowers AABs for the two variables employment

Table 4. Change in absolute attrition biases (AAB) and coefficients of variation (CV) under both weighting schemes; averages over eight waves.

Variable	Sample mean/ percentage All wave average	Absolute attrition bias (AAB) All wave average			CV change using administrative variables All wave average
		No weights	Weighted (orig)	Weighted (orig + admin)	
Currently employed (%)	76.38	1.5	1.4	0.8	-0.3
Currently marginally employed (%)	11.28	0.3	0.6	0.5	-0.1
Receiving UB II (%)	4.51	0.2	0.5	0.9	+0.0
Ever received UB I in lifetime (%)	56.61	1.3	1.2	0.9	-0.1
Ever received UB II in lifetime (%)	14.08	1.7	0.7	0.7	-0.1
Commuting to work (%)	30.97	0.7	0.9	0.6	-0.1
Average daily wage (in EUR)	69.88	3.2	3.5	1.9	-0.0
Number of employment spells last five years	4.75	0.1	0.1	0.1	-0.0
Became employed (%)	4.87	0.3	0.4	0.4	-0.0
Left employment (%)	4.58	0.5	0.6	0.5	-0.0
Started commuting to workplace (%)	3.93	0.3	0.5	0.4	-0.0
Different employer compared to last wave (%)	8.09	0.6	0.7	0.8	-0.1

Notes: CV change in percentage points, negative value means CV is reduced under the original-administrative weighting scheme; AAB for indicator variables in percentage points.

and daily wages, but for the remaining variables there is little difference between both weighting schemes.

Table 5 summarizes the precision effects of introducing administrative variables differentiating between waves. The results suggest, on average, small decreases in the estimates' sampling variance in later waves compared to the original weighting scheme.

4.4. Impact of Administrative Data on Weighted Survey Estimates

Finally, we assess the impact of the original-administrative weighting scheme on estimates of actual survey variables (RQ4). Altogether, we estimate proportions and means of 16 survey variables, nine of them over all eight waves. The other seven variables stem from competence measurements that were carried out in selected waves only. As an example, Table 6 shows the estimated proportion of regularly employed persons in the NEPS population. As a first result, each wave's estimated proportions based on no attrition adjustment, the original NEPS weighting scheme, and the combined original-administrative weights are rather similar. In every wave, all three 95% confidence

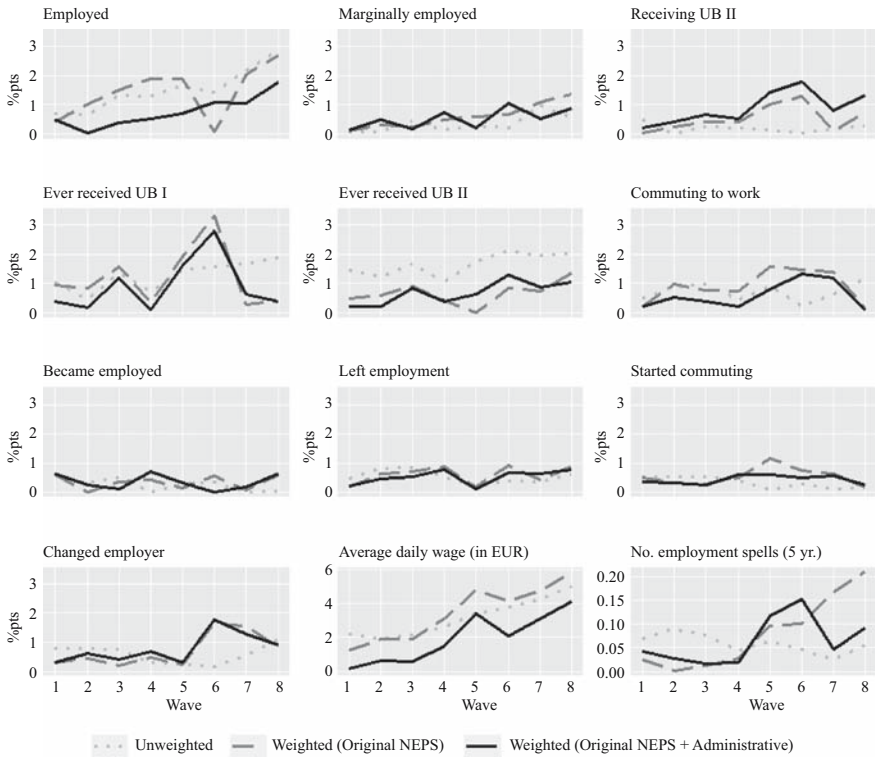

Fig. 4. Absolute attrition biases (AAB) by wave and weighting scheme.

Table 5. Average changes in coefficients of variation (CV) using the original-administrative weighting adjustment compared to the original NEPS adjustment by wave; averages calculated over 11 variables.

NEPS-Wave	1	2	3	4	5	6	7	8
CV change using administrative variables	+0.0	+0.0	+0.0	-0.0	-0.1	-0.1	-0.1	-0.2

intervals of the estimates overlap. This overlap can also be found for the remaining 15 variables (see Tables S27 to S35 in the Online supplemental data). From this we conclude, that including administrative variables in the weighting scheme does not seem to produce substantially different results.

Table 6 also shows the difference between the unweighted estimates and the original- and original-administrative-weighted estimates. For the regular employment variable, using the combined original-administrative adjustment results in slightly larger deviations from the unweighted estimate. As already mentioned before, we generally do not have benchmark estimates unaffected by attrition. However, for this variable correlation patterns shown in Subsection 4.1 imply similarity in measurement to the employment variable in the administrative data. In both cases, using the original-administrative weight shifts the estimated employment rate downwards (see Table S15 in the Online supplemental data). The smaller bias in the administrative counterpart lends support to the assumption that point estimates of the survey estimates are less biased when using the combined weighting scheme. Still, this reasoning is restricted to the employment variable, since we do not have

Table 6. Differences in estimated proportions of regular employment and coefficients of variation (CV) between both weighting schemes.

NEPS-Wave	Linked respondents	Linked respondents			Difference between weighted and unweighted estimates		% CV			CV change using admin
		No weights	Weighted (orig)	Weighted (orig + admin)	Weighted (orig)	Weighted (orig + admin)	No weights	Weighted (orig)	Weighted (orig + admin)	
1	5196	0.76 (0.75, 0.77)	0.76 (0.74, 0.77)	0.75 (0.73, 0.76)	0.00	-0.01	2.9	3.0	3.0	-0.0
2	4492	0.77 (0.75, 0.78)	0.77 (0.76, 0.79)	0.76 (0.75, 0.78)	0.00	-0.01	3.2	3.5	3.4	-0.0
3	4269	0.79 (0.78, 0.80)	0.80 (0.78, 0.81)	0.78 (0.77, 0.80)	0.00	-0.01	3.5	3.9	3.9	-0.0
4	3882	0.79 (0.77, 0.80)	0.79 (0.77, 0.80)	0.77 (0.76, 0.79)	0.00	-0.01	3.6	4.2	4.1	-0.1
5	3617	0.80 (0.78, 0.81)	0.80 (0.77, 0.82)	0.79 (0.76, 0.81)	0.00	-0.01	3.8	6.4	5.8	-0.5
6	3338	0.79 (0.78, 0.81)	0.78 (0.75, 0.81)	0.77 (0.74, 0.80)	-0.01	-0.02	4.0	6.7	6.3	-0.4
7	3267	0.80 (0.78, 0.81)	0.80 (0.77, 0.83)	0.79 (0.76, 0.81)	0.01	-0.01	4.1	7.4	6.9	-0.5
8	3054	0.81 (0.79, 0.82)	0.79 (0.76, 0.82)	0.78 (0.75, 0.81)	-0.01	-0.03	4.3	8.2	7.0	-1.2

Notes: 95% confidence intervals of the estimates in parentheses.

further similarly-measured administrative counterparts for the remaining survey variables. If the differences in the estimates between both weighting schemes were not generally small (see Table 7 for a summary of all 16 survey variables), then subject matter expertise could be a means to conclude whether the estimates might be affected by attrition bias and which weighting scheme is more effective at reducing the bias.

Finally, the results in Subsection 4.3 suggested small decreases in the estimates' CVs under the original-administrative weighting scheme. Table 8 summarizes the changes in coefficients of variation over the nine survey variables for which we have data from each wave. Across all waves, the largest mean change in percentage points is -0.3, which occurs in NEPS-wave 8. This suggests that time-varying and up-to-date variables derived from the administrative data might be more correlated with the later waves' survey variables than are the mostly time-constant original weighting variables. Looking at wave-specific correlation patterns as in Subsection 4.1 does not however yield clear-cut evidence for this. The larger change in average CV in wave 8 seems to be driven, for the most part, by the regular employment variable (see Table 6). In general, effects on variances are small.

Table 7. Means and percentages of survey outcomes, and absolute differences to unweighted estimates (AD) under both weighting schemes; all available wave averages.

Variable	Unweighted mean/percentage All available wave average	Absolute difference (AD) All available wave average	
		Weighted (orig)	Weighted (orig + admin)
Regular employment (%)	78.75	0.55	1.28
Self-employed (%)	14.09	1.39	0.80
Economic sector: manufacturing (%)	21.36	0.38	0.52
Economic sector: trade (%)	7.94	0.32	0.45
Economic sector: health/social (%)	11.75	0.64	0.66
Having children younger than six years (%)	13.68	1.20	1.24
Gross monthly earned income (in EUR)	3054.32	60.15	84.00
Number of further education courses last year	0.72	0.04	0.04
CAMSIS status score	52.57	0.50	0.35
Math score	-0.11	0.05	0.04
Reading score	-0.03	0.05	0.04
ICT literacy score	0.08	0.00	0.01
Scientific literacy score	0.02	0.03	0.04
Vocabulary comprehension	73.42	1.16	0.97
Reasoning	8.75	0.04	0.01
Cognition speed	34.70	0.24	0.35

Notes: AD for indicator variables in percentage points.

Table 8. Average changes in coefficients of variation (CV) using the original-administrative weighting adjustment compared to the original NEPS adjustment by wave; Averages calculated over nine survey variables.

NEPS-Wave	1	2	3	4	5	6	7	8
CV change using administrative variables	+0.0	-0.0	+0.0	+0.0	-0.1	+0.0	-0.1	-0.3

5. Discussion

The aim of this study was to evaluate whether a specific type of longitudinal survey design (a so-called piggyback design) has, besides various cost saving opportunities, potential for improving nonresponse and attrition adjustments through the use of existing linked-administrative data. We investigated four research questions using a combination of a general population panel survey (the NEPS) and federal administrative data. These administrative data include detailed longitudinal labor market and social security related biographies. If not already part of the sampling frame, such auxiliary administrative data are typically unavailable for initial nonrespondents and attritors of a longitudinal survey. Here, however, the sample was recruited from a preceding cross-sectional survey (the ALWA), where consenting respondents' interview data had already been linked to the administrative data. We deem this setting of using independent cross-sectional surveys as a basis to recruit subsequent longitudinal study participants, as well as the general linkage of survey and administrative data, to be relatively common practices, which are likely to become more frequent as survey budgets become more constrained. Combining both practices in order to improve the quality of longitudinal survey data is a new concept, which we evaluated in the present study.

Within our first research question (RQ1), we compared correlation patterns of the survey's original nonresponse weighting variables and a set of linked-administrative variables each with the response indicator and a selection of substantive survey outcome variables across all survey waves. With one exception, correlations of both variable sets with response were rather low. Regarding the substantive survey variables, both sets showed low-to-moderate correlations. Some of the administrative variables showed similar and even higher correlations with (construct-similar) survey variables.

RQ2 checked relationships of both variable sets with panel willingness and response in the wave-specific selection models. Adding the 16 administrative variables to each of the models showed only small gains in model fit. Although the panel willingness model yielded only few significant effects, several of the employment biography-related variables were significant in different response models across the eight study waves. Two associations proved to be relatively stable over the investigated timespan: Current employment and having left employment since the previous interview showed significant associations (positive and negative, respectively) with response in three out of eight waves. Proxy-variables for moving, such as change of employer or starting to commute to a different district, were not consistent significant predictors of response, in contrast to findings reported in related studies (e.g., Trappmann et al. 2015; Kroh 2010; Short and McArthur 1986; Watson and Wooden 2009).

Using the administrative data as outcome variables to assess attrition bias, we compared adjustment effects of alternative weighting schemes – with and without adding administrative variables into the weighting procedure (RQ3). Although attrition biases were rather low for many of the investigated variables, we observed a clear reduction of bias for the variables most affected by attrition: either the combined original-administrative weights outperformed the original NEPS weighting procedure (for employment and wage variables) or both weighting schemes performed similarly well in reducing the bias (for Unemployment Benefit I/II receipt variables). The same applied to

the adjustment of panel willingness bias. An important result for the application of the piggyback design is that actual panel willingness biases are generally small, and both weighting schemes successfully reduce it even further. Concerning RQ4, we found that adding administrative variables into the weighting procedure had only minor effects on point estimates of substantive survey variables; again, this could be because there was little attrition bias in the survey to begin with. However, there was some evidence of a bias reduction for the survey-measured current employment status variable, which had one of the largest biases in its administrative variable counterpart. In terms of variance, we observed a tendency towards slightly smaller coefficients of variation in later waves when the administrative variables were included in the weighting scheme.

We note that the dropout in the NEPS sample was quite substantial, after eight waves only about 43% of the wave 1 sample continued to respond. With regard to the linked-administrative variables that were available for both respondents and nonrespondents, the evidence suggested substantial attrition bias for some variables. For some policy-relevant indicators (employment, marginal employment, wages), we saw an increasing trend in bias over time. For instance, the average daily wage in wave 8 was (unadjusted) estimated at EUR 83.2, which is roughly 6% more than the EUR 78.2 estimated from the full sample unaffected by attrition. As mentioned before, here the combined original-administrative weights reduce some of the bias (producing an estimate of EUR 82.3) compared to the original-weighted estimate (EUR 84.1). The remaining bias suggests, however, that the assumption of respondents and nonrespondents being equal given the auxiliary variables does not entirely hold, and that the procedure does not eliminate all bias.

With modest benefits, the utility of linked-administrative records for attrition adjustment is even more sensitive to its costs. This study simulated the application of a piggyback design, also with reference to its cost-saving potentials. In a piggyback design, with administrative data already linked to the preceding survey, the threshold to implement a weighting adjustment using the administrative data is possibly lower than if the data are not already linked. However, one must consider the potential loss of sample units due to unsuccessful linkage and panel non-willingness (e.g., Sin 2006), which could offset the benefits of augmented weighting if not properly accounted for. Moreover, it should be mentioned that administrative data can contain specific measurement error (e.g., Pavlopoulos and Vermunt 2015; Pankowska et al. 2018) which might attenuate correlations that are crucial for the proposed procedures.

Assuming that these issues can be addressed, this approach is worth considering when setting up longitudinal surveys using a piggyback design that includes linked-administrative data from the preceding survey. Depending on the scope of the study, this approach seems likely to yield advantages from having rich information on initial sample units that can be used to compensate for their decreasing participation over time. As shown, augmented nonresponse weights can be useful for bias adjustment without increasing sampling variance. Moreover, we envision the suggested strategy to help with monitoring nonresponse (with regard to certain subgroups) over the course of the panel, even if the administrative variables are not included in any adjustment procedure. This could help to assess the sample's longitudinal representativeness and potentially substantiate design decisions regarding future waves or refreshment samples.

6. References

American Association for Public Opinion Research. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, (9th edition). Lanexa: AAPOR. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed October 2021).

Antoni, M., and A. Bethmann. 2019. "PASS-ADIAB – Linked Survey and Administrative Data for Research on Unemployment and Poverty." *Journal of Economics and Statistics* 239(4): 747–756. DOI: https://doi.org/10.1515/jbnst-2018-0002.

Antoni, M., K. Drasch, C. Kleinert, B. Matthes, M. Ruland, and A. Trahms. 2010. *Working and Learning in a Changing World, part I: Overview of the study, FDZ Methodenreport No. 5/2010 (en)*. Nürnberg: Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (IAB). Available at: http://doku.iab.de/fdz/reporte/2010/MR_05-10_EN.pdf (accessed October 2021).

Antoni, M., and S. Seth. 2011. *ALWA-ADIAB – Linked individual Survey and Administrative Data for Substantive and Methodological Research, FDZ Methodenreport No. 12/2011 (en)*. Nürnberg: Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (IAB). Available at: http://doku.iab.de/fdz/reporte/2011/MR_12-11_EN.pdf (accessed October 2021).

Bee, C.A., G.M.R. Gathright, and B.D. Meyer. 2015. "Bias from Unit Non-Response in the Measurement of Income in Household Surveys." Paper presented at the Joint Statistical Meetings of the American Statistical Association, August 9, 2015, Seattle, USA. Available at: http://www.solejole.org/16068.pdf (accessed February 2020).

Behrens, K., and K. Wiese. 2019. "BIK-Regionen." In *Regionale Standards: Ausgabe 2019*, edited by K. Beckmann, K. Behrens, H. Hoffmann, M. Pfister, K. Wiese, J.H.P. Hoffmeyer-Zlotnik, G. Rösch, P. Siegers, W. Sodeur, U. Hanefeld, R. Herter-Eschweiler, and E. Krack-Roberg: 114–126. Cologne, Germany: GESIS – Leibniz-Institut für Sozialwissenschaften.

Blossfeld, H.P., H.G. Roßbach, and J. von Maurice (Eds.). 2011. *Education as a Lifelong Process – The German National Educational Panel Study (NEPS)*, Zeitschrift für Erziehungswissenschaft: Sonderheft 14. Wiesbaden: Springer VS.

Burgette, L.F., and J.P. Reiter. 2010. "Multiple Imputation for Missing Data via Sequential Regression Trees." *American Journal of Epidemiology* 172(9): 1070–1076. DOI: https://doi.org/10.1093/aje/kwq260.

Burkam, D.T., and V.E. Lee. 1998. "Effects of Monotone and Nonmonotone Attrition on Parameter Estimates in Regression Models with Educational Data." *The Journal of Human Resources* 33(2): 555–574. DOI: https://doi.org/10.2307/146441.

Calderwood, L., and C. Lessof. 2009. "Enhancing Longitudinal Surveys by Linking to Administrative Data." In *Methodology of Longitudinal Surveys*, edited by P. Lynn: 55–72. West Sussex, England: John Wiley & Sons.

Cheshire, H., D. Hussey, J. Medina, K. Pickering, N. Wood, K. Ward, K. Taylor, and C. Lessof. 2012. *Financial circumstances, health and well-being of the older population in England: The 2008 English Longitudinal Study of Ageing: Wave 4 Technical Report*. London: National Centre for Social Research. Available at: https://ifs.org.uk/elsa/report10/w4_tech.pdf (accessed October 2021).

Christen. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Germany: Springer. DOI: https://doi.org/10.1007/978-3-642-31164-2_2.

Cohen, S.B. 2005. "Integrated survey designs: A framework for nonresponse bias reduction." *Journal of Economic and Social Measurement* 30(2–3): 101–114. DOI: https://doi.org/10.3233/JEM-2005-0244.

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. West Sussex, England: John Wiley & Sons.

Couper, M.P. 2017. "New Developments in Survey Data Collection." *Annual Review of Sociology* 43(1): 121–145. DOI: https://doi.org/10.1146/annurev-soc-060116-053613.

Davern, M.E., B.D. Meyer, and N.K. Mittag. 2019. "Creating Improved Survey Data Products Using Linked Administrative-Survey Data." *Journal of Survey Statistics and Methodology* 7(3): 440–463. DOI: https://doi.org/10.1093/jssam/smy017.

Edwards, B., L. Branden, and M. Stange. 2011. "Piggyback Survey Respondents and Mode: Lessons Learned from Design and Operations." In *JSM Proceedings*, Survey Research Methods Section, August 1, 2011: 5009–5021. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2011/Files/302992_69483.pdf (accessed October 2021).

Ezzati-Rice, T.M., F. Rohde, and J. Greenblatt. 2008. *Sample Design of the Medical Expenditure Panel Survey Household Component, 1998–2007. Methodology Report No. 22. March 2008*. Rockville, MD, USA: Agency for Healthcare Research and Quality. Available at: http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.shtml (accessed February 2020).

Freedman, V.A., K. McGonagle, and P. Andreski. 2014. *The Panel Study of Income Dynamics: Linked Medicare Claims Data.* PSID Technical Series Paper No. 14-01. University of Michigan. Available at: https://psidonline.isr.umich.edu/publications/Papers/tsp/2014-01_PSIDMedicare.pdf (accessed October 2021).

Groves, R., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*, (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Hammon, A., S. Zinn, C. Aßmann, and A. Würbach. 2016. *Samples, Weights, and Nonresponse: The Adult cohort of the National Educational Panel Study (Wave 2 to 6)*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. NEPS Survey Paper No. 7. DOI: https://doi.org/10.5157/NEPS:SP07:1.0.

Hammon, A. 2018. *Samples, Weights, and Nonresponse: The Adult Cohort of the National Educational Panel Study (Wave 7 to 9)*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study. Available at: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/9-0-0/SC6_9-0-0_W.pdf (accessed October 2021).

Hoonhout, P., and G. Ridder. 2019. "Nonignorable Attrition in Multi-Period Panels With Refreshment Samples." *Journal of Business & Economic Statistics* 37(3): 377–390. DOI: https://doi.org/10.1080/07350015.2017.1345744.

Jacobebbinghaus, P., and S. Seth. 2007. "The German Integrated Employment Biographies Sample IEBS." *Schmollers Jahrbuch: Journal of Applied Social Science*

Studies 127: 335–342. Available at: https://www.ratswd.de/download/schmollers/2007_127/Schmollers_2007_2_S335.pdf (accessed October 2021).

Korbmacher, J., and C. Czaplicki. 2013. "Linking SHARE Survey Data with Administrative Records: First Experiences from SHARE-Germany." In *SHARE Wave 4: Innovations & Methodology*, edited by F. Malter and A. Börsch-Supan: 47–53. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Knies, G., J. Burton, and E. Sala. 2012. "Consenting to Health Record Linkage: Evidence from a Multipurpose Longitudinal Survey of a General Population." *BMC Health Services Research* 12(52) . DOI: https://doi.org/10.1186/1472-6963-12-52.

Knies, G., and J. Burton. 2014. "Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys." *BMC Medical Research Methodology* 14(125). DOI: https://doi.org/10.1186/1471-2288-14-125.

Kreuter, F., and U. Kohler. 2009. "Analyzing Contact Sequences in Call Record Data: Potential and Limitations of Sequence Indicators for Nonresponse Adjustments in the European Social Survey." *Journal of Official Statistics* 25(2): 203–226.

Kreuter, F., G. Müller, and M. Trappmann. 2010a. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74: 880–906. DOI: https://doi.org/10.1093/poq/nfq060.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T.M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R.M. Groves, and T.E. Raghunathan. 2010b. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173: 389–407. DOI: https://doi.org/10.1111/j.1467-985X.2009.00621.x.

Kreuter, F. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: Wiley.

Kroh, M. 2010. *DIW Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) (1984 until 2009). DIW Data Documentation 50*. Berlin: German Institute for Economic Research. Available at: https://www.diw.de/documents/publikationen/73/diw_01.c.359697.de/diw_datadoc_2010-050.pdf (accessed October 2021).

Lemay, M. 2009. *Understanding the Mechanisms of Panel Attrition.* PhD Dissertation. University of Maryland, College Park. Available at: http://hdl.handle.net/1903/9631 (accessed October 2021).

Lepkowski, J.M., and M.C. 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.A. Little: 259–272. New York, NY: John Wiley & Sons.

Lin, I., and N.C. Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59: 236–258. DOI: https://doi.org/10.1086/269471.

Little, R.J.A., 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/pub/12-001-x/2005002/article/9046-eng.pdf?st=3epbdduy (accessed October 2021).

Meyer, B.D., and N. Mittag. 2019. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net." *American Economic Journal: Applied Economics* 11(2): 176–204. DOI: https://doi.org/10.1257/app.20170478.

Neukirch, T. 2002. "Nonignorable Attrition and Selectivity Biases in the Finnish Subsample of the ECHP: an Empirical Study Using Additional Register Information." *Chintex Working Paper 5*. Frankfurt a. M.: J.W. Goethe-Universität. Available at: http://hdl.handle.net/10068/92455 (accessed October 2021).

Olsen, R. 2018. "Panel Attrition." In *The Palgrave Handbook of Survey Research*, edited by D. Vannette and J. Krosnick: 509–517. DOI: https://doi.org/10.1007/978-3-319-54395-6_59.

Olson, J.A. 1999. "Linkages with Data from Social Security Administrative Records in the Health and Retirement Study." *Social Security Bulletin* 62(2): 73–85. Available at: https://www.ssa.gov/policy/docs/ssb/v62n2/v62n2p73.pdf (accessed October 2021).

Pankowska, P., B. Bakker, D.L. Oberski, and D. Pavlopoulos. 2018. "Reconciliation of two data sources by correction for measurement error: the feasibility of parameter re-use." *Statistical Journal of the IAOS* 34(3): 317–329. DOI: https://doi.org/10.3233/SJI-170368.

Pavlopoulos, D., and J.K. Vermunt. 2015. "Measuring temporary employment. Do survey or register data tell the truth?" *Survey Methodology* 41(1): 197–214. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015001/article/14151-eng.pdf?st=tAsC-b20 (accessed October 2021).

Sakshaug, J.W., M. Antoni, and R. Sauckel. 2017. "The Quality and Selectivity of Linking Federal Administrative Records to Respondents and Nonrespondents in a General Population Survey in Germany." *Survey Research Methods* 1: 63–80. DOI: https://doi.org/10.18148/srm/2017.v11i1.6718.

Sakshaug, J.W., and M. Antoni. 2019. "Evaluating the Utility of Indirectly Linked Federal Administrative Records for Nonresponse Bias Adjustment." *Journal of Survey Statistics and Methodology* 7: 227–249. DOI: https://doi.org/10.1093/jssam/smy009.

Sakshaug, J.W., and M. Huber. 2016. "An Evaluation of Panel Nonresponse and Linkage Consent Bias in a Survey of Employees in Germany." *Journal of Survey Statistics and Methodology* 4(1): 71–93. DOI: https://doi.org/10.1093/jssam/smv034.

Sakshaug, J.W., and F. Kreuter. 2012. "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data." *Survey Research Methods* 6(2): 113–122. DOI: https://doi.org/10.18148/srm/2012.v6i2.5094.

Schaurer, I., and K. Weyandt. 2018. *GESIS Panel Technical Report: Recruitment 2016 (Wave d11 and d12)*. Cologne, Germany: GESIS – Leibniz-Institut für Sozialwissenschaften.

Scholz, J., A. Seshadri, and S. Khitatrakun. 2006. "Are Americans Saving 'Optimally' for Retirement?" *Journal of Political Economy* 114: 607–643. DOI: https://doi.org/10.1086/506335.

Short, K., and E. McArthur. 1986. "Life Events and Sample Attrition in the Survey of Income and Program Participation." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*: 200–205. Available at: http://www.asasrms.org/Proceedings/y1986f.html (accessed October 2021).

Sin, C.H. 2006. "The feasibility of using national surveys to derive samples of older people from different ethnic groups in Britain: Lessons from 'piggy-backing' on the Family Resources Survey." *International Journal of Social Research Methodology* 9(1): 15–28. DOI: https://doi.org/10.1080/13645570500435264.

Sinibaldi, J., M. Trappmann, and F. Kreuter. 2014. "Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary data or collecting interviewer observations?" *Public Opinion Quarterly* 78(2): 440–473. DOI: https://doi.org/10.1093/poq/nfu003.

Smith, T.W. 2011. "The Report of the International Workshop on Using Multi-Level Data from Sample Frames, Auxiliary Databases, Paradata and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys." *International Journal of Public Opinion Research* 23: 389–402. DOI: https://doi.org/10.1093/ijpor/edr035.

Smith, T.W., and J. Kim. 2013. "An Assessment of the Multi-Level Integrated Database Approach." *The Annals of the American Academy of Political and Social Science* 645: 185–221. DOI: https://doi.org/10.1177/0002716212463340.

Spieß, M., and M. Kroh. 2004. "Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (GSOEP)." *Research Note No. 28a*. Berlin, Germany: German Institute for Economic Research (DIW). Available at: https://www.diw.de/documents/publikationen/73/diw_01.c.43407.de/diw_datadoc_2004-001.pdf (accessed October 2021).

StataCorp. 2017. *Stata 15 Base Reference Manual*. College Station, TX: Stata Press.

Stöckinger, C., S. Kretschmer, and C. Kleinert. 2018. *Panel Attrition in NEPS Starting Cohort 6: A Description of Attrition Processes in Waves 2 to 7 with Regard to Nonresponse Bias, NEPS Survey Paper No. 35*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

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

Taylor, R., L. Conway, L. Calderwood, C. Lessof, H. Cheshire, K. Cox, and S. Scholes. 2007. *Health, wealth and lifestyles of the older population in England: The 2002 English Longitudinal Study of Ageing: Technical Report*. London: Institute of Fiscal Studies. Available at: https://ifs.org.uk/elsa/report03/w1_tech.pdf (accessed October 2021).

Trappmann, M., T. Gramlich, and A. Mosthaf. 2015. "The Effect of Events Between Waves on Panel Attrition." *Survey Research Methods* 9(1): 31–43. DOI: https://doi.org/10.18148/srm/2015.v9i1.5849.

UNECE (United Nations Economic Commission for Europe). 2007. *Register-Based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics.* Technical Report E.07.II.E.11. New York/Geneva: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf (accessed February 2020).

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer. DOI: https://doi.org/10.1007/978-3-319-93632-1_13.

Watson, N., and M. Wooden. 2009. "Identifying Factors Affecting Longitudinal Survey Response." In *Methodology of Longitudinal Surveys*, edited by P. Lynn: 157–181. West Sussex, England: John Wiley & Sons.

Weinert, S., C. Artelt, M. Prenzel, M. Senkbeil, T. Ehmke, C.H. Carstensen, and K. Lockl. 2019. "Development of Competencies Across the Life Course." In *Education as a Lifelong Process: The German National Educational Panel Study (NEPS), Edition ZfE Vol. 3*, edited by H.P. Blossfeld and H.G. Roßbach: 57–81. Wiesbaden: Springer VS. DOI: https://doi.org/10.1007/978-3-658-23162-0.

West, B.T., F. Kreuter, and M. Trappmann. 2014. "Is the Collection of Interviewer Observations Worthwhile in an Economic Panel Survey? New Evidence from the German Labor Market and Social Security (PASS) Study." *Journal of Survey Statistics and Methodology* 2: 159–181. DOI: https://doi.org/10.1093/jssam/smu002.

West, B.T., J. Wagner, F. Hubbard, and H. Gu. 2015. "The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth." *Journal of Survey Statistics and Methodology* 3: 240–264. DOI: https://doi.org/10.1093/jssam/smv004.

Young, A.F., A.J. Dobson, and J.E. Byles. 2001. "Health Services Research Using Linked Records: Who Consents and What is the Gain?" *Australian and New Zealand Journal of Public Health* 25(5): 417–420. DOI: https://doi.org/10.1111/j.1467-842X.2001.tb00284.x.

Received October 2020
Revised March 2021
Accepted May 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 865–905, http://dx.doi.org/10.2478/JOS-2021-0038

Combining Cluster Sampling and Link-Tracing Sampling to Estimate Totals and Means of Hidden Populations in Presence of Heterogeneous Probabilities of Links

Martín Humberto Félix-Medina[1]

We propose Horvitz-Thompson-like and Hájek-like estimators of the total and mean of a response variable associated with the elements of a hard-to-reach population, such as drug users and sex workers. A portion of the population is assumed to be covered by a frame of venues where the members of the population tend to gather. An initial cluster sample of elements is selected from the frame, where the clusters are the venues, and the elements in the sample are asked to name their contacts who belong to the population. The sample size is increased by including in the sample the named elements who are not in the initial sample. The proposed estimators do not use design-based inclusion probabilities, but model-based inclusion probabilities which are derived from a Rasch model and are estimated by maximum likelihood estimators. The inclusion probabilities are assumed to be heterogeneous, that is, they depend on the sampled people. Variance estimates are obtained by bootstrap and are used to construct confidence intervals. The performance of the proposed estimators and confidence intervals is evaluated by two numerical studies, one of them based on real data, and the results show that their performance is acceptable.

Key words: Capture-recapture; Hájek estimator; Horvitz-Thompson estimator; maximum likelihood estimator; snowball sampling.

1. Introduction

The problem of selecting samples from hidden or hard-to-detect populations, such as drug users, sex workers and homeless people, that allow reasonably valid statistical inferences is challenging because of the following six factors (1) lack of appropriate sampling frames for those populations; (2) rareness of those populations; (3) elusiveness of their members to be sampled; (4) difficulty in identifying their elements due to a stigmatized or illegal behavior; (5) difficulty in locating their members, and (6) difficulty in persuading their elements to participate in the study, among others. See Tourangean (2014) for a detailed discussion about these and other issues. Because of these factors, conventional sampling methods are not appropriate for this type of population, and consequently several specially tailored sampling methods that take into account the particular characteristics of those populations have been proposed. Among these methods we can mention multiplicity sampling, venue-based sampling, link-tracing sampling and capture-recapture sampling. (For descriptions of these methods, see Spreen 1992; Magnani et al. 2005; Kalton 2009;

[1] Facultad de Ciencias Fisico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Av. de Las Américas y Universitarios, Culiacán, Sinaloa 80013, México. Email: mhfelix@uas.edu.mx
Acknowledgments: This research was supported by Grant PROFAPI-2011/057 of Universidad Autónoma de Sinaloa.

Marpsat and Razafindratsima 2010; UNAIDS/WHO 2010; Lee et al. 2014; Spreen and Bogaerts 2015; Heckathorn and Cameron 2017). In addition, special estimation methods based on conventional samples, such as the scale up method (Killworth et al. 1998a, 1998b; Bernard et al. 2010; McCormick et al. 2010; Maltiel et al. 2015) or a combination of some of the previously mentioned sampling designs, such as the multiplier method (UNAIDS/WHO 2010; Johnston et al. 2013; Men and Gustafson 2017) and the one step network based method (Dombrowski et al. 2012; Khan et al. 2018) have been proposed.

It is worth noting that most of the recently published research papers on sampling from hidden populations focus on estimating the population size. (See Cheng et al. 2020 for a review and analysis, from an asymptotic approach, of most of these methods.) Thus, the scale up, multiplier and one step network methods were developed with this goal in mind. The interest in developing methods for estimating the size of a hidden population is mainly because information about this parameter allows the design of appropriate plans to address the problems associated with this type of population. However, information about other population parameters, such as average monthly spending on drugs and average age at which consumption begins in a population of drug addicts, and average weekly number of clients and average weekly income in a population of sex workers, is also important. This is because this type of information increases our knowledge about the population, and in addition, this knowledge could be used to improve the plans for its care that are based only on its population size.

On the other hand, among the sampling designs for hidden populations that allow estimating parameters different from or in addition to the population size, we have venue-based sampling and link-tracing sampling. Venue-based sampling (MacKellar et al. 1996) is a probability sampling method specifically developed to estimate the means of variables of interest, and particularly proportions. The method consists in carrying out an ethnographic study to construct a sampling frame of venues where the members of the population tend to gather. Venues are not only sites, such as bars, parks and street locations, but could be combinations of sites, days of the week and time segments. For instance, a venue could be a specific bar from 4:00 p.m. to 12:00 a.m. on Fridays and Saturdays, whereas another venue could be the same bar, over the same time segment, but from Monday to Thursday. Furthermore, some venues could be events such as gay parades. A probability sample of venues is selected, and from each chosen venue a sort of systematic sample of members of the population who are present at the venue is selected. For each sampled element the values of the variables of interest associated with that person are recorded, and in addition, information about his or her attendance to the venues in the sampling frame is obtained so that his or her inclusion probability can be estimated, and consequently, suitable unbiased weighted estimators of the means can be computed. It is evident that the estimates obtained by using this sampling design are valid only for the portion of the population that attends the venues in the frame. Therefore, the extension of the results to the entire population requires the assumption that with respect to the distributions of the variables of interest there are no differences between the members who visit the places in the frame and those who do not.

Link-tracing sampling (LTS) is an umbrella term that encompasses a set of sampling designs in which an initial sample of elements from the target population is selected and every sampled person is asked to name his or her contacts (defined according to a certain

criterion) who are also members of the target population. The elements in the initial sample form wave zero, and the named elements who are not in the initial sample form wave one. People in wave one might also be asked to name their contacts. The named elements who have not been previously sampled form wave two. The sampling procedure might continue in this way until a stopping rule is satisfied. For instance, a specified number of waves or a specified sample size. Several variants of LTS have been proposed. For example, Klovdahl (1989) developed a variant, known as random walk. Heckathorn (1997, 2002) proposed a variant of LTS, called respondent driven sampling (RDS) to estimate proportions of some subpopulations of the population of interest. It is worth noting that in subsequent works, such as Volz and Heckathorn (2008), Handcock et al. (2014) and Crawford et al. (2018) estimators of several population parameters have been proposed to be used along with this sampling design. Thompson and Frank (2000), Chow and Thompson (2003) and St. Clair and O'Connell (2012) have also proposed estimators of population parameters to be used with a pretty general variant of LTS. Finally, a variant of LTS related to the one considered in this work is the one proposed by Frank and Snijders (1994). In this variant, the initial sample is assumed to be selected by Bernoulli sampling, (Särndal et al. 1992, subsec. 3.2) that is, every element of the population has the same probability of being included in the initial sample and the inclusions are independent. Furthermore, those authors assumed that the probability that a specific element of the population be named as a contact by a particular person in the initial sample, which in this document is called link probability, is a constant, that is, it does not depend on the named person nor on the person who names. This supposition is known as the homogeneity assumption.

In this article, we consider the problem of estimating the total and the mean of a variable of interest, such as the weekly drug expense of a drug user, the number of weekly clients of a sex worker, and an indicator (positive = 1/negative = 0) of a person's drug use, from a sample selected by the LTS variant proposed by Félix-Medina and Thompson (2004). This sampling variant was devised to avoid the assumption of an initial Bernoulli sample required by the sampling variant proposed by Frank and Snijders (1994), which is difficult to satisfy in real-world applications. To achieve this goal, those authors proposed that a sampling frame of venues where the members of the population tend to gather be constructed, as in venue-based sampling. Those authors made the assumptions that the frame does not cover the whole population and that an element in the portion covered by the frame is assigned to only one venue. The last assumption can be achieved by means of a specified criterion, for instance, the venue where the person spends most of his or her time. Then, a simple random sample without replacement (SRSWOR) of venues is selected and every one of the members of the population who belongs to each sampled venue is sampled and interviewed. Notice that the assumption that each element in the frame is assigned to only one venue implies that if a person were found in a sampled place where he or she does not meet the criterion to be assigned to, that person should not be included in the initial sample. Next, from each sampled venue its elements are asked to name their contacts who are also members of the population, either they belong or not to the portion covered by the frame. Félix-Medina and Thompson (2004) proposed maximum likelihood estimators (MLEs) of the population size which were derived under the assumption that the probability that a person is linked to a sampled venue, that is, that he or she is a contact of an element in that venue, depends on the venue, but not on the named

person. This means that the estimators were derived under the assumption of homogeneous link probabilities. This assumption might be unrealistic in most actual populations because people with many social relationships may be more likely to be linked to a given venue than those with few social relationships. In order to avoid the homogeneity assumption, Félix-Medina et al. (2015) derived MLEs of the population size under the assumption that the link probabilities also depend on the named persons, that is, that the probabilities are heterogeneous. In their work, the authors showed by means of a Monte Carlo study, that if the assumption of homogeneous link probabilities is not satisfied, the estimators derived under that assumption are negatively biased. This result agrees with those reported in capture-recapture studies (see e.g., Burnham and Overton 1978; Hwang and Huggins 2005) and it should be expected as the estimators that have been proposed to be used with this LTS variant are similar to those used in capture-recapture studies. Since the LTS variant proposed by Félix-Medina and Thompson (2004) has not been used in any study with a real population, there are no results about the effect of the heterogeneity of the link probabilities on the estimators derived under the homogeneity assumption based on real populations. However, Dávid and Snijders (2002) used the variant of Frank and Snijders (1994) to estimate the number of homeless in Hungary and obtained an underestimation. They attributed this result to the failure of the initial Bernoulli sample assumption. However, it is possible that the estimate was also affected by not satisfying the assumption of homogeneity of the link probabilities on which the estimator they used is based.

In this article, we used the model proposed by Félix-Medina et al. (2015) to construct model-based Horvitz-Thompson-like estimators (HTLEs) and model-based Hájek-like estimators (HKLEs) of the total and the mean of a variable of interest. It should be noted that Félix-Medina and Monjardin (2010) also considered the problem addressed in this article, but they proposed estimators of the total and the mean derived under the assumption of homogeneous link probabilities. Thus, our work is an extension of theirs. The structure of this article is as follows. In Section 2, we introduce the LTS variant proposed by Félix-Medina and Thompson (2004), as well as the notation to be used throughout this article. In Section 3, we present the models and the MLEs of the population sizes proposed by Félix-Medina et al. (2015). In Section 4, we develop the strategy to construct the proposed model-based HTLEs and HKLEs of the total and the mean. In Section 5, we describe the construction of confidence intervals for the total and mean based on the proposed HTLEs and HKLEs of these parameters and on estimates of the standard deviations of these estimators obtained by a variant of bootstrap proposed in this article. In Section 6, we present the results of two numerical studies carried out to observe the performance of the proposed estimators and confidence intervals and to compare their performance with that of the proposed by Félix-Medina and Monjardin (2010). In Section 7, we state some conclusions and suggestions for future research. Finally, in the Appendix (Section 8) we described the technical aspects of the proposed bootstrap procedure.

2. Sampling Design and Notation

In this article, we consider the variant of LTS proposed by Félix-Medina and Thompson (2004) which we will describe next. Let U be a finite population of an unknown number τ

of people. A portion U_1 of U is assumed to be covered by a sampling frame of N venues $A_1, \ldots, A_N$, where the members of the population can be found with high probability. As in ordinary cluster sampling, each person in U_1 is assumed that can be assigned, by means of a specified criterion, to only one venue in the frame, for instance, the venue where he or she spends most of his or her time. Let m_i denote the number of members of the population that belong, that is, that are assigned to the venue $A_i, i = 1, \ldots, N$. From the previous assumption it follows that the number of people in U_1 is $\tau_1 = \sum_1^N m_i$ and the number of people in the portion $U_2 = U - U_1$ of U that is not covered by the frame is $\tau_2 = \tau - \tau_1$.

The first step of the sampling design is to select a SRSWOR S_A of n venues $A_1, \ldots, A_n$ from the frame. The m_i members of the population who belong to the sampled venue A_i are identified and their associated y-values of a variable of interest y are recorded, $i = 1, \ldots, n$. Notice that the assumption that a person in U_1 is assigned to only one venue by a specified criterion implies that if a person were found in a sampled venue where he or she does not meet the criterion to be assigned, that person should not be included in the initial sample. Let S_0 be the set of people in the initial sample. Notice that the size of S_0 is $m = \sum_1^n m_i$. The second step is to ask the people in each sampled venue to name other members of the population. We will say that a person and a venue are linked if any of the people who belong to that venue names him or her. Let $x_{ij}^{(k)} = 1$ if person $j \in U_k - A_i$ is linked to venue $A_i \in S_A$ and $x_{ij}^{(k)} = 0$ if $j \in A_i$ or j is not linked to A_i, $i = 1, \ldots, n$; $k = 1, 2$. For each named person, the following information is recorded: the value of the variable of interest y associated with him or her, the sampled venues that are linked to him or her, and the subset of U: $U_1 - S_0$, a specific $A_i \in S_A$ or U_2, that contains him or her. Let S_1 be the set of people in $U_1 - S_0$ who are linked to at least one venue in S_A, and let S_2 be the set of people in U_2 who are linked to at least one venue in S_A. We will denote by r_k the size of S_k, $k = 1, 2$. Finally, let $S_1^* = S_0 \cup S_1$ and $S_2^* = S_2$ be the sets of the sampled people from U_1 and U_2, respectively. Notice that the respective sizes of these sets are $m + r_1$ and r_2.

3. Maximum Likelihood Estimators of the Population Sizes

Félix-Medina et al. (2015) proposed MLEs of the population sizes τ_1, τ_2 and τ, which were derived from the following assumptions. The values $m_1, \ldots, m_N$ are considered as realizations of the random variables $M_1, \ldots, M_N$, which are supposed to be independent and identically distributed Poisson random variables with mean λ_1. This implies that the joint conditional distribution of the vector of variables $\mathbf{M}_s = (M_1, \ldots, M_n, \tau_1 - M)$, where $M = \sum_1^n M_t$ given that $\sum_1^n M_t = \tau_1$, is multinomial with parameter of size τ_1 and vector of probabilities $(1/N, \ldots, 1/N, 1 - n/N)$. The assumption that the M_is are independent and identically distributed Poisson random variables is not as restrictive as it seems at first glance. This assumption contributes to the likelihood function with the previously indicated multinomial distribution through the term $[\tau_1!/(\tau_1 - m)!](1 - n/N)^{\tau_1}$, which depends on the M_is by means of the value m of $M = \sum_1^n M_t$. Since NM/n is a design-based estimator of τ_1 (based only on the information contained in the initial sample), it follows that the Poisson assumption does not weaken the robustness of the maximum likelihood estimators. However, it does affect the variability of the estimators. For this reason, variance estimators and confidence intervals must be constructed without taking into account this assumption.

The values $x_{ij}^{(k)}$s are assumed to be realizations of the random variables $X_{ij}^{(k)}$s, which given the sample S_A of venues, are supposed to be independent Bernoulli random variables with means $p_{ij}^{(k)}$s, where the means or link probabilities $p_{ij}^{(k)}$s are given by the following Rasch model:

$$p_{ij}^{(k)} = \Pr\left(X_{ij}^{(k)} = 1 \middle| S_A, \alpha_{(k)i}, \beta_{(k)j}\right) = \frac{\exp\left(\alpha_{(k)i} + \beta_{(k)j}\right)}{1 + \exp\left(\alpha_{(k)i} + \beta_{(k)j}\right)}, \; j \in U_k - A_i; \; i = 1, \ldots, n. \tag{1}$$

As is indicated in Félix-Medina et al. (2015), this model was considered by Coull and Agresti (1999) in the context of multiple capture-recapture sampling. The parameter $\alpha_{(k)i}$ is a fixed (not random) effect that represents the potential that the venue A_i has of forming links with people in $U_k - A_i$, and $\beta_{(k)j}$ is a random effect that represents the propensity of the person $j \in U_k$ to be linked to a sampled venue. Those authors suppose that $\beta_{(k)j}$ is normally distributed with mean 0 and unknown variance σ_k^2 and that these variables are independent. The parameter σ_k^2 determines the degree of heterogeneity of the $p_{ij}^{(k)}$s: great values of σ_k^2 imply high degrees of heterogeneity.

Henceforth, all probability statements will be conditioned on the sample S_A of venues unless otherwise is specified. Let $\mathrm{X}_j^{(k)} = (X_{1j}^{(k)}, \ldots, X_{nj}^{(k)})$ be the n-dimensional vector of link indicator variables $\mathrm{X}_{ij}^{(k)}$s associated with the j-th person in $U_k - S_0$, and let $\Omega = \{\mathrm{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_i = 0 \text{ or } x_i = 1; \, i = 1, \ldots, n\}$. Then, the probability that $\mathrm{X}_j^{(k)}$ equals $\mathrm{x} = (x_1, \ldots, x_n) \in \Omega$, that is, the probability that the j-th person in $U_k - S_0$ is linked to only the venues $A_i \in S_A$ such that the i-th element x_i of x equals 1, is

$$\Pr\left(\mathbf{X}_j^{(k)} = \mathbf{x} | \boldsymbol{\alpha}_k, \beta_{(k)j}\right) = \prod_{i=1}^{n} [p_{ij}^{(k)}]^{x_i} [1 - p_{ij}^{(k)}]^{1-x_i} = \prod_{i=1}^{n} \frac{\exp\left[x_i(\alpha_{(k)i} + \beta_{(k)j})\right]}{1 + \exp\left(\alpha_{(k)i} + \beta_{(k)j}\right)},$$

where $\boldsymbol{\alpha}_k = (\alpha_{(k)1}, \ldots, \alpha_{(k)n})$. Therefore, the probability that the vector of link indicator variables associated with a randomly selected person in $U_k - S_0$ equals x is

$$\pi_{(k)\mathbf{x}} = \pi_{(k)\mathbf{x}}(\boldsymbol{\alpha}_k, \sigma_k) = \int \prod_{i=1}^{n} \frac{\exp\left[x_i(\alpha_{(k)i} + \sigma_k z)\right]}{1 + \exp\left(\alpha_{(k)i} + \sigma_k z\right)} \phi(z) dz,$$

where $\phi(\cdot)$ denotes the probability density function of the standard normal distribution $[N(0, 1)]$.

Félix-Medina et al. (2015), following Coull and Agresti (1999), used the Gaussian quadrature method to obtain the following approximation to $\pi_{(k)\mathrm{x}}$:

$$\tilde{\pi}_{(k)\mathbf{x}} = \tilde{\pi}_{(k)\mathbf{x}}(\boldsymbol{\alpha}_k, \sigma_k) = \sum_{t=1}^{q} \prod_{i=1}^{n} \frac{\exp\left[x_i(\alpha_{(k)i} + \sigma_k z_t)\right]}{1 + \exp\left(\alpha_{(k)i} + \sigma_k z_t\right)} \nu_t, \tag{2}$$

where q is a fixed constant and $\{z_t\}$ and $\{\nu_t\}$ are obtained from tables (see Table 25.5 in Abramowitz and Stegun 1964) or statistical software (see R library statmod developed by Giner and Smyth 2016).

Similarly, for person j in $A_i \in S_A$, let $\mathbf{X}_j^{(A_i)} = (X_{1j}^{(A_i)}, \ldots, X_{i-1j}^{(A_i)}, X_{i+1j}^{(A_i)}, \ldots, X_{nj}^{(A_i)})$ be the $(n - 1)$-dimensional vector of link indicator variables $X_{i'j}^{(A_i)}$s associated with that person, and let $\Omega_{-i} = \{\mathrm{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \mathbb{R}^{n-1} : x_j = 0 \text{ or } x_j = 1; j = 1, \ldots, n; j \neq i\}$. Then, the probability that $\mathrm{X}_j^{(Ai)}$ equals $\mathrm{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \Omega_{-i}$,

that is, the probability that the j-th person in A_i is linked to only the venues $A_{i'} \in S_A, i' \neq i$, such that the i'-th element $x_{i'}$ of x equals 1, is

$$\Pr\left(\mathbf{X}_j^{(A_i)} = \mathbf{x} | \boldsymbol{\alpha}_1, \beta_{(1)j}\right) = \prod_{i' \neq 1}^{n} [p_{i'j}^{(1)}]^{x_{i'}} [1 - p_{i'j}^{(1)}]^{1 - x_{i'}} = \prod_{i' \neq 1}^{n} \frac{\exp\left[x_{i'}(\alpha_{(1)i'} + \beta_{(1)j})\right]}{1 + \exp\left(\alpha_{(1)i'} + \beta_{(k)j}\right)}$$

and the Gaussian quadrature approximation to the probability $\pi_{(A_i)\mathrm{x}} = \pi_{(A_i)\mathrm{x}}(\boldsymbol{\alpha}_1, \sigma_1)$ that the vector of link indicator variables associated with a randomly selected person from the sampled venue A_i equals the $(n-1)$-dimensional vector $\mathrm{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ is

$$\tilde{\pi}_{(A_i)\mathbf{x}} = \tilde{\pi}_{(A_i)\mathbf{x}}(\boldsymbol{\alpha}_1, \sigma_1) = \sum_{t=1}^{q} \prod_{i' \neq i}^{n} \frac{\exp\left[x_{i'}(\alpha_{(1)i'} + \sigma_1 z_t)\right]}{1 + \exp\left(\alpha_{(1)i'} + \sigma_1 z_t\right)} v_t. \tag{3}$$

Under the previous assumptions, Félix-Medina et al. (2015) constructed the likelihood function of τ_k, $\boldsymbol{\alpha}_k$ and σ_k, $k = 1, 2$, which is proportional to a product of several multinomial distributions. One multinomial distribution is the conditional distribution of the vector of variables $\mathrm{M}_s = (M_1, \ldots, M_n, \tau_1 - M)$, given that $\sum_1^N M_i = \tau_1$, and that was indicated at the beginning of this section. Other two multinomial distributions are obtained by considering each one of the 2^n vectors $\mathrm{x} \in \Omega$ as the label of a cell of a contingency table. Then, the vectors of cell-frequencies $(R_{\mathbf{x}}^{(1)})_{\mathbf{x}\in\Omega}$ and $(R_{\mathbf{x}}^{(2)})_{\mathbf{x}\in\Omega}$, where $R_{\mathbf{x}}^{(1)}$ and $R_{\mathbf{x}}^{(2)}$ denote the numbers of people in $U_1 - S_0$ and in U_2 whose vectors $\mathrm{X}^{(1)}$ and $\mathrm{X}^{(2)}$ of link indicator variables are equal to $\mathrm{x} \in \Omega$, have joint multinomial distributions with parameters of size $\tau_1 - m$ and τ_2, and vectors of probabilities $(\pi_{(1)\mathbf{x}})_{\mathbf{x}\in\Omega}$ and $(\pi_{(2)\mathbf{x}})_{\mathbf{x}\in\Omega}$, respectively. The last n multinomial distributions, one for each $A_i \in S_A$, are obtained as the previous two. Thus, for each $i = 1, \ldots, n$, it follows that the vector of variables $(R_{\mathbf{x}}^{(A_i)})_{\mathbf{x}\in\Omega_{-i}}$, where $R_{\mathbf{x}}^{(A_i)}$ denotes the number of people in A_i whose vectors of link indicator variables $\mathrm{X}^{(A_i)}$ are equal to the vector $\mathrm{x} \in \Omega_{-i}$ has a multinomial distribution with parameter of size m_i and vector of probabilities $(\pi_{(A_i)\mathrm{x}})_{\mathrm{x}\in\Omega_{-i}}$.

Those authors proposed maximum likelihood estimators of τ_k, $\boldsymbol{\alpha}_k$ and σ_k, $k = 1, 2$, whose values are obtained by numerically maximizing the likelihood function expressed in terms of the Gaussian quadrature approximations $\tilde{\pi}_{(k)\mathrm{x}}$ and $\tilde{\pi}_{(A_i)\mathrm{x}}$ to the probabilities $\pi_{(k)\mathrm{x}}$ and $\pi_{(A_i)\mathrm{x}}$. They called these estimators unconditional maximum likelihood estimators (UMLE) and denoted them as $\hat{\tau}_k^{(U)}$, $\hat{\boldsymbol{\alpha}}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$, $k = 1, 2$. Although these estimators do not have closed forms, the authors provided the following asymptotic approximations for $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_2^{(U)}$:

$$\hat{\tau}_1^{(U)} = \frac{M + R_1}{1 - (1 - n/N)\tilde{\pi}_{(1)\mathbf{0}}\left(\hat{\alpha}_1^{(U)}, \hat{\sigma}_1^{(U)}\right)} \quad \text{and} \quad \hat{\tau}_2^{(U)} = \frac{R_2}{1 - \tilde{\pi}_{(2)\mathbf{0}}\left(\hat{\alpha}_2^{(U)}, \hat{\sigma}_2^{(U)}\right)}, \tag{4}$$

where R_1 and R_2 denote the numbers of distinct people in $U_1 - S_0$ and U_2, respectively, that are linked to at least one venue in S_A. Notice that these are not close forms because $\hat{\boldsymbol{\alpha}}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ depend on $\hat{\tau}_k^{(U)}$. Once $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_2^{(U)}$ are obtained, the UMLE of τ is $\hat{\tau}^{(U)} = \hat{\tau}_1^{(U)} + \hat{\tau}_2^{(U)}$.

Also, Félix-Medina et al. (2015), following Coull and Agresti (1999), used Sanathanan's (1972) approach to derive conditional MLEs $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$ of $\boldsymbol{\alpha}_k$ and σ_k, given R_k, $k = 1, 2$. The conditional likelihood function of these parameters is obtained by considering the conditional distribution of the vector $\left(R_{\mathbf{x}}^{(k)}\right)_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}$ given that $R_k = r_k$,

which is multinomial with parameter of size r_k and vector of probabilities $(\pi^{(c)}_{(\mathbf{k})\mathbf{x}})_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}$, where $\pi^{(c)}_{(k)\mathbf{x}} = \pi_{(k)\mathbf{x}}/[1 - \pi_{(k)\mathbf{0}}]$, $k = 1, 2$. The product of these two distributions and the n multinomial distributions of the vectors $(R^{(A_i)}_{\mathbf{x}})_{\mathbf{x}\in\Omega_{-i}}$, $i = 1, \ldots, n$, forms the conditional likelihood function. The values of the estimators $\hat{\boldsymbol{\alpha}}^{(C)}_k$ and $\hat{\sigma}^{(C)}_k$ are obtained by maximizing numerically this function expressed in terms of the Gaussian quadrature approximations to the probabilities $\pi_{(k)\mathrm{x}}$ and $\pi_{(A_i)\mathrm{x}}$. The values of the conditional estimators $\hat{\tau}^{(C)}_1$ and $\hat{\tau}^{(C)}_2$ of τ_1 and τ_2 are obtained by using the fact that the respective distributions of R_1 and R_2 are binomial with parameters of size $\tau_1 - m$ and τ_2 and probabilities $1 - \pi_{(k)\mathbf{0}}$, $k = 1, 2$. The product of these two distributions and the conditional multinomial distribution of M_s forms the likelihood function of τ_1 and τ_2. The maximization of this function, after replacing the values of $\boldsymbol{\alpha}_k$ and σ_k by their estimates $\hat{\boldsymbol{\alpha}}^{(C)}_k$ and $\hat{\sigma}^{(C)}_k$, yields the conditional MLEs $\hat{\tau}^{(C)}_1$ and $\hat{\tau}^{(C)}_2$, which are given by Equation (4), but replacing $\hat{\boldsymbol{\alpha}}^{(U)}_k$ and $\hat{\sigma}^{(U)}_k$ by $\hat{\boldsymbol{\alpha}}^{(C)}_k$ and $\hat{\sigma}^{(C)}_k$. Note that in this case the expressions (4) are closed forms. The conditional MLE of τ is $\hat{\tau}^{(C)} = \hat{\tau}^{(C)}_1 + \hat{\tau}^{(C)}_2$.

It is worth noting that for each $k = 1, 2$, the assumption that the link probabilities follow the Rasch model (1) could be tested by a Pearson chi-square statistic. One possibility is to use the previously indicated conditional multinomial distribution of $\left(R^{(k)}_{\mathbf{x}}\right)_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}$ given that $R_k = r_k$. In this case, the Pearson chi-square statistic to test the hypothesis that the link probabilities associated with the elements in $U_k - S_0$ follow the Rasch model (1) is

$$X^2_k = r_k \sum_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[R^{(k)}_{\mathbf{x}}/r_k - \tilde{\pi}^{(c)}_{(k)\mathbf{x}}\right]^2 / \tilde{\pi}^{(c)}_{(k)\mathbf{x}}, k = 1, 2,$$

where $\tilde{\pi}^{(c)}_{(k)\mathbf{x}} = \tilde{\pi}_{(k)\mathbf{x}}(\tilde{\alpha}^{(a)}_k, \tilde{\sigma}^{(a)}_k)/\left[1 - \tilde{\pi}_{(k)\mathbf{0}}\left(\hat{\boldsymbol{\alpha}}^{(a)}_k, \hat{\sigma}^{(a)}_k\right)\right]$, and $a = U$ or $a = C$ depending on whether UMLEs or CMLEs are used. Since the contingency tables on which these statistics are based are sparse, their distributions are not well approximated by chi-square distributions. One alternative to compute the p-values of these tests is by bootstrap. See Reiser (2019) for this and other alternatives. Specific assumptions about the Rasch model (1) such as the non-interaction between the link indicator variables that form the vector $\mathrm{X}^{(k)}$ could be tested using Rasch models that take into account these interactions. See Fienberg et al. (1999) for more general Rasch models than the one used in this article.

4. Estimators of the Total and Mean

In this section, we will focus on the problem of estimating the total and the mean of the values of a variable of interest y. Let $y^{(k)}_j$ be the value of y associated with the j-th element of U_k, $j = 1, \ldots, \tau_k$, $k = 1, 2$. In this work we will suppose that the y-values are fixed numbers and not random variables. Note that this assumption is the one made in traditional sampling, in which inferences are made from the design-based approach. (See Thompson 2012, 2) Then $Y_k = \sum_{j\in U_k} y^{(k)}_j$ and $\bar{Y}_k = Y_k/\tau_k$ represent the total and the mean of the portion U_k, $k = 1, 2$, of the population. Similarly, $Y = Y_1 + Y_2$ and $\bar{Y} = Y/\tau$ represent the total and the mean of the whole population U.

We cannot compute the design-based inclusion probabilities of the sampled elements because we do not know the venues in the frame that are linked to each sampled person. However, we can compute conditional model-based inclusion probabilities given the venues $A_i \in S_A$. These probabilities are given by

$$\pi_{(1)j}\left(\boldsymbol{\alpha}_1, \sigma_1, \beta_{(1)j}\right) = 1 - (1 - n/N)\prod_{i=1}^{n}(1 - p_{ij}^{(1)}) \text{ if } j \in U_1, \text{ and} \tag{5}$$

$$\pi_{(2)j}\left(\boldsymbol{\alpha}_2, \sigma_2, \beta_{(2)j}\right) = 1 - \prod_{i=1}^{n}\left(1 - p_{ij}^{(2)}\right) \text{ if } j \in U_2. \tag{6}$$

We do not know the probabilities $\pi_{(k)j}(\alpha_k, \sigma_k, \beta_{(k)j})$ because they depend on unknown parameters. However, we could estimate them by estimating those parameters and replacing the parameters in Equations (5) and (6) by their estimates. The computation of both UMLEs and CMLEs of $\boldsymbol{\alpha}_k$ and σ_k was described in the previous section. See also Félix-Medina et al. (2015). Next, we will derive a predictor of the random effect $\beta_{(k)j}$.

Thus, if the element $j \in U_k - S_0$, $k = 1, 2$, then the conditional joint probability density function of the vector $\mathrm{X}_j^{(k)}$ of link indicator variables associated with that element and the random effect $\beta_{(k)j}$ is

$$f(x_j^{(k)}, \beta_{(k)j}|\boldsymbol{\alpha}_k, \sigma_k) = \Pr(\mathrm{X}_j^{(k)} = \mathrm{x}_j^{(k)}|\beta_{(k)j}, \boldsymbol{\alpha}_k) f(\beta_{(k)j}|\sigma_k)$$

$$\propto \prod_{i=1}^{n}[p_{ij}^{(k)}]^{x_{ij}^{(k)}}[1 - p_{ij}^{(k)}]^{1-x_{ij}^{(k)}} \exp[-(\beta_{(k)j})^2/2\sigma_k^2],$$

whereas, if the element $j \in A_{i'} \in S_A$, $i' = 1, \ldots, n$, then

$$f(x_j^{(A_{i'})}, \beta_{(1)j}|\boldsymbol{\alpha}_1, \sigma_1) \propto \prod_{i \neq i'}^{n}[p_{ij}^{(1)}]^{x_{ij}^{(A_{i'})}}[1 - p_{ij}^{(1)}]^{1-x_{ij}^{(A_{i'})}} \exp[-(\beta_{(1)j})^2/2\sigma_1^2].$$

We will propose as a predictor of $\beta_{(k)j}$ the conditional expected value of $\beta_{(k)j}$ given $\mathrm{X}_j^{(k)} = \mathrm{x}_j^{(k)}$, evaluated either at the UMLEs $\hat{\boldsymbol{\alpha}}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ or at the CMLEs $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$, that is

$$\hat{\tilde{\beta}}_{(k)_j}^{(a)} = E\left(\beta_{(k)j}|\mathbf{x}_j^{(k)}, \hat{\boldsymbol{\alpha}}_k^{(a)}, \hat{\sigma}_k^{(a)}\right) = \frac{\int \beta_{(k)j} f\left(\mathbf{x}_j^{(k)}, \beta_{(k)j}|\hat{\boldsymbol{\alpha}}_k^{(a)}, \hat{\sigma}_k^{(a)}\right) d\beta_{(k)j}}{\int f\left(\mathbf{x}_j^{(k)}, \beta_{(k)j}|\hat{\boldsymbol{\alpha}}_k^{(a)}, \hat{\sigma}_k^{(a)}\right) d\beta_{(k)j}}, \quad a = U, C.$$

We will approximate $\hat{\tilde{\beta}}_{(k)j}^{(a)}$ by using the Gaussian quadrature method, that is by

$$\hat{\tilde{\beta}}_{(k)j}^{(a)} = \frac{\hat{\sigma}_k^{(a)}\sum_{t=1}^{q} z_t \prod_{i=1}^{n}\left\{\exp\left[x_i(\hat{\alpha}_{(k)i}^{(a)} + \hat{\sigma}_k^{(a)} z_t)\right]/\left[1 + \exp(\hat{\alpha}_{(k)i}^{(a)} + \hat{\sigma}_k^{(a)} z_t)\right]\right\} v_t}{\sum_{t=1}^{q}\prod_{i=1}^{n}\left\{\exp\left[x_i(\hat{\alpha}_{(k)i}^{(a)} + \hat{\sigma}_k^{(a)} z_t)\right]/\left[1 + \exp(\hat{\alpha}_{(k)i}^{(a)} + \hat{\sigma}_k^{(a)} z_t)\right]\right\} v_t} \tag{7}$$

$$= \frac{\hat{\sigma}_k^{(a)}\sum_{t=1}^{q} z_t \left\{\exp\left(\hat{\sigma}_k^{(a)} z_t \sum_{i=1}^{n} x_i\right)/\prod_{i=1}^{n}\left[1 + \exp(\hat{\alpha}_{(k)_i}^{(a)} + \hat{\sigma}_k^{(a)} z_t)\right]\right\} v_t}{\sum_{t=1}^{q}\left\{\exp\left(\hat{\sigma}_k^{(a)} z_t \sum_{i=1}^{n} x_i\right)/\prod_{i=1}^{n}\left[1 + \exp(\hat{\alpha}_{(k)_i}^{(a)} + \hat{\sigma}_k^{(a)} z_t)\right]\right\} v_t}, a = U, C,$$

if $j \in U_k - S_0$, $k = 1, 2$, and by

$$\tilde{\tilde{\beta}}_{(1)j}^{(a)} = \frac{\hat{\sigma}_1^{(a)} \sum_{t=1}^{q} z_t \left\{ \exp\left(\hat{\sigma}_1^{(a)} z_t \sum_{i \neq i'}^{n} x_i \right) / \prod_{i \neq i'}^{n} \left[1 + \exp\left(\hat{\alpha}_{(1)i}^{(a)} + \hat{\sigma}_1^{(a)} z_t \right) \right] \right\} v_t}{\sum_{t=1}^{q} \left\{ \exp\left(\hat{\sigma}_1^{(a)} z_t \sum_{i \neq i'}^{n} x_i \right) / \prod_{i \neq i'}^{n} \left[1 + \exp\left(\hat{\alpha}_{(1)i}^{(a)} + \hat{\sigma}_1^{(a)} z_t \right) \right] \right\} v_t}, a = U, C,$$

if $j \in A_{i'} \in S_{\mathrm{A}}$, $i' = 1, \ldots, n$.

The previous expressions imply that $\tilde{\tilde{\beta}}_{(1)j}^{(a)}$ depends on the x_is through their sum, that is, on the number of venues that are linked to the element j, but not on the particular venues to which that element is linked. Thus, if two persons j and j' in $U_k - S_0$ are linked to the same number of venues in S_A, the predictors $\tilde{\tilde{\beta}}_{(k)j}$ and $\tilde{\tilde{\beta}}_{(k)j'}$ are equal one another. The same happens for two persons in $A_i \in S_A$.

Thus, model-based Horvitz-Thompson-like estimators (HTLEs) of the totals $Y_k, k = 1, 2$, and Y are

$$\hat{Y}_{HT.k}^{(a)} = \sum_{j \in S_k^*} y_j^{(k)} / \hat{\pi}_{(k)j}^{(a)} \left(\hat{\boldsymbol{\alpha}}_k^{(a)}, \hat{\sigma}_k^{(a)}, \tilde{\tilde{\beta}}_{(k)j}^{(a)} \right), k = 1, 2, \text{ and } \hat{Y}_{HK}^{(a)} = \hat{Y}_{HT.1}^{(a)} + \hat{Y}_{HT.2}^{(a)}, a = U, C. \tag{8}$$

Similarly, model-based HTLEs of the means $\bar{Y}_k$ and $\bar{Y}$ are

$$\hat{\bar{Y}}_{HT.k}^{(a)} = \hat{Y}_{HT.k}^{(a)} / \hat{\tau}_k^{(a)}, k = 1, 2, \quad \text{and} \quad \hat{\bar{Y}}_{HT}^{(a)} = \hat{Y}_{HT}^{(a)} / \hat{\tau}^{(a)}, \ a = U, C.$$

Notice that if we set $y_j^{(k)} = 1$, for $j = 1, \ldots, \tau_k$, then $Y_k = \tau_k$, $k = 1, 2$, and $Y = \tau$. Therefore, HTLEs of τ_k and τ are $\hat{\tau}_{HT.k}^{(a)} = \hat{Y}_{HT.k}^{(a)}, k = 1, 2$, and $\hat{\tau}_{HT}^{(a)} = \hat{Y}_{HT}^{(a)}, a = U, C$, where $\hat{Y}_{HT.k}^{(a)}$ and $\hat{Y}_{HT}^{(a)}$ are given by (8) with $y_j^{(k)} = 1$.

We could also define Hájek-like estimators (HKLEs) of the population totals and means. Thus, HKLEs of the means $\bar{Y}_k$ and $\bar{Y}$ are

$$\hat{\bar{Y}}_{HK.k}^{(a)} = \hat{Y}_{HT.k}^{(a)} / \hat{\tau}_{HT.k}^{(a)}, k = 1, 2, \quad \text{and} \quad \hat{\bar{Y}}_{HK}^{(a)} = \hat{Y}_{HT}^{(a)} / \hat{\tau}_{HT}^{(a)}, \ a = U, C,$$

and HKLEs of the totals Y_k and Y are

$$\hat{Y}_{HK.k}^{(a)} = \hat{\bar{Y}}_{HK.k}^{(a)} \hat{\tau}_k^{(a)}, k = 1, 2, \quad \text{and} \quad \hat{Y}_{HK}^{(a)} = \hat{\bar{Y}}_{HK}^{(a)} \hat{\tau}^{(a)}, \ a = U, C.$$

5. Bootstrap Variance Estimators and Confidence Intervals

We propose the use of bootstrap to construct estimators of the variances of the proposed estimators of the totals and means, as well as confidence intervals (CIs) for those population parameters. The proposed bootstrap variant is obtained by combining the bootstrap version for finite populations proposed by Booth et al. (1994) and the parametric bootstrap variant (see Davison and Hinkley, 1997, chap. 2). This version of bootstrap is an extension of the one used by Félix-Medina et al. (2015) to construct CIs based on their proposed MLEs of the population sizes. We will describe the technical aspects of the proposed boostrap variant in Appendix and for now it is enough to indicate that after applying that procedure to an estimator $\hat{\theta}$ of a population parameter θ, we will get a bootstrap sample of B values $\hat{\theta}_1, \ldots \hat{\theta}_B$.

To construct the CIs for the population totals and means we could use any of the different bootstrap alternatives that have been proposed. For instance, if we did not want to assume any probability distribution for an estimator, we could use the basic or the percentile method. (See Davison and Hinkley 1997, chap. 5) Although this type of alternative has good properties of robustness, it requires a large number B of bootstrap samples, say $B \geq 1000$, and this might be a serious problem if the estimator requires much time to be computed. On the other hand, if we were willing to assume a distribution probability for an estimator, we could use the B values $\hat{\theta}_1, \ldots \hat{\theta}_B$ to estimate the variance of $\hat{\theta}$ and construct the CI using the estimated variance and the assumed distribution of $\hat{\theta}$. In this case, the number B of required bootstrap samples is not so large, say $50 \leq B \leq 200$ is generally enough. We will follow this approach, using some ideas taken from Félix-Medina et al. (2015).

Thus, as in that article, we will estimate the variance of an estimator $\hat{\theta}$ of the population parameter θ, by using Huber's proposal 2 to jointly estimate the parameters of location and scale from the bootstrap sample of B values $\hat{\theta}_b$. (See Staudte and Sheather 1990, subsec. 4.5). In particular, the estimate of the parameter of scale is an estimate of the standard deviation $\sqrt{\hat{V}(\hat{\theta})}$ of $\hat{\theta}$. The idea behind the use of this estimator is that it yields an estimate of the standard deviation that is robust to very large values $\hat{\theta}_b$ which are likely to occur.

To construct the CIs we will use the following approach. (1) If the parameter is τ_k, $k = 1$, 2, or τ, then, as in Félix-Medina et al. (2015), we will assume that $\hat{\tau}_k^{(a)} - \nu_k$ is lognormally distributed, where $\hat{\tau}_k^{(a)}$, $a = U, C$, is an estimator of τ_k and ν_k is the number of sampled elements from U_k. Thus, a CI for τ_k is $(\nu_k + (\hat{\tau}_k^{(a)} - \nu_k)/c_k,\ \nu_k + (\hat{\tau}_k^{(a)} - \nu_k) \times c_k)$, where $c_k = \exp\left\{z_{\alpha/2}\sqrt{\ln\left[1 + \hat{V}(\hat{\tau}_k^{(a)})/(\hat{\tau}_k^{(a)} - \nu_k)^2\right]}\right\}$, $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution and $\hat{V}(\hat{\tau}_k^{(a)})$ is an estimate of the variance of $\hat{\tau}_k^{(a)}$. A CI for τ is built analogously. The values of ν_1, ν_2 and ν that are used in the CIs for τ_1, τ_2 and τ are $m + r_1$, r_2 and $m + r_1 + r_2$, respectively. This type of CI was considered by Chao (1987), in the context of capture-recapture studies, as an alternative to the ordinary Wald CI to take into account the usually right skewed distribution of the estimator of the size and avoid the problem of obtaining CIs with lower bounds smaller than the number of sampled elements. She found that this CI generally performs better that the Wald CI. (See Williams et al., 2002, subsec. 14.2). (2) If the parameter is $\bar{Y}_k$, $k = 1$, 2, or $\bar{Y}$, and it is a proportion, that is, the y-value associated with an element is equal to one if the element has a characteristic of interest and is equal to zero otherwise, then we will assume that the number of sampled elements with the characteristic of interest has a binomial distribution and a CI for $\bar{Y}$ will be constructed using the proposal of Korn and Graubard (1998), which is an adaptation of the Clopper-Pearson CI for a proportion in the case of complex samples. Thus, a CI for $\bar{Y}_k$ is $\left(v_{(k)1}^{(a)} F_{v_{(k)1}^{(a)}, v_{(k)2}^{(a)}}(\alpha/2)/[v_{(k)2}^{(a)} + v_{(k)1}^{(a)} F_{v_{(k)1}^{(a)}, v_{(k)2}^{(a)}}(\alpha/2)],\ v_{(k)3}^{(a)} F_{v_{(k)3}^{(a)}, v_{(k)4}^{(a)}}(1-\alpha/2)/[v_{(k)4}^{(a)} + v_{(k)3}^{(a)} F_{v_{(k)3}^{(a)}, v_{(k)4}^{(a)}}(1-\alpha/2)]\right)$, where $v_{(k)1}^{(a)} = 2y_k^{(a)}$, $v_{(k)2}^{(a)} = 2(n_k^{(a)} - y_k^{(a)} + 1)$, $v_{(k)3}^{(a)} = 2(n_k^{(a)} + 1)$, $v_{(k)4}^{(a)} = 2(n_k^{(a)} - y_k^{(a)})$, $y_k^{(a)} = n_k^{(a)} \hat{\bar{Y}}_k^{(a)}$, $n_k^{(a)} = \hat{\bar{Y}}_k^{(a)}(1 - \hat{\bar{Y}}_k^{(a)})/\hat{V}(\hat{\bar{Y}}_k^{(a)})$, $\hat{\bar{Y}}_k^{(a)}$ is an estimator of $\bar{Y}_k$, $\hat{V}(\hat{\bar{Y}}_k^{(a)})$ is an estimate of the variance of $\hat{\bar{Y}}_k^{(a)}$ and $F_{d_1,d_2}(\beta)$ is the β quantile of the F distribution with d_1 and d_2 degrees of freedom. A CI for $\bar{Y}$ is built analogously. (3) If the parameter is $\bar{Y}_k$, $k = 1$, 2, or $\bar{Y}$, and it is the mean of the y-values a continuous variable of interest or if it is Y_k, $k = 1$, 2, or Y, that is, it is the total of the

y-values of a continuous or a binary variable of interest, we will assume that the estimator $\hat{\bar{Y}}_k^{(a)}$ (or $\hat{Y}_k^{(a)}$) is normally distributed. Thus, a CI for $\bar{Y}_k$ is $\left(\hat{\bar{Y}}_k^{(a)} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\bar{Y}}_k^{(a)})}, \hat{\bar{Y}}_k^{(a)} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\bar{Y}}_k^{(a)})}\right)$, where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution and $\hat{V}(\hat{\bar{Y}}_k^{(a)})$ is an estimate of the variance of $\hat{\bar{Y}}_k^{(a)}$. CIs for Y_k, $\bar{Y}$ and Y are built analogously.

6. Monte Carlo Studies

In order to observe the performance of the proposed estimators and CIs and to compare their performance with the ones proposed by Félix-Medina and Monjardin (2010), which were derived under the assumption of homogeneity of the link probabilities, we carried out two numerical studies. In the first study, we used artificial data to construct two populations with specific characteristics, whereas in the second one we used data from the National Longitudinal Study of Adolescent Health collected during the 1994–1995 school year to construct a population. Both studies were carried out using the R software environment for statistical computing (R Core Team, 2018).

6.1. *Populations Constructed Using Artificial Data*

We constructed two populations whose characteristics are described in Table 1. The difference between the two populations is that in Population I the link probabilities were generated by using Expression (1), that is, under the assumed model, whereas in Population II they were generated by the following latent-class model used by Pledger (2000) in the context of capture-recapture studies: $p_{ij}^{(k)} = \exp[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}]/\{1 + \exp[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}]\}$, $i = 1, \ldots, n$; $j = 1, 2$, and $k = 1, 2$. In this model, the people in U_k are divided into two classes according to their propensities to be linked to the sample clusters. The probability that a person in U_k is in class j is $p_j^{(k)}$ and it is the same for each person in U_k. The values of the parameters that appear in each of the two expressions of the link probabilities were set so that when the size of the initial sample of clusters is $n = 15$, in both populations the sampling fractions were $f_1 = 0.5$ in U_1 and $f_2 = 0.4$ in U_2. Note from Table 1 that associated with each element of each population, there are two values of two response variables. One variable is a continuous variable whose value associated with the j-th element of U_k was obtained by sampling from a non-central chi-square distribution with two degrees of freedom and non-centrality parameter $\psi_j^{(k)}$. The other variable is a binary variable whose value associated with that element was obtained from a Bernoulli distribution with mean $\varphi_j^{(k)}$. The values of the parameters that appear in the expressions of $\psi_j^{(k)}$ and $\phi_j^{(k)}$ were set so that the values of the population means of the continuous variable in both populations were $\bar{Y}_1 \approx 50$ and $\bar{Y}_2 \approx 40$, whereas the corresponding values of the binary variable were $\bar{Y}_1 \approx 0.3$ and $\bar{Y}_2 \approx 0.2$. Furthermore, they were set so that for $n = 15$, in Population I the values of the Pearson correlation coefficients between the values of the continuous response variable and those of the inclusion probabilities were $\rho(y^{(1)}, \pi_{(1)}) \approx 0.8$ and $\rho(y^{(2)}, \pi_{(2)}) \approx 0.7$, whereas the corresponding values for the binary response variable were $\rho(y^{(1)}, \pi_{(1)}) \approx 0.3$ and $\rho(y^{(2)}, \pi_{(2)}) \approx 0.27$. In the case of Population II and continuous response variable those values were $\rho(y^{(1)}, \pi_{(1)}) \approx 0.15$ and $\rho(y^{(2)}, \pi_{(2)}) \approx 0.1$, whereas the corresponding values for the binary variable were $\rho(y^{(1)}, \pi_{(1)}) \approx 0$ and $\rho(y^{(2)}, \pi_{(2)}) \approx 0.1$.

Table 1. Parameters of simulated populations

Population I	Population II
$N = 150$	$N = 150$
$M_i \sim$ zero truncated neg. binom. distribution	$M_i \sim$ zero truncated neg. binom. distribution
$E(M_i) = 8, \quad V(M_i) = 24$	$E(M_i) = 8, \quad V(M_i) = 24$
$\tau_1 = 1208,\ \tau_2 = 400,\ \tau = 1608$	$\tau_1 = 1208,\ \tau_2 = 400,\ \tau = 1608$
Rasch model for $p_{ij}^{(k)}$:	Latent class model for $p_{ij}^{(k)}$:
$p_{ij}^{(k)} = \frac{\exp\left(\alpha_i^{(k)}+\beta_j^{(k)}\right)}{1+\exp\left(\alpha_i^{(k)}+\beta_j^{(k)}\right)}$	$p_{ij}^{(k)} = \frac{\exp\left(\mu^{(k)}+\alpha_i^{(k)}+\beta_j^{(k)}+(\alpha\beta)_{ij}^{(k)}\right)}{1+\exp\left(\mu^{(k)}+\alpha_i^{(k)}+\beta_j^{(k)}+(\alpha\beta)_{ij}^{(k)}\right)}$
$i = 1, \ldots, N,\ j = 1, \ldots, \tau_k,\ k = 1, 2$	$i = 1, \ldots, N,\ j = 1, 2,\ k = 1, 2$
	$\mu^{(1)} = 0.25, \quad \mu^{(2)} = 0.05$
$\alpha_i^{(k)} = \frac{c_k}{0.001+M_i^{1/4}},\ i = 1, \ldots, N,\ k = 1, 2$	$\alpha_i^{(k)} = \frac{c_k}{0.001+M_i^{1/4}},\ i = 1, \ldots, N,\ k = 1, 2$
$c_1 = -5.45, \quad c_2 = -5.85$	$c_1 = -12, \quad c_2 = -12$
$\beta_j^{(k)} \sim N(0, 1),\ j = 1, \ldots, \tau_k$	$\beta_1^{(k)} = 1.5,\ \beta_2^{(k)} = 0,\ k = 1, 2$
	$(\alpha\beta)_{i1}^{(k)} \sim N(0, 1.25^2),\ (\alpha\beta)_{i2}^{(k)} = 0,\ k = 1, 2$
	$i = 1, \ldots, N$
	$p_1^{(k)} = 0.3,\ p_2^{(k)} = 0.7,\ k = 1, 2$
Continuous response variable:	Continuous response variable:
$Y_j^{(k)} \sim \chi^2_2(\psi_j^{(k)}),\ j = 1, \ldots, \tau_k,\ k = 1, 2$	$Y_j^{(k)} \sim \chi^2_2(\psi_j^{(k)}),\ j = 1, \ldots, \tau_k,\ k = 1, 2$
$\psi_j^{(k)} = 5 + \frac{d_k \exp(\beta_j^{(k)})}{1+\exp(\beta_j^{(k)})},\ j = 1, \ldots, \tau_k,$ $k = 1, 2$	$\psi_j^{(k)} = 5 + \frac{d_k \exp(\mu_k+\beta_j^{(k)})}{1+\exp(\mu_k+\beta_j^{(k)})},\ j = 1, \ldots, \tau_k,$ $k = 1, 2$
$d_1 = 87, \quad d_2 = 65$	$d_1 = 65.05, \quad d_2 = 50.05$
$Y_1 = 60390.34,\ Y_2 = 15945.89,\ Y = 76336.22$	$Y_1 = 60289.03,\ Y_2 = 16112.78,\ Y = 76401.81$
$\bar{Y}_1 = 49.99,\ \bar{Y}_2 = 39.87,\ \bar{Y} = 47.47$	$\bar{Y}_1 = 49.91,\ \bar{Y}_2 = 40.28,\ \bar{Y} = 47.51$
$\rho(y^{(1)}, \pi_{(1)}) = 0.79, \quad \rho(y^{(2)}, \pi_{(2)}) = 0.72$	$\rho(y^{(1)}, \pi_{(1)}) = 0.16, \quad \rho(y^{(2)}, \pi_{(2)}) = 0.11$
Binary response variable:	Binary response variable:
$Y_j^{(k)} \sim \text{Bernoulli}(\phi_j^{(k)}),\ j = 1, \ldots, \tau_k,\ k = 1, 2$	$Y_j^{(k)} \sim \text{Bernoulli}(\phi_j^{(k)}),\ j = 1, \ldots, \tau_k,\ k = 1, 2$
$\phi_j^{(k)} = \frac{g_k \exp(\beta_j^{(k)})}{1+\exp(\beta_j^{(k)})},\ j = 1, \ldots, \tau_k,\ k = 1, 2$	$\phi_j^{(k)} = \frac{g_k \exp(\mu_k+\beta_j^{(k)})}{1+\exp(\mu_k+\beta_j^{(k)})},\ j = 1, \ldots, \tau_k,\ k = 1, 2$
$g_1 = 0.6, \quad g_2 = 0.39$	$g_1 = 0.46, \quad g_2 = 0.33$
$Y_1 = 366,\ Y_2 = 82,\ Y = 448$	$Y_1 = 365,\ Y_2 = 79,\ Y = 444$
$\bar{Y}_1 = 0.303,\ \bar{Y}_2 = 0.205,\ \bar{Y} = 0.279$	$\bar{Y}_1 = 0.302,\ \bar{Y}_2 = 0.198,\ \bar{Y} = 0.276$
$\rho(y^{(1)}, \pi_{(1)}) = 0.29, \quad \rho(y^{(2)}, \pi_{(2)}) = 0.27$	$\rho(y^{(1)}, \pi_{(1)}) = -0.01, \quad \rho(y^{(2)}, \pi_{(2)}) = 0.09$

Note: Correlations coefficients $\rho(y^{(k)}, \pi_{(k)})$ were computed assuming a sample with initial size $n = 15$.

The Monte Carlo study was carried out by repeatedly selecting r samples from the population U using the sampling design described in Section 2. Thus, a SRSWOR of $n = 15$ values m_i was selected from the population of $N = 150$ values. For each selected value m_i, the value $x_{ij}^{(k)}$ of the link indicator variable $X_{ij}^{(k)}$ was generated from the Bernoulli distribution with mean $p_{ij}^{(k)}$ (see its expression in Table 1). From each selected sample, the estimators indicated in Table 2 were computed. The performance of an estimator $\hat{\theta}$ of a parameter θ was evaluated by its relative bias (r-bias), the square root of its relative mean

Table 2. Point estimators and standard deviation estimators included in the Monte Carlo studies.

Estimators of population sizes			
Type of estimator	Notation	Type of standard deviation estimator	Proposed by
UMLEs	$\hat{\tau}_1^{(U)}$, $\hat{\tau}_2^{(U)}$ and $\hat{\tau}^{(U)}$	Bootstrap	Félix-Medina et al. (2015)
CMLEs	$\hat{\tau}_1^{(C)}$, $\hat{\tau}_2^{(C)}$ and $\hat{\tau}^{(C)}$		
HTLEs based on UMLEs of inclusion probabilities	$\hat{\tau}_{HT.1}^{(U)}$,$\hat{\tau}_{HT.2}^{(U)}$ and $\hat{\tau}_{HT}^{(U)}$	Bootstrap	This work
HTLEs based on CMLEs of inclusion probabilities	$\hat{\tau}_{HT.1}^{(C)}$,$\hat{\tau}_{HT.2}^{(C)}$ and $\hat{\tau}_{HT}^{(C)}$		
MLEs derived under the homogeneity assumption	$\hat{\tau}_{ML.1}^{(H)}$ $\hat{\tau}_{ML.2}^{(H)}$ and $\hat{\tau}_{ML}^{(H)}$	Linearization	Félix-Medina and Monjardin (2006)
Bayesian-assisted estimators derived under the homogeneity assumption	$\hat{\tau}_{BA.1}^{(H)}$,$\hat{\tau}_{BA.2}^{(H)}$ and $\hat{\tau}_{BA}^{(H)}$		
Estimators of population totals and means			
Type of estimator	Notation	Type of standard deviation estimator	Proposed by
HTLEs based on UMLEs of inclusion probabilities	$\hat{Y}_{HT.1}^{(U)}$, $\hat{Y}_{HT.2}^{(U)}$ and $\hat{Y}_{HT}^{(U)}$ $\hat{\bar{Y}}_{HT.1}^{(U)}$, $\hat{\bar{Y}}_{HT.2}^{(U)}$ and $\hat{\bar{Y}}_{HT}^{(U)}$	Bootstrap	This work
HTLEs based on CMLEs of inclusion probabilities	$\hat{Y}_{HT.1}^{(C)}$, $\hat{Y}_{HT.2}^{(C)}$ and $\hat{Y}_{HT}^{(C)}$ $\hat{\bar{Y}}_{HT.1}^{(C)}$, $\hat{\bar{Y}}_{HT.2}^{(C)}$ and $\hat{\bar{Y}}_{HT}^{(C)}$		
HKLEs based on UMLEs of inclusion probabilities	$\hat{Y}_{HK.1}^{(U)}$, $\hat{Y}_{HK.2}^{(U)}$ and $\hat{Y}_{HK}^{(U)}$ $\hat{\bar{Y}}_{HK.1}^{(U)}$, $\hat{\bar{Y}}_{HK.2}^{(U)}$ and $\hat{\bar{Y}}_{\mathrm{HK}}^{(\mathrm{U})}$		
HKLEs based on CMLEs of inclusion probabilities	$\hat{Y}_{HK.1}^{(C)}$, $\hat{Y}_{HK.2}^{(C)}$ and $\hat{Y}_{HK}^{(C)}$ $\hat{\bar{Y}}_{HK.1}^{(C)}$, $\hat{\bar{Y}}_{HK.2}^{(C)}$ and $\hat{\bar{Y}}_{HK}^{(C)}$		
HTLEs based on MLEs and derived under the homogeneity assumption	$\hat{Y}_{ML.1}^{(H)}$, $\hat{Y}_{ML.2}^{(H)}$ and $\hat{Y}_{ML}^{(H)}$ $\hat{\bar{Y}}_{ML.1}^{(H)}$, $\hat{\bar{Y}}_{ML.2}^{(H)}$ and $\hat{\bar{Y}}_{ML}^{(H)}$	Linearization	Félix-Medina and Monjardin (2010)
HTLEs based on Bayesian assisted estimators derived under the homogeneity assumption	$\hat{Y}_{BA.1}^{(H)}$, $\hat{Y}_{BA.2}^{(H)}$ and $\hat{Y}_{BA}^{(H)}$ $\hat{\bar{Y}}_{BA.1}^{(H)}$, $\hat{\bar{Y}}_{BA.2}^{(H)}$ and $\hat{\bar{Y}}_{BA}^{(H)}$		

square error ($\sqrt{\text{r-mse}}$), the median of its relative estimation error (mdre), and the median of its absolute relative estimation error (mdare) defined as r-bias $= \sum_1^r(\hat{\theta}_i - \theta)/(r\theta)$, $\sqrt{\text{r-mse}} = \sqrt{\sum_1^r(\hat{\theta}_i - \theta)^2/(r\theta^2)}$, mdre $=$ median$\{(\hat{\theta}_i - \theta)/\theta\}$ and mdare $=$ median $\{|(\hat{\theta}_i - \theta)/\theta|\}$, respectively, where $\hat{\theta}_i$, is the value of $\hat{\theta}$ obtained in the i-th sample, $i = 1, \ldots, r$. In the case of the point estimators of the population sizes, totals and means their performance was evaluated using $r = 5{,}000$ samples.

We also computed estimators of the standard deviation of the point estimators, which are indicated in Table 2. The performance of a standard deviation estimator $\widehat{sd}(\hat{\theta})$ of the standard deviation $sd(\hat{\theta})$ of $\hat{\theta}$ was also evaluated by its r-bias, $\sqrt{\text{r-mse}}$, mdre and mdare, where $sd(\hat{\theta})$ was computed by the sample standard deviation of the $\hat{\theta}_i$, $i = 1, \ldots, r$. Because of the time required to compute the bootstrap standard deviation estimators, we used bootstrap samples of size $B = 50$ and their performance was evaluated using $r = 500$ samples, whereas in the case of the linearization standard deviation estimators using $r = 5{,}000$ samples.

From each point estimator and its associated standard deviation estimator a 95% CI was compute for the corresponding parameter. In the case of the estimators based on the assumption of heterogeneous link probabilities the CIs were computed as was described in Section 5, whereas in the case of those based on the homogeneity assumption the CIs were Wald type CIs. The performance of a CI was evaluated by its coverage probability (cp) defined as the proportion of samples in which the parameter is inside the interval, and by both its mean relative length (mrl) and median relative length (mdrl) defined as the sample mean and median of the lengths of the r intervals divided by the value of the parameter, respectively. In the case of the CIs based on point estimators derived under the assumption of heterogeneous link probabilities, their performance was evaluated using $r = 500$ samples, whereas in the case of the estimators derived under the assumption of homogeneous link probabilities using $r = 5{,}000$ samples.

In this and in the following study, and in the case of the estimators that were derived under the assumption of heterogeneous link probabilities, we only present the outcomes corresponding to the estimators based on the UMLEs of these probabilities because their performance was very similar to that of the estimators based on the CMLEs of the link probabilities. In the case of the estimators derived under the assumption of homogeneous link probabilities, we only present the outcomes corresponding to the estimators based on the MLEs of these probabilities because their performance was very similar to that of the estimators based on the Bayesian assisted estimators of the link probabilities. In addition, in the descriptions of the results of the numerical studies we will use the convention that the performance of a point estimator will be considered as acceptable if both its r-bias (or mdre) and its $\sqrt{\text{r-mse}}$ (or mdare) are around or are lesser than 0.1. Similarly, we will use the convention that the performance of a 95% CI is acceptable if its cp is around or greater than 0.9 and its mrl (or mdrl) is around or is lesser than 0.4 ($= 4 \times 0.1$).

The results of the study about the inferences on the population sizes are shown in Figure 1 and in Table 3. We see that in both populations, the distributions of the UMLE $\hat{\tau}_1^{(U)}$ were symmetrical, and those of $\hat{\tau}_2^{(U)}$ and $\hat{\tau}^{(U)}$ were skewed to the right with long tails. In Population I, the UMLEs did not show bias problems, but in Population II, they presented slight negative biases, except for the estimator $\hat{\tau}_2^{(U)}$ which presented a

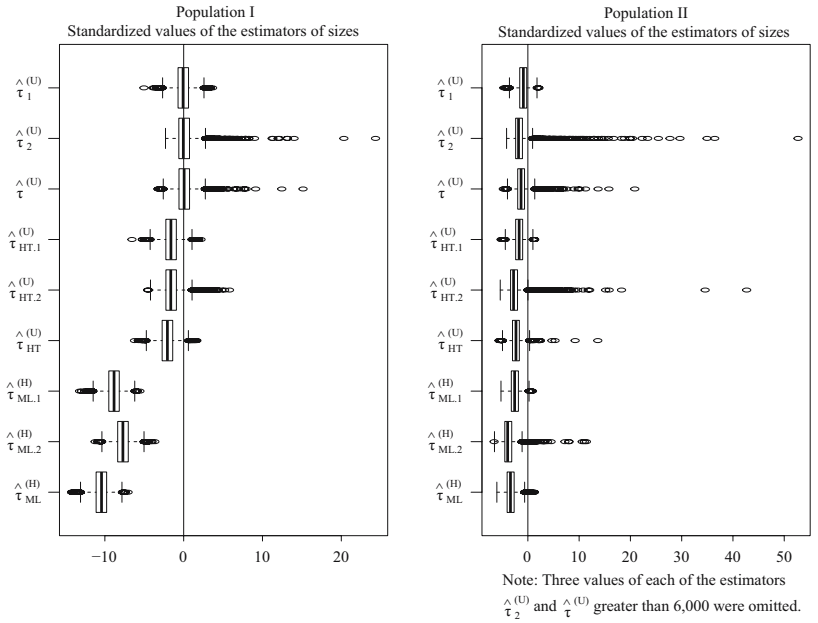

Fig. 1. Boxplots of standardized values of the estimators of sizes in Populations I and II.

moderate negative bias. In both populations, the bootstrap estimators of the standard deviations of the UMLEs showed problems of bias. The confidence intervals based on these estimators had good coverage probabilities in Population I, although the interval for τ_2 was very long, whereas in Population II their coverage probabilities were small. The HTLEs derived under the heterogeneity assumption exhibited approximately symmetric distributions, except for the estimators $\hat{\tau}_{HT.2}^{(U)}$ and $\hat{\tau}_{HT}^{(U)}$ which exhibited right skewed distributions in Population II. In both populations these estimators underestimated, with moderate biases, their corresponding parameters. The bootstrap estimators of their standard deviations exhibited bias problems, and the confidence intervals based on these estimators presented low coverage probabilities, especially in the case of Population II, where they were very small. Finally, the MLEs derived under the homogeneity assumption seriously underestimated their corresponding sizes. Their large biases affected the performance of the confidence intervals based on these estimators in such a way that their coverage probabilities were practically null. Thus, in summary, the UMLEs had the best performance, which was good in Population I, and regular in Population II. The MLEs derived under the homogeneity assumption had the worst performance, which was poor.

The results of the study on the point estimators of totals and means are shown in Figure 2 and in Table 4. We see that in Population I, and regardless of the type of variable, continuous or binary, the distributions of the HTLEs of the totals were relatively symmetrical, whereas in Population II their distributions were skewed to the right, except for that of $\hat{Y}_{HT.1}^{(U)}$ which was approximately symmetric. These estimators did not have problems of bias in Population I, and presented moderate negative biases in Population II. The distributions of the HKLEs of the totals were right skewed, except for

Table 3. Simulation results on point estimators, standard deviation estimators and 95% confidence intervals for population sizes.

Results on point estimators			UMLEs			HTLEs			Homogeneous MLEs		
			$\hat{\tau}_1^{(U)}$	$\hat{\tau}_2^{(U)}$	$\hat{\tau}^{(U)}$	$\hat{\tau}_{HT.1}^{(U)}$	$\hat{\tau}_{HT.2}^{(U)}$	$\hat{\tau}_{HT}^{(U)}$	$\hat{\tau}_{ML.1}^{(H)}$	$\hat{\tau}_{ML.2}^{(H)}$	$\hat{\tau}_{ML}^{(H)}$
	Sampling	rbias	-.00	.06	.01	-.11	-.19	-.13	-.31	-.40	-.33
Population	rates	$\sqrt{\text{rmse}}$	.08	.37	.11	.13	.24	.14	.31	.41	.33
I	$f_1 = 0.5$	mdre	-.01	-.01	.01	-.11	-.21	-.13	-.31	-.40	-.33
	$f_2 = 0.4$	mdare	.05	.16	.06	.11	.21	.13	.31	.40	.33
	Sampling	rbias	-.09	-.16	-.10	-.15	-.28	-.18	-.25	-.35	-.27
Population	rates	$\sqrt{\text{rmse}}$	.13	1.9	.48	.18	.35	.20	.27	.36	.29
II	$f_1 = 0.5$	mdre	-.08	-.27	-.12	-.15	-.32	-.18	-.26	-.37	-.28
	$f_2 = 0.4$	mdare	.09	.28	.12	.15	.32	.18	.26	.37	.28
Results on standard deviation estimators			UMLEs			HTLEs			Homogeneous MLEs		
			$\hat{\tau}_1^{(U)}$	$\hat{\tau}_2^{(U)}$	$\hat{\tau}^{(U)}$	$\hat{\tau}_{HT.1}^{(U)}$	$\hat{\tau}_{HT.2}^{(U)}$	$\hat{\tau}_{HT}^{(U)}$	$\hat{\tau}_{ML.1}^{(H)}$	$\hat{\tau}_{ML.2}^{(H)}$	$\hat{\tau}_{ML}^{(H)}$
	Sampling	rbias	.15	2.0	1.8	.15	.85	.55	-.42	-.20	-.41
Population	rates	$\sqrt{\text{rmse}}$	.29	6.7	5.6	.29	1.6	.90	.42	.26	.42
I	$f_1 = 0.5$	mdre	.12	.22	.35	.11	.42	.36	-.42	-.22	-.42
	$f_2 = 0.4$	mdare	.16	.58	.39	.15	.48	.36	.42	.23	.42
	Sampling	rbias	-.19	.26	.23	-.30	-.11	-.18	-.74	-.48	-.69
Population	rates	$\sqrt{\text{rmse}}$	.27	10.5	7.3	.35	1.9	.94	.75	.66	.71
II	$f_1 = 0.5$	mdre	-.20	-.59	-.30	-.31	-.48	-.32	-.75	-.59	-.71
	$f_2 = 0.4$	mdare	.21	.64	.33	.31	.55	.35	.75	.60	.71

Table 3. Continued.

Results on 95% confidence intervals			UMLEs			HTLEs			Homogeneous MLEs		
			$\hat{\tau}_1^{(U)}$	$\hat{\tau}_2^{(U)}$	$\hat{\tau}^{(U)}$	$\hat{\tau}_{HT.1}^{(U)}$	$\hat{\tau}_{HT.2}^{(U)}$	$\hat{\tau}_{HT}^{(U)}$	$\hat{\tau}_{ML.1}^{(H)}$	$\hat{\tau}_{ML.2}^{(H)}$	$\hat{\tau}_{ML}^{(H)}$
	Sampling rates:	cp	.95	.97	.98	.78	.85	.78	.00	.00	.00
Population	$f_1 = 0.5$	mrl	.37	5.6	1.4	.31	1.2	.39	.08	.17	.07
I	$f_2 = 0.4$	mdrl	.36	1.8	.52	.30	.85	.33	.08	.16	.07
	Sampling rates:	cp	.76	.51	.63	.44	.36	.28	.05	.09	.02
Population	$f_1 = 0.5$	mrl	.29	1.7	.54	.24	.66	.27	.10	.24	.10
II	$f_2 = 0.4$	mdrl	.29	.44	.26	.24	.34	.21	.10	.19	.09

Notes: Results on point estimators are based on 5,000 samples. In Population I the percentages of samples in which the estimators of τ_1, τ_2 and τ derived under the heterogeneity assumption that were not computed because of numerical convergence problems were 0%, 0.02%, and 0.02%, respectively, whereas in Population II the corresponding percentages were 0.36%, 8.7%, and 9.0%. Bootstrap standard deviation estimators corresponding to estimators of sizes derived under the heterogeneity assumption were computed using 50 bootstrap samples, and their results, as well as those of the 95% confidence intervals are based on 500 replicated samples. In Population II the percentages of samples in which the bootstrap standard deviation estimators and confidence intervals were not computed because of convergence problems were 0.8%, 8.2% and 8.8% in U_1, U_2 and U, respectively, whereas in Population I the respective percentages were all 0%. Results on point estimators, standard deviation estimators and confidence intervals derived under the homogeneity assumption are based on 5,000 samples and no convergence problems occurred.

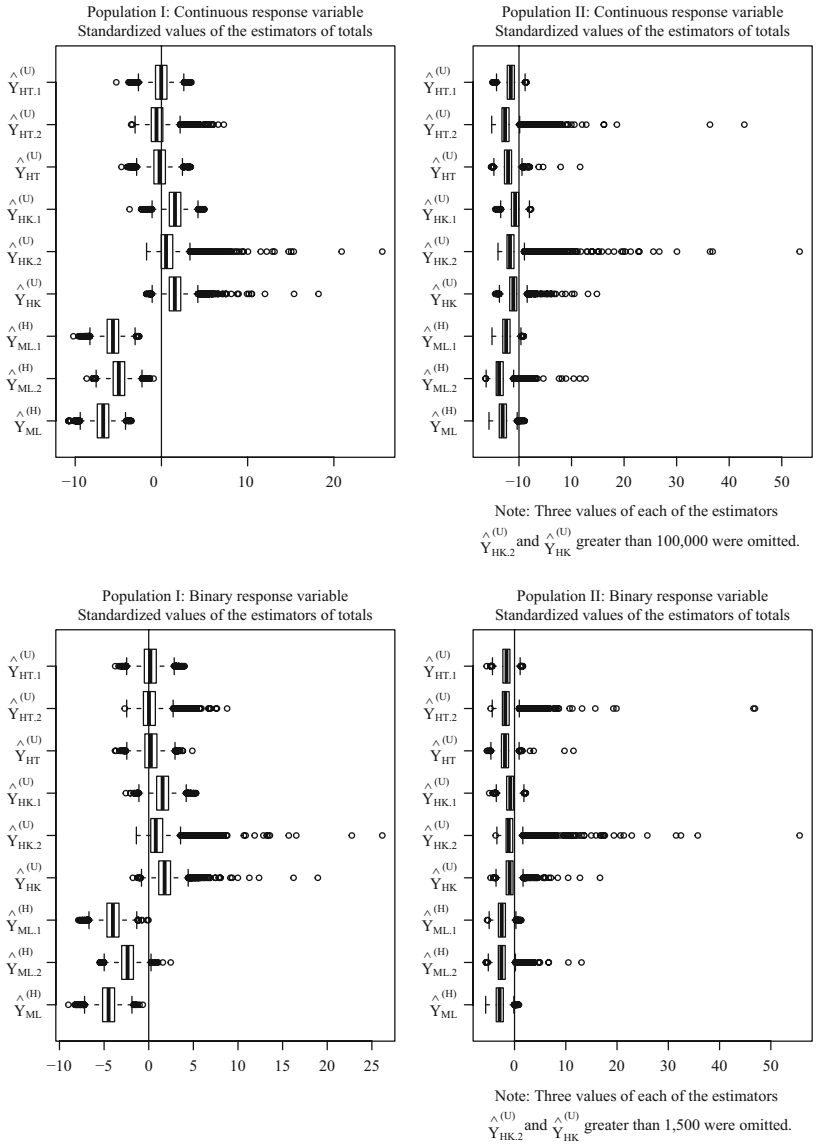

Fig. 2. Boxplots of standardized values of the estimators of totals in Populations I and II.

those of $\hat{Y}^{(U)}_{HK.1}$ which were more or less symmetrical. The distributions presented this shape regardless of the population and type of variable. These estimators showed moderate positive biases in Population I, and moderate negative biases in Population II. Note that in Population II the magnitudes of the biases of these estimators were smaller than those of the HTLEs. The HTLEs of the totals derived under the homogeneity assumption exhibited serious underestimation problems in both populations and with both types of response variables.

In the case of the estimators of the population means (see Figure 3 and Table 4), we have that the HTLEs exhibited more or less symmetric distributions, except for $\hat{\bar{Y}}^{(U)}_{HT}$ whose

Table 4. Relative biases, square roots of relative mean square errors, medians of relative errors and medians of absolute relative errors of the point estimators of the totals and means.

Population		I								II							
Sampling rates		$f_1 = 0.5$				$f_2 = 0.4$				$f_1 = 0.5$				$f_2 = 0.4$			
Response variable		Continuous				Binary				Continuous				Binary			
Estimator		rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare
HTLEs of totals	$\hat{Y}^{(U)}_{HT.1}$	-.00	.06	-.00	.04	.01	.07	.01	.05	-.14	.17	-.14	.14	-.16	.19	-.16	.16
	$\hat{Y}^{(U)}_{HT.2}$	-.06	.17	-.08	.12	.03	.19	.01	.11	-.27	.34	-.30	.31	-.21	.32	-.25	.26
	$\hat{Y}^{(U)}_{HT}$	-.01	.06	-.01	.04	.02	.07	.02	.05	-.17	.18	-.17	.17	-.16	.19	-.17	.17
HKLEs of totals	$\hat{Y}^{(U)}_{HK.1}$	.11	.14	.11	.11	.13	.16	.13	.13	-.07	.12	-.07	.08	-.09	.14	-.09	.10
	$\hat{Y}^{(U)}_{HK.2}$	.24	.47	.15	.18	.35	.57	.25	.25	-.15	1.9	-.26	.27	-.06	2.7	-.20	.22
	$\hat{Y}^{(U)}_{HK}$	.15	.18	.14	.14	.18	.21	.17	.17	-.08	.45	-.09	.10	-.08	.49	-.09	.10
Homo. HTLEs totals	$\hat{Y}^{(H)}_{ML.1}$	-.19	.19	-.19	.19	-.17	.17	-.17	.17	-.24	.25	-.24	.24	-.26	.28	-.26	.26
	$\hat{Y}^{(H)}_{ML.2}$	-.29	.29	-.29	.29	-.21	.23	-.21	.21	-.33	.35	-.35	35	-.27	.31	-.29	.30
	$\hat{Y}^{(H)}_{ML}$	-.21	.21	-.21	.21	-.18	.18	-.18	.18	-.26	.27	-.26	.26	-.26	.27	-.26	.26
HTLEs of means	$\hat{\bar{Y}}^{(U)}_{HT.1}$	.00	.03	.00	.02	.02	.05	.02	.03	-.06	.07	-.07	.07	-.08	.09	-.08	.08
	$\hat{\bar{Y}}^{(U)}_{HT.2}$	-.08	.14	-.07	.08	.01	.16	.01	.10	-.06	.10	-.05	.05	.02	.14	.03	.09
	$\hat{\bar{Y}}^{(U)}_{HT}$	-.02	.08	-.01	.03	.01	.09	.01	.04	-.06	.11	-.06	.06	-.06	.15	-.06	.06

Table 4. Continued.

Population		I								II							
Sampling rates		$f_1 = 0.5$				$f_2 = 0.4$				$f_1 = 0.5$				$f_2 = 0.4$			
Response variable		Continuous				Binary				Continuous				Binary			
Estimator		rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare
HKLEs of means	$\hat{\bar{Y}}^{(U)}_{HK.1}$	.12	.12	.12	.12	.14	.14	.14	.14	.02	.02	.02	.02	-.01	.04	-.01	.03
	$\hat{\bar{Y}}^{(U)}_{HK.2}$	.17	.17	.17	.17	.27	.30	.27	.27	.02	.03	.02	.02	.11	.16	.11	.11
	$\hat{\bar{Y}}^{(U)}_{HK}$	.13	.13	.13	.13	.17	.17	.17	.17	.02	.03	.02	.02	.03	.05	.03	.03
Homo. HTLEs means	$\hat{\bar{Y}}^{(H)}_{ML.1}$	.17	.17	.17	.17	.20	.21	.20	.20	.02	.02	.02	.02	-.01	.04	-.01	.03
	$\hat{\bar{Y}}^{(H)}_{ML.2}$	.20	.20	.20	.20	.33	.35	.32	.32	.02	.03	.02	.02	.12	.16	.12	.12
	$\hat{\bar{Y}}^{(H)}_{ML}$	.18	.12	.18	.18	.24	.14	.23	.23	.03	.03	.03	.03	.02	.08	.02	.03

Notes: Results are based on 5,000 samples. In Population I the percentages of samples in which the estimators of the totals and means of U_1, U_2 and U derived under the heterogeneity assumption that were not computed because of numerical convergence problems were 0%, 0.02% and 0.02%, respectively, whereas in Population II the corresponding percentages were 0.36%, 8.7% and 9.0%. In the computation of the estimators derived under the homogeneity assumption no convergence problems occurred.

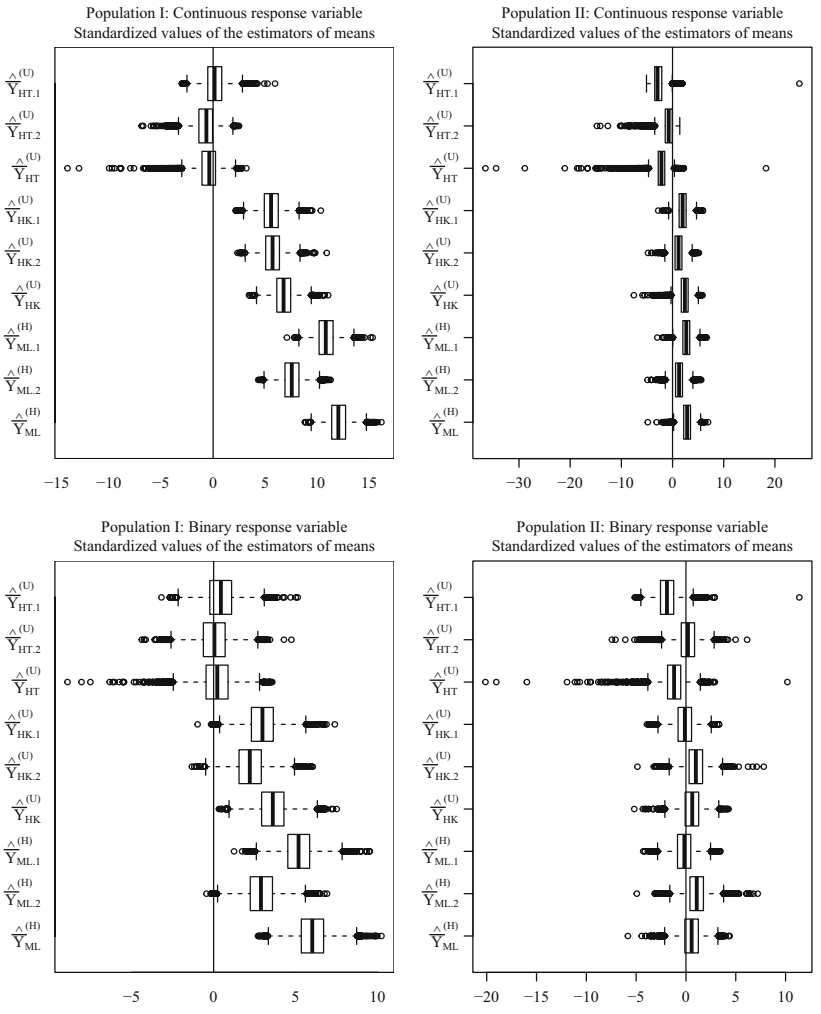

Fig. 3. Boxplots of standardized values of the estimators of means in Populations I and II.

distributions were skewed to the left. These estimators did not present problems of bias, regardless of the type of response variable and population. The HKLEs showed symmetrical distributions. In the case of Population I, they exhibited moderate positive biases, whereas in Population II, they were practically unbiased. Finally, the HTLEs derived under the homogeneity assumption performed very similarly to the HKLEs, but the magnitudes of their biases were slightly larger than those of the HKLEs.

Thus, in summary, the HTLEs of totals and means derived under the heterogeneity assumption had the best performance in Population I, and their performance was good, whereas in Population II, the HKLEs had the best performance and their performance was acceptable. These results are the consequence of the high correlations between the response variables and the inclusion probabilities in Population I, which favors the HTLEs, and the small correlations in Population II, which favors the HKLEs. (See Thompson 2012, subsec. 6.2; Särndal et al. 1992, subsec. 5.7). The HTLEs derived under

the homogeneity assumption performed very badly, except in the case of estimating the means of the Population II, where their performance was fairly good. The small biases of these estimators could be explained by the fact that the biases of this type of estimator when totals were estimated were practically the same as the biases of this type of estimator when sizes were estimated, and consequently, their biases were canceled out when the quotients were computed to form the estimators of the means.

The results of the estimators of the standard deviations of the estimators of the totals and means are shown in Table 5. In general, each one of the estimators exhibited problems of bias. The estimators of the standard deviations of the HKLEs had the best performance. Estimators of the standard deviations of the estimators of means performed better than those of the estimators of totals. In addition, their performance was better in the case of the binary response variable than in the case of the continuous response variable. Furthermore, their performance was better in Population I than in Population II.

The results on the confidence intervals for the totals and means are shown in Table 6. In Population I, the confidence intervals for the totals based on the HTLEs had good values of the coverage probabilities and relative lengths regardless of the type of response variable, except for the interval for Y_2, which had large relative lengths. In Population II, the coverage probabilities of these intervals were low. This result is a consequence of the moderate biases exhibited by the HTLEs of the totals. The intervals for the totals based on the HKLEs had, in general, relatively low coverage probabilities regardless of the type of response variable and population, except for the interval for Y_2, which showed large coverage probabilities, as well as large lengths, in Population I. The intervals for the totals based on the HTLES derived under the homogeneity assumption showed very low coverage probabilities, which was a consequence of the large biases exhibited by these estimators. In the case of the confidence intervals for the means, we have that the intervals based on the HTLEs derived under the heterogeneity assumption had good performance in Population I. In Population II, the coverage probabilities of the intervals for $\bar{Y}_1$ and $\bar{Y}$ exhibited low coverage probabilities. The intervals based on the HKLEs had very low coverage probabilities in Population I. In Population II, and particularly in the case of the binary response variable, these intervals had good performance. Finally, the intervals based on the HTLEs derived under the homogeneity assumption showed very low coverage probabilities in each of the situations that were considered in this study, except in Population II and binary response variable, where the intervals performed acceptably.

Thus, in summary, the intervals for the totals and means based on the HTLEs derived under the heterogeneity assumption had the best performance, but their performance was good only in the case of Population I (although the intervals for Y_2 were very long). The intervals based on the HKLEs and on the HTLEs derived under the homogeneity assumption did not perform well, except in the case of Population II and binary response variable, where their performance was acceptable.

6.2. *Populations Constructed Using Data from the National Longitudinal Study of Adolescent Health*

In this Monte Carlo study, we used data from the National Longitudinal Study of Adolescent Health (Add Health) to construct a population. The Add Health is a longitudinal study of a

Table 5. Relative biases, square roots of relative mean square errors, medians of relative errors and medians of absolute relative errors of the standard deviation estimators of the estimators of the totals and means.

Population		I								II							
Sampling rates		$f_1 = 0.5$				$f_2 = 0.4$				$f_1 = 0.5$				$f_2 = 0.4$			
Response variable		Continuous				Binary				Continuous				Binary			
Standard deviation estimator		rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare
HTLEs of totals	$\widehat{sd}_B(\hat{Y}^{(U)}_{HT.1})$	.20	.32	.16	.19	.12	.26	.10	.15	-.29	.35	-.30	.30	-.23	.30	-.24	.25
	$\widehat{sd}_B(\hat{Y}^{(U)}_{HT.2})$	.78	1.5	.33	.45	.63	1.3	.26	.41	-.09	1.9	-.47	.54	.02	1.6	-.35	.44
	$\widehat{sd}_B(\hat{Y}^{(U)}_{HT})$	.57	.89	.40	.40	.46	.72	.34	.34	-.18	.82	-.31	.34	-.13	.59	-.22	.26
HKLEs of totals	$\widehat{sd}_B(\hat{Y}^{(U)}_{HK.1})$	.20	.32	.16	.19	.12	.26	.09	.15	-.18	.27	-.19	.21	-.16	.26	-.16	.19
	$\widehat{sd}_B(\hat{Y}^{(U)}_{HK.2})$	1.9	6.3	.14	.56	1.6	5.9	.09	.52	.26	10.3	-.59	.63	.32	9.6	-.50	.57
	$\widehat{sd}_B(\hat{Y}^{(U)}_{HK})$	1.9	5.6	.34	.37	1.7	5.2	.33	.37	.22	6.7	-.27	.30	.18	5.2	-.21	.26
Homo. HTLEs totals	$\widehat{sd}_B(\hat{Y}^{(H)}_{ML.1})$	-.24	.25	-.24	.24	-.18	.19	-.18	.18	-.69	.71	-.72	.72	-.62	.64	-.65	.65
	$\widehat{sd}_B(\hat{Y}^{(H)}_{ML.2})$	-.11	.21	-.14	.17	-.08	.19	-.10	.14	-.47	.65	-.58	.59	-.31	.54	-.42	.45
	$\widehat{sd}_B(\hat{Y}^{(H)}_{ML})$	-.25	.27	-.26	.26	-.17	.19	-.18	.18	-.66	.68	-.69	.69	-.58	.60	-.61	.61
HTLEs of means	$\widehat{sd}_B(\hat{\bar{Y}}^{(U)}_{HT.1})$	.01	.18	-.01	.11	.16	.23	.15	.16	-.50	.54	-.51	.51	-.11	.27	-.13	.17
	$\widehat{sd}_B(\hat{\bar{Y}}^{(U)}_{HT.2})$	.35	.65	.22	.35	.24	.44	.17	.21	-.20	.67	-.38	.44	.06	.50	-.06	.24
	$\widehat{sd}_B(\hat{\bar{Y}}^{(U)}_{HT})$	.84	1.8	.20	.43	.65	1.3	.19	.28	-.35	.73	-.50	.52	-.02	.60	-.16	.22

Table 5. Continued

Population		I								II							
Sampling rates		$f_1 = 0.5$				$f_2 = 0.4$				$f_1 = 0.5$				$f_2 = 0.4$			
Response variable		Continuous				Binary				Continuous				Binary			
Standard deviation estimator		rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare
HTLEs of means	$\widehat{sd}_B(\hat{\bar{Y}}^{(U)}_{HK,1})$	-.12	.18	-.13	.14	.11	.18	.10	.12	-.10	.20	-.12	.15	.05	.23	.04	.15
	$\widehat{sd}_B(\hat{\bar{Y}}^{(U)}_{HK,2})$	.09	.23	.06	.15	.11	.25	.09	.15	.07	.35	.00	.20	.03	.33	-.03	.20
	$\widehat{sd}_B(\hat{\bar{Y}}^{(U)}_{HK})$	.06	.30	-.01	.14	.16	.29	.13	.14	-.13	.47	-.24	.29	.08	.32	.06	.16
Homo. HTLEs means	$\widehat{sd}_B(\hat{\bar{Y}}^{(H)}_{ML,1})$	-.38	.39	-.39	.39	-.16	.18	-.16	.16	-.21	.29	-.23	.24	-.09	.25	-.11	.18
	$\widehat{sd}_B(\hat{\bar{Y}}^{(H)}_{ML,2})$	-.19	.22	-.20	.20	-.11	.15	-.12	.13	-.12	.29	-.16	.22	-.12	.29	-.15	.22
	$\widehat{sd}_B(\hat{\bar{Y}}^{(H)}_{ML})$	-.37	.38	-.38	.38	-.16	.18	-.17	.17	-.26	.36	-.30	.31	-.13	.25	-.15	.19

Notes: Bootstrap standard deviation estimators sd_B were computed using 50 bootstrap samples and their results are based on 500 replicated samples. In Population II the percentages of replicated samples in which the estimators sd_B were not computed because of convergence problems were 0.8%, 8.2% and 8.8% in U_1, U_2 and U, respectively, whereas in Population I the respective percentages were all 0%. Results on standard deviation estimators derived under the homogeneity assumption are based on 5000 replicated samples and no convergence problems occurred.

Table 6. Coverage probabilities and means and medians of the relative lengths of the 95% confidence intervals for the totals and means.

Population		I						II					
Sampling rates		$f_1 = 0.5$			$f_2 = 0.4$			$f_1 = 0.5$			$f_2 = 0.4$		
Response variable		Continuous			Binary			Continuous			Binary		
95% CI		cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl
HTLEs of totals	$CI(\hat{Y}^{(U)}_{HT.1})$	.96	.29	.28	.96	.32	.32	.42	.24	.23	.53	.29	.28
	$CI(\hat{Y}^{(U)}_{HT.2})$	.90	1.1	.81	.98	1.4	1.0	.28	.56	.33	.63	.82	.49
	$CI(\hat{Y}^{(U)}_{HT})$	.96	.35	.31	.98	.38	.35	.27	.25	.21	.47	.30	.26
HKLEs of totals	$CI(\hat{Y}^{(U)}_{HK.1})$	.80	.34	.33	.66	.38	.37	.74	.28	.28	.78	.33	.33
	$CI(\hat{Y}^{(U)}_{HK.2})$	.98	4.1	1.6	.99	6.4	2.0	.45	1.3	.43	.75	1.9	.60
	$CI(\hat{Y}^{(U)}_{HK})$	.94	1.1	.50	.79	1.3	.54	.64	.44	.26	.75	.53	.31
Homo. HTLEs totals	$CI(\hat{Y}^{(H)}_{ML.1})$	.00	.10	.10	.01	.14	.14	.09	.12	.11	.10	.15	.14
	$CI(\hat{Y}^{(H)}_{ML.2})$	.00	.21	.20	.32	.32	.32	.10	.25	.20	.26	.38	.32
	$CI(\hat{Y}^{(H)}_{ML})$	.00	.09	.09	.00	.13	.13	.03	.11	.10	.06	.15	.14
HTLES of means	$CI(\hat{\bar{Y}}^{(U)}_{HT.1})$	.95	.10	.09	.96	.19	.19	.19	.08	.07	.55	.18	.18
	$CI(\hat{\bar{Y}}^{(U)}_{HT.2})$	.98	.57	.52	.96	.79	.75	.91	.23	.18	.95	.58	.51
	$CI(\hat{\bar{Y}}^{(U)}_{HT})$	.98	.33	.21	.96	.41	.29	.38	.11	.08	.84	.22	.19

Table 6. Continued.

Population		I						II					
Sampling rates		$f_1 = 0.5$			$f_2 = 0.4$			$f_1 = 0.5$			$f_2 = 0.4$		
Response variable		Continuous			Binary			Continuous			Binary		
95% CI		cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl
HKLEs of means	$\mathrm{CI}(\hat{\hat{Y}}^{(U)}_{HK.1})$	.00	.07	.07	.17	.19	.19	.48	.03	.03	.96	.17	.17
	$\mathrm{CI}(\hat{\hat{Y}}^{(U)}_{HK.2})$	.00	.13	.12	.48	.55	.54	.77	.07	.07	.84	.47	.44
	$\mathrm{CI}(\hat{\hat{Y}}^{(U)}_{HK})$	.00	.08	.08	.08	.21	.20	.30	.04	.04	.92	.18	.17
Homo. HTLEs means	$\mathrm{CI}(\hat{\hat{Y}}^{(H)}_{ML.1})$	.00	.00	.04	.00	.13	.13	.17	.03	.03	.91	.14	.13
	$\mathrm{CI}(\hat{\hat{Y}}^{(U)}_{ML.2})$	.00	.08	.08	.11	.39	.39	.66	.06	.06	.75	.40	.38
	$\mathrm{CI}(\hat{\hat{Y}}^{(H)}_{ML})$	.00	.04	.04	.00	.13	.13	.13	.03	.03	.86	.14	.13

Notes: Results on confidence intervals derived under the assumption of heterogeneous link probabilities are based on 500 samples. In Population II the percentages of replicated samples in which the CIs were not computed because of convergence problems were 0.8%, 8.2% and 8.8% in U_1, U_2 and U, respectively, whereas in Population I the respective percentages were all 0%. Results on CIs derived under the homogeneity assumption are based on 5,000 replicated samples and no convergence problems occurred.

representative sample of more than 90,000 adolescents who in the years 1994–1995 were in grades 7–12 in the United States. The participants were followed through adolescence and the transition to adulthood with the goal of helping to explain the causes of adolescent health and health behavior. The sample of students was selected by a stratified probability proportional to size cluster sampling design, where the clusters were the high schools and the strata were defined in terms of region, urbanicity, school type and so on. For each of the 84 selected high schools, one of its feeder middle schools was selected with probability proportional to the number of contributed students to the high school.

Each student in the representative sample was asked to named up to five male and five female friends within his or her high school or in the feeder school, and in addition, to complete an in-school questionnaire. Thus, the collected information can be modeled as a directed network, where the nodes are the sampled students and their referred friends, and a directed arc from node i to node j is considered to exist if student i names student j as a friend. See Harris (2013) for a description of this study.

A subset of the data obtained in the Add Health study is contained in Linton Freeman's web page: See Freeman. In our numerical study we used data from this subset corresponding to the high school and its feeder in Community 50 to construct a population U of $\tau = 2{,}497$ elements divided into subpopulations U_1 and U_2 of sizes $\tau_1 = 1{,}800$ and $\tau_2 = 697$, respectively. The elements assigned to U_1 were those at positions labeled with odd numbers in the data file and that named at least one friend plus a simple random sample of elements at positions labeled with even numbers and that named at least one friend. These elements were grouped into $N = 150$ clusters of sizes m_i, $i = 1, \ldots, N$, obtained by sampling from a negative binomial distribution with mean and variance equal to 12 and 24, respectively. The elements assigned to U_2 were the remaining elements in the data file that were named as a friend by at least one element in U_1. Once the subpopulation U_k was constructed, the $N \times \tau_k$ matrix X_k of values $x_{ij}^{(k)}$s of the link indicator variables $X_{ij}^{(k)}$s was constructed, $k = 1, 2$. We considered as response variables the following: "Number of friends" (named by each element) and "Sex" (1 = male, 0 = female). The totals and means of the variable "Number of friends" were $(Y_1, Y_2, Y) = (10{,}101, 2{,}729, 12{,}830)$ and $(\bar{Y}_1, \bar{Y}_2, \bar{Y}) = (5.612, 3.915, 5.138)$, and those of the variable "Sex" were $(Y_1, Y_2, Y) = (838, 361, 1{,}200)$ and $(\bar{Y}_1, \bar{Y}_2, \bar{Y}) = (0.466, 0.518, 0.481)$. For an initial sample size $n = 20$, which was the size used in this study, the values of the Pearson correlation coefficients between the values of the variable "Number of friends" and those of the inclusion probabilities associated with subpopulations U_1 and U_2 were $\rho(y^{(1)}, \pi_{(1)}) \approx 0.36$ and $\rho(y^{(2)}, \pi_{(2)}) \approx 0.29$, respectively, whereas the corresponding values for the variable "Sex" were $\rho(y^{(1)}, \pi_{(1)}) \approx -0.04$ and $\rho(y^{(2)}, \pi_{(2)}) \approx -0.07$.

The Monte Carlo study was carried out as in the previous study, except that the size of the initial sample of clusters was $n = 20$ and that for each selected value m_i, the values $x_{ij}^{(k)}$s of the link indicator variables $X_{ij}^{(k)}, j = 1, \ldots, \tau_k$, were obtained from the matrix X_k, $k = 1, 2$. Furthermore, for each element $j \in U_k$ that was sampled, the values $y_j^{(k)}$s of both response variables were recorded.

The results about the inferences on the population sizes are shown in Figure 4 and in Table 7. We can see that the UMLES performed well: they did not have problems of bias or instability. The estimators of their standard deviations presented problems of bias, but the confidence intervals based on these estimators performed acceptably. The HTLEs also

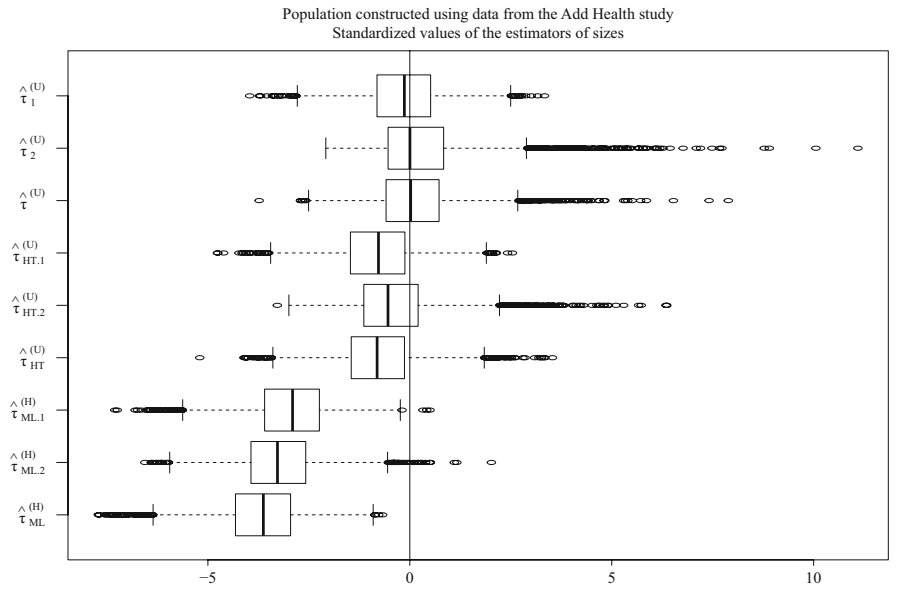

Fig. 4. Boxplots of standardized values of the estimators of sizes in the population constructed using data from the National Longitudinal Study of Adolescent Health.

had acceptable performance, although the magnitudes of their biases were slightly greater than those of the UMLEs, and these biases affected the coverage probabilities of the intervals based on these estimators. Finally, the MLEs derived under the homogeneity assumption had large biases that affected their performance and that of the confidence intervals based on these estimators. Thus, in summary, the UMLEs had the best performance, which was good; the HTLEs performed acceptably, whereas the MLEs derived under the homogeneity assumption performed poorly.

The results of the point estimators of the totals and means and of the estimators of their standard deviations are shown in Table 8 and Figure 5. We can see that the HTLEs performed well regardless of the response variable: they did not present problems of bias or instability. In the case of the variable "Sex" the magnitudes of their biases were slightly greater than in the case of the variable "Number of friends". The HKLEs also performed well, although slightly less well than the HTLEs, except in the estimation of the mean of the variable "Sex", where the HKLEs performed better than the HTLEs. The HTLEs of the totals derived under the homogeneity assumption presented problems of underestimation, although in the case of the variable "Number of friends" the magnitudes of their biases were not very great. These estimators performed well in the estimation of the mean, especially of the variable "Sex", where these estimators, along with the HKLEs, had the best performance. The estimators of the standard deviations of the point estimators considered in this study presented problems of biases, with some exceptions. However, the magnitudes of their biases were not so great.

The results of the confidence intervals for the totals and means are shown in Table 9. We can see that the confidence intervals for the totals of the variable "Number of friends" based on the HTLEs performed well. In the case of the variable "Sex" the coverage probabilities of these intervals were somewhat low. The intervals for the totals based on the HKLEs had

Table 7. Simulation results obtained for the point estimators, standard deviation estimators and 95% confidence intervals for the sizes in a population constructed using data from the national longitudinal study of adolescent health.

		Point estimators				Standard deviation estimators				Confidence intervals		
		rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	cp	mrl	mdrl
UMLEs of sizes	$\hat{\tau}_1^{(U)}$	-.01	.06	-.01	.04	-.15	.23	-.17	.19	.89	.22	.21
	$\hat{\tau}_2^{(U)}$	.06	.25	.00	.13	.08	1.3	-.29	.43	.95	1.1	.68
	$\hat{\tau}^{(U)}$	.01	.08	.00	.05	.05	.99	-.24	.33	.93	.36	.26
HTLEs of sizes	$\hat{\tau}_{HT.1}^{(U)}$	-.04	.07	-.04	.05	-.15	.23	-.16	.18	.82	.20	.19
	$\hat{\tau}_{HT.2}^{(U)}$	-.05	.15	-.07	.10	.13	.81	-.10	.31	.90	.67	.52
	$\hat{\tau}_{HT}^{(U)}$	-.05	.08	-.05	.05	-.04	.47	-.17	.25	.82	.24	.21
Homo. MLEs of sizes	$\hat{\tau}_{ML.1}^{(H)}$	-.14	.15	-.14	.14	-.42	.43	-.42	.42	.04	.11	.11
	$\hat{\tau}_{ML.2}^{(H)}$	-.24	.25	-.24	.24	-.24	.29	-.26	.26	.10	.22	.22
	$\hat{\tau}_{ML}^{(H)}$	-.17	.18	-.17	.17	-.45	.45	-.45	.45	.01	.10	.10

Notes: Results for point estimators are based on 5,000 samples; those for standard deviation estimators and confidence intervals derived under the heterogeneity assumption are based on 500 samples and those under the homogeneity assumption are based on 5,000 samples. Average sampling fractions are $f_1 = 0.46$ and $f_2 = 0.40$. No convergence problems were observed, except in one sample in which the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_{HT.2}^{(U)}$ could not be computed.

Table 8. Simulation results obtained for the point and standard deviation estimators of totals and means in a population constructed using data from the National Longitudinal Study of Adolescent Health.

Response variable		Point estimators								Standard deviation estimators							
		Number of friends				Sex				Number of friends				Sex			
Estimator		rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare
HTLEs of totals	$\hat{Y}_{HT.1}^{(U)}$	.00	.06	.01	.04	-.07	.09	-.07	.07	-.14	.22	-.16	.17	-.15	.22	-.16	.17
	$\hat{Y}_{HT.2}^{(U)}$	.07	.17	.05	.09	-.06	.16	-.08	.11	.12	.75	-.10	.31	.19	.81	-.04	.29
	$\hat{Y}_{HT}^{(U)}$	.02	.06	.02	.04	-.07	.09	-.07	.07	-.07	.36	-.16	.21	-.02	.45	-.13	.23
HKLEs of totals	$\hat{Y}_{HK.1}^{(U)}$	.04	.08	.04	.05	-.03	.08	-.03	.05	-.15	.23	-.17	.18	-.16	.23	-.18	.19
	$\hat{Y}_{HK.2}^{(U)}$	.19	.33	.13	.15	.04	.24	-.01	.13	.05	1.2	-.30	.43	.12	1.3	-.26	.40
	$\hat{Y}_{HK}^{(U)}$	.08	.11	.07	.07	-.01	.09	-.02	.06	.03	.89	-.22	.30	.06	.96	-.21	.31
Homo. HTLEs totals	$\hat{Y}_{ML.1}^{(H)}$	-.09	.11	-.09	.09	-.16	.17	-.16	.16	-.20	.21	-.19	.19	-.22	.23	-.22	.22
	$\hat{Y}_{ML.2}^{(H)}$	-.13	.16	-.14	.14	-.25	.26	-.25	.25	-.17	.23	-.19	.20	-.19	.25	-.21	.22
	$\hat{Y}_{ML}^{(H)}$	-.10	.11	-.10	.10	-.19	.20	-.19	.19	-.28	.28	-.27	.27	-.31	.31	-.31	.31
HTLEs of means	$\hat{\bar{Y}}_{HT.1}^{(U)}$	.01	.02	.01	.02	-.06	.07	-.06	.06	-.15	.19	-.15	.16	-.03	.13	-.03	.09
	$\hat{\bar{Y}}_{HT.2}^{(U)}$	.03	.10	.04	.08	-.10	.13	-.09	.09	-.14	.42	-.27	.32	.03	.42	-.10	.24
	$\hat{\bar{Y}}_{HT}^{(U)}$	.01	.08	.02	.03	-.07	.06	-.07	.07	-.09	.70	-.33	.40	.10	.67	-.09	.20

Table 8. Continued.

Response variable		Point estimators								Standard deviation estimators							
		Number of friends				Sex				Number of friends				Sex			
Estimator		rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare	rbias	$\sqrt{\text{rmse}}$	mdre	mdare
HKLEs of means	$\hat{\bar{Y}}^{(U)}_{HK.1}$	.05	.05	.05	.05	-.02	.04	-.02	.03	-.12	.16	-.13	.13	-.06	.14	-.06	.09
	$\hat{\bar{Y}}^{(U)}_{HK.2}$	.13	.13	.13	.13	-.02	.04	-.02	.03	.01	.16	.00	.10	.31	.37	.30	.30
	$\hat{\bar{Y}}^{(U)}_{HK}$	.07	.07	.07	.07	-.02	.03	-.02	.02	-.09	.29	-.15	.20	.01	.16	-.00	.10
Homo. HTLEs means	$\hat{\bar{Y}}^{(H)}_{ML.1}$	.06	.06	.06	.06	-.02	.04	-.02	.03	-.18	.19	-.18	.18	-.13	.14	-.13	.13
	$\hat{\bar{Y}}^{(H)}_{ML.2}$	.14	.14	.14	.14	-.02	.04	-.02	.03	-.10	.12	-.11	.11	.12	.14	.12	.12
	$\hat{\bar{Y}}^{(H)}_{ML}$	.08	.02	.08	.08	-.02	.03	-.02	.03	-.25	.26	-.26	.26	-.13	.14	-.13	.13

Notes: Results for point estimators are based on 5,000 samples; those for standard deviations estimators derived under the heterogeneity assumption are based on 500 samples and those the homogeneity assumption are based on 5,000 samples. Average sampling fractions are $f_1 = 0.46$ and $f_2 = 0.40$. No convergence problems were observed, except in one sample in which the estimators of Y_2 and $\bar{Y}$ derived under the heterogeneity assumption could not be computed.

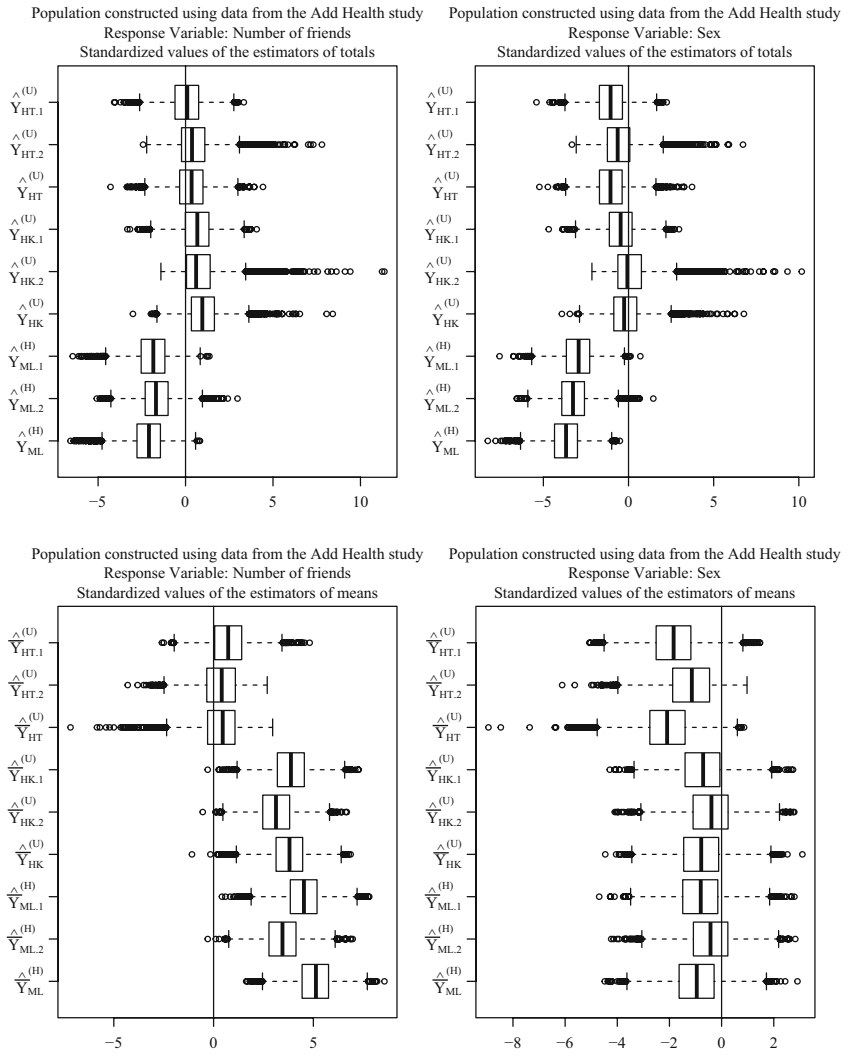

Fig. 5. Boxplots of standardized values of the estimators of totals and means in the population constructed using data from the National Longitudinal Study of Adolescent Health.

slightly low coverage probabilities, regardless of the response variable. In addition, the lengths of the intervals for Y_2 were very large. The intervals for the totals based on the HTLEs derived under the homogeneity assumption showed very low coverage probabilities, especially in the case of the variable "Sex", where the values of the probabilities were about 0.1. In the case of the intervals for the means, we have that the intervals based on the HTLEs derived under the heterogeneity assumption presented coverage probabilities that were somewhat low in the case of the variable "Number of friends", and clearly low in the case of the variable "Sex", except for the interval for $\bar{Y}_2$. These low probabilities were a consequence of the moderate biases exhibited by these estimators. The confidence intervals based on the HKLEs showed very low coverage probabilities in the case of the variable "Number of friends", which was also a consequence

Table 9. Simulation results obtained for the 95% confidence intervals in a population constructed using data from the National Longitudinal Study of Adolescent Health.

Response variable		CIs based on HTLEs of totals			CIs based on HKLEs of totals			CIs based on homogeneous HTLEs of totals		
		$\hat{Y}_{HT.1}^{(U)}$	$\hat{Y}_{HT.2}^{(U)}$	$\hat{Y}_{HT}^{(U)}$	$\hat{Y}_{HK.1}^{(U)}$	$\hat{Y}_{HK.2}^{(U)}$	$\hat{Y}_{HK}^{(U)}$	$\hat{Y}_{ML.1}^{(H)}$	$\hat{Y}_{ML.2}^{(H)}$	$\hat{Y}_{ML}^{(H)}$
Number of friends	cp	.90	.97	.94	.84	.98	.88	.42	.49	.28
	mrl	.20	.68	.23	.22	1.1	.33	.16	.28	.14
	mdrl	.20	.55	.21	.22	.71	.25	.16	.27	.14
Sex	cp	.73	.85	.74	.84	.92	.88	.10	.11	.02
	mrl	.22	.66	.26	.24	1.0	.37	.17	.25	.14
	mdrl	.22	.54	.23	.23	.68	.28	.17	.25	.14
		CIs based on HTLEs of means			CIs based on HKLEs of means			CIs based on homogeneous HTLEs of means		
		$\hat{\bar{Y}}_{HT.1}^{(U)}$	$\hat{\bar{Y}}_{HT.2}^{(U)}$	$\hat{\bar{Y}}_{HT}^{(U)}$	$\hat{\bar{Y}}_{HK.1}^{(U)}$	$\hat{\bar{Y}}_{HK.2}^{(U)}$	$\hat{\bar{Y}}_{HK}^{(U)}$	$\hat{\bar{Y}}_{ML.1}^{(H)}$	$\hat{\bar{Y}}_{ML.2}^{(H)}$	$\hat{\bar{Y}}_{ML}^{(H)}$
Number of friends	cp	.81	.77	.77	.02	.16	.09	.01	.05	.00
	mrl	.06	.32	.15	.05	.16	.07	.04	.14	.05
	mdrl	.06	.27	.11	.05	.16	.06	.04	.14	.05
Sex	cp	.53	.93	.53	.85	.99	.86	.81	.96	.77
	mrl	.12	.33	.15	.11	.19	.10	.10	.16	.09
	mdrl	.12	.29	.13	.11	.19	.10	.10	.16	.08

Notes: Results for confidence intervals derived under the heterogeneity assumption are based on 500 samples and those derived under the homogeneity assumption are based on 5,000 samples. Sampling fractions are $f_1 = 0.46$ and $f_2 = 0.40$. No convergence problems were observed.

of the biases exhibited by these estimators (see Figure 5). However, in the case of the variable "Sex" the coverage probabilities of these intervals were just somewhat low. The intervals based on the HTLEs derived under the homogeneity assumption performed very similarly to those based on the HKLEs, that is, they showed very low coverage probabilities in the case of the variable "Number of friends" and somewhat low coverage probabilities in the case of the variable "Sex", although in both cases the coverage probabilities were slightly lower than those of the intervals based on the HKLEs. Thus, in summary, the best intervals for the total and mean of the variable "Number of friends" were the ones based on the HTLEs. Their performance was good in the case of the total and acceptable in the case of the mean. On the other hand, the best intervals for the total and mean of the variable "Sex" were the ones based on the HKLEs, and their performance was acceptable.

7. Conclusions and Suggestions for Future Research

In this article, we have considered the link-tracing sampling variant proposed by Félix-Medina and Thompson (2004) and have proposed Horvitz-Thompson-like and Hájek-like estimators of population totals and means. This work extends that of Félix-Medina and Monjadin (2010) by assuming heterogeneous, rather than homogeneous, link probabilities which are modeled by a Rasch model used by Félix-Medina et al. (2015). The variances of the proposed estimators are estimated by a variant of bootstrap which extends the variant used by the previously cited authors by estimating the variances of estimators of totals and means, in addition to the variances of estimators of population sizes. This variant of bootstrap allows the estimation of variances when the response variable is either continuous, discrete or binary. Large sample confidence intervals for the population parameters are constructed by assuming that the estimators of the parameters have normal distributions, except for the estimators of the population sizes, which are assumed to have log-normal distributions. In addition, confidence intervals for proportions are constructed using Korn and Graubard's (1998) proposal.

We evaluated the performance of the proposed estimators by means of two Monte Carlo studies. In the first study, two populations were constructed using simulated data, and the results show that erroneous inferences might be obtained if some model assumptions were not satisfied. In particular, we found that if any of the assumptions are satisfied, then reliable inferences about population sizes, totals and means are obtained. Furthermore, we found that the assumption of the Poisson distribution of the sizes M_is of the venues A_is does not need to be satisfied to obtain reliable inferences. Nevertheless, we also found out that severe deviations from the Rasch model of the link probabilities lead to erroneous inferences, and that inferences about population sizes and totals are affected in greater extent than inferences about population means. In fact, inferences about the population means seem to be robust to deviations from the assumed models. In addition, we came upon that in any situation, the performance of the proposed bootstrap variance estimators is at most just good. However, in the second study, in which a finite population was constructed using data from the Add Health study, the results are promising. Thus, if this population were more or less representative of the populations that could be found in applications of this methodology, then reliable inferences would be expected to be obtained.

In the light of these results, we consider that the following issues are worthy of future research: (1) To develop a model for the link probabilities $p_{ij}^{(k)}$s that is robust to deviations from the assumed model. For instance, to model $p_{ij}^{(k)}$ as a quadratic function of $\alpha_i^{(k)}$ and $\beta_j^{(k)}$, or to assume that the $\beta_j^{(k)}$s are T-Student distributed instead of normally distributed, or to change the Rasch model by the latent classes model proposed by Pledger (2000) and which was used in the first Monte Carlo study to generate the values $x_{ij}^{(k)}$s in Population II. (2) To improve the proposed bootstrap variance estimators. For example, to predict the values of the response variable associated with the non-sampled elements by using a quadratic or a nonparametric regression model instead of a simple linear regression model. (3) To enhance the proposed CIs for the population sizes, totals and means. For instance, using the bootstrap percentile method which does not require assuming a probability distribution for the estimator of the parameter of interest.

8. Appendix: Bootstrap Procedure

In this section, we will describe the bootstrap variant that is proposed to construct a sample of values of an estimator $\hat{\theta}$ of a parameter θ (population size, total or mean), from which an estimate of the variance and/or the standard deviation of $\hat{\theta}$, as well as a confidence interval for θ can be computed. Thus, hereinafter, we will denote by $\lfloor x \rfloor$, the greatest integer less than or equal to $x \in \mathbb{R}$. The steps of the proposed bootstrap procedure are as follows. (1) Construct a population vector m_{Boot} of N values of m_is by means of the following procedure. If N/n is an integer, repeat N/n times the observed sample of n cluster sizes $\mathrm{m}_s = \{m_1, \ldots, m_n\}$. If N/n is not an integer, that is, if $N = an + b$, where a and b, $b < n$, are positive integers, then repeat a times m_s and add to this set a SRSWOR of b values of m_is selected from m_s. If the sum of the elements of the vector m_{Boot} is greater than the value $\hat{\tau}_1^{(a)}$, $a = U, C$, (depending of the type of estimator that is being considered), delete one element at a time from m_{Boot} starting from the N-th element until the sum is less than or equal to $\hat{\tau}_1^{(a)}$. Let N_{Boot} be the final number of elements in m_{Boot}. Note that this procedure for constructing the vector m_{Boot} avoids the assumption that the M_is be independent and identically distributed Poisson random variables. Therefore, the resulting bootstrap variance estimates and confidence intervals are robust to deviations from this assumption. In addition, the condition that the sum of the elements in m_{Boot} be less than or equal to $\hat{\tau}_1^{(a)}$ guarantees that every bootstrap initial sample $m'_1, \ldots, m'_n$, satisfies $\sum_1^n m'_i < \hat{\tau}_1^{(a)}$, and consequently, that no initial bootstrap sample contains the whole bootstrap population U_1^{Boot} of size $\hat{\tau}_1^{(a)}$. See step (6) of the procedure. (2) For each $k = 1, 2$, construct a population vector $\hat{\alpha}_{(k)Boot}^{(a)}$ of length N_{Boot} whose elements are the estimates $\hat{\alpha}_{(k)i}^{(a)}s$ of the $\alpha_{(k)i}$s associated with the clusters whose sizes m_is are in m_{Boot}. (3) For each $k = 1, 2$, construct a population vector $\hat{\beta}_{(k)Boot}^{(a)}$ of length $\hat{\tau}_k^{(a)}$ whose first $m + r_1$ elements in the case of $k = 1$, or whose first r_2 elements in the case of $k = 2$, are the estimates $\hat{\tilde{\beta}}_{(k)j}^{(a)}$s of the $\beta_{(k)j}$s associated with the people in S_k^*, and each one of the remaining elements is the estimate $\hat{\tilde{\beta}}_{(k)0}^{(a)}$ of $\beta_{(k)j}$ obtained using Equation (7) with $x_i = 0$, $i = 1, \ldots, n$. (4) For each $k = 1, 2$, construct a population vector $\hat{y}_{(k)Boot}^{(a)}$ of length $\hat{\tau}_{(k)}^{(a)}$ whose first $m + r_1$ elements in the case of $k = 1$, or whose first r_2 elements in the case of $k = 2$, are the y-values associated with the elements in S_k^*, and the remaining elements are estimates of the y-values associated with the people in $U_k - S_k^*$ and obtained using the following procedure. If the variable of interest y is continuous, then fit a simple linear

regression model to the data $(\hat{\pi}^{(a)}_{(k)j}(\hat{\boldsymbol{\alpha}}^{(a)}_k, \hat{\sigma}^{(a)}_k, \tilde{\tilde{\beta}}^{(a)}_{(k)j}), y^{(k)}_j)\ j \in S^*_k$. Next, predict the y-value associated with the j-th element in $U_k - S^*_k$ by using a value sampled from the normal distribution with mean equals to the quantity obtained by evaluating the fitted model at the estimate $\hat{\pi}^{(a)}_{(k)0}(\hat{\boldsymbol{\alpha}}^{(a)}_k, \hat{\sigma}^{(a)}_k, \tilde{\tilde{\beta}}^{(a)}_{(k)0})$ of the inclusion probability of an element in $U_k - S^*_k$ and variance equals to the estimate of the variance of the error terms of the regression model. If the design matrix is numerically singular, then predict the y-value associated with $j \in U_k - S^*_k$ by a value sampled from the normal distribution with mean and variance given by the sample mean and sample variance, respectively, of the y-values associated with the elements in S^*_k. On the other hand, if the variable of interest y is binary, then fit a simple logistic regression model to the data $(\hat{\pi}^{(a)}_{(k)j}(\hat{\boldsymbol{\alpha}}^{(a)}_k, \hat{\sigma}^{(a)}_k, \tilde{\tilde{\beta}}^{(a)}_{(k)j}), y^{(k)}_j), j \in S^*_k$. Next, predict the y-value associated with the j-th element in $U_k - S^*_k$ by using a value sampled from the Bernoulli distribution with success probability equals to the quantity obtained by evaluating the fitted model at $\hat{\pi}^{(a)}_{(k)0}(\hat{\boldsymbol{\alpha}}^{(a)}_k, \hat{\sigma}^{(a)}_k, \tilde{\tilde{\beta}}^{(a)}_{(k)0})$. If the design matrix is numerically singular, then predict the y-value associated with $j \in U_k - S^*_k$ by a value sampled from the Bernoulli distribution with success probability equals to the sample mean of the y-values of the elements in S^*_k. (5) Select a SRSWOR of n values m_i from m_{Boot}. Let $S^{Boot}_A = \{i_1, \ldots, i_n\}$ be the set of indices of the m_is in the sample. In addition, let $A^{Boot}_i = (\sum_{t=1}^{i-1} m_t, \sum_{t=1}^{i} m_t) \cap \mathbb{Z}$ be the set of indices j associated with the elements in the cluster whose index is $i \in S^{Boot}_A$, where m_t is the t-th element of m_{Boot} and $\mathbb{Z}$ is the set of the integer numbers. Finally, let $S^{Boot}_0 = \cup_{i \in S^{Boot}_A} A^{Boot}_i$ be the set of indices j associated with the elements in the clusters whose indices are in S^{Boot}_A. (6) For each $k = 1, 2, i \in S^{Boot}_A$ and $j \in \{1, \ldots, \lfloor \hat{\tau}^{(a)}_1 \rfloor\} - A^{Boot}_i$ in the case of $k = 1$, or $j \in \{1, \ldots, \lfloor \hat{\tau}^{(a)}_2 \rfloor\}$ in the case of $k = 2$, generate a value $x^{(k)}_{ij}$ by sampling from the Bernoulli distribution with success probability equals to the value obtained by evaluating (1) at the i-th element of the vector $\hat{\boldsymbol{\alpha}}^{(a)}_{(k)Boot}$ and the j-th element of the vector $\hat{\beta}^{(a)}_{(k)Boot}$. (7) Compute the estimates of the sizes τ_1, τ_2 and τ; those of the totals Y_1, Y_2 and Y, and those of the means $\bar{Y}_1$, $\bar{Y}_2$ and $\bar{Y}$ using the same procedure as that used to compute the original estimates. (8) Repeat the steps (5)-(7) a large enough number B of times.

It is worth noting that in the proposed variant of bootstrap, we are using a simple linear regression model or a simple logistic regression model to roughly approximate a potential relation between the y-values and the estimated inclusion probabilities. This does not mean that there exists such a relation. However, if it existed, we would expect that the employed simple regression models yield predicted y-values that allow us to compute acceptable variance estimators and confidence intervals. Note that if there were not a relation, the predicted y-values would be basically obtained, in the case of a quantitative response variable, by sampling from a normal distribution with mean and variance equal to the sample mean and variance of the observed y-values, and in the case of a binary response variable, by sampling from a Bernoulli distribution with mean equals to the sample mean of the y-values.

9. References

Abramowitz, M., and I.A. Stegun. 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Ninth Dover printing, tenth GPO printing. New York: Dover.

Bernard, H.R., T. Hallett, A. Iovita, E.C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M.J. Salganik, T. Saliuk, O. Scutelniciuc, G.A. Shelley, P. Sirinirund, S. Weir, and D.F. Stroup. 2010. "Counting hard-to-count populations: The network scale-up method for public health." *Sexually Transmitted Infections* 86 (Suppl. 2): 11–15. DOI: http://dx.doi.org/10.1136/sti.2010.044446.

Booth, J.G., R.W. Butler, and P. Hall. 1994. "Bootstrap methods for finite populations." *Journal of the American Statistical Association* 89: 1282–1289. DOI: https://doi.org/10.1080/01621459.1994.10476868.

Burnham, K.P., and W.S. Overton. 1978. "Estimation of the size of a closed population when capture probabilities vary among animals." *Biometrika* 65: 625–633. DOI: https://doi.org/10.1093/biomet/65.3.625.

Chao, A. 1987. "Estimating the population size for capture-recapture data with unequal catchability." *Biometrics* 43: 783–791. DOI: https://doi.org/10.2307/2531532.

Cheng, S., D.J. Eck, and F.W. Crawford. 2020. "Estimating the size of a hidden finite set: Large-sample behavior of estimators." *Statistics Surveys* 14: 1–31. DOI: https://doi.org/10.1214/19-SS127.

Chow, M., and S.K. Thompson. 2003. "Estimation with link-tracing sampling designs – A Bayesian approach." *Survey Methodology* 29: 197–205. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9553-eng.pdf?st=6cXXjDD2 (accessed April 2020).

Coull, B.A., and A. Agresti. 1999. "The use of mixed logit models to reflect heterogeneity in capture-recapture studies." *Biometrics* 55: 294–301. DOI: https://doi.org/10.1111/j.0006-341X.1999.00294.x.

Crawford, F.W., J. Wu, and R. Heimer. 2018. "Hidden population size estimation from respondent-driven sampling: A network approach." *Journal of the American Statistical Association* 113: 755–766. DOI: https://doi.org/10.1080/01621459.2017.1285775.

Dávid, B., and T.A.B. Snijders. 2002. "Estimating the Size of the Homeless Population in Budapest, Hungary." *Quality & Quantity* 36: 291–303. DOI: https://doi.org/10.1023/A:1016080606287.

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

Dombrowski, K., B. Khan, T. Wendel, K. McLean, E. Misshula, and R. Curtis. 2012. "Estimating the Size of the Methamphetamine-Using Population in New York City Using Network Sampling Techniques." *Advances in Applied Sociology* 2: 245–252. DOI: https://doi.org/10.4236/aasoci.2012.24032.

Félix-Medina, M.H., and P.E. Monjardin. 2006. "Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A Bayesian assisted approach." *Survey Methodology* 32: 187–195. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9553-eng.pdf?st=6cXXjDD2 (accessed April 2020).

Félix-Medina, M.H., and P.E. Monjardin. 2010. "Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations." *Journal of Official Statistics* 26(4): 603–631. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/20200206/felix-medina.pdf (accessed April 2020).

Félix-Medina, M.H., P.E. Monjardin, and A.N. Aceves-Castro. 2015. "Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population in presence of heterogeneous link-probabilities." *Survey Methodology*: 349–376. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14238-eng.pdf?st=ki7rx2GZ (accessed April 2020).

Félix-Medina, M.H., and S.K. Thompson. 2004. "Combining Cluster Sampling and Link-Tracing Sampling to Estimate the Size of Hidden Populations." *Journal of Official Statistics* 20(1): 19–38. Available at: https://www.scb.se/contentassets/ca21efb41-fee47d293bbee5bf7be7fb3/combining-link-tracing-sampling-and-cluster-sampling-to-estimate-the-size-of-hidden-populations.pdf (accessed April 2020).

Fienberg, S.E., M.S. Johnson, and B.W. Junker. 1999. "Classical multilevel and Bayesian approaches to population size estimation using multiple lists." *Journal of the Royal Statistical Society. Series A* 162: 383–405. DOI: https://doi.org/10.1111/1467-985X.00143.

Frank, O., and T. Snijders. 1994. "Estimating the Size of Hidden Populations Using Snowball Sampling." *Journal of Official Statistics* 10 (1): 53–67. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/estimating-the-size-of-hidden-populations-using-snowball-sampling.pdf (accessed April 2020).

Freeman L.C. (n.d.) Network data sets repository. Available at: http://moreno.ss.uci.edu/data (accessed June 2018).

Giner, G., and G.K. Smyth. 2016. "statmod: Probability calculations for the inverse Gaussian distribution." *The R Journal* 8: 339–351. DOI: https://doi.org/10.32614/RJ-2016-024.

Handcock, M.S., K.J. Gile, and C.M. Mar. 2014. "Estimating hidden population size using respondent-driven sampling data." *Electronic Journal of Statistics* 8: 1491–1521. DOI: https://doi.org/10.1214/14-EJS923.

Harris, K.M. 2013. "The add health study: Design and accomplishments." Available at: www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf (accessed September 2017).

Heckathorn, D.D. 1997. "Respondent driven sampling: a new approach to the study of hidden samples." *Social Problems* 44: 174–199. DOI: https://doi.org/10.2307/3096941.

Heckathorn, D.D. 2002. "Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations." *Social Problems* 49: 11–34. DOI: https://doi.org/10.1525/sp.2002.49.1.11.

Heckathorn, D.D., and C.J. Cameron. 2017. "Network sampling: From snowball and multiplicity to respondent-driven sampling." *Annual Review of Sociology* 43: 101–119. DOI: https://doi.org/10.1146/annurev-soc-060116-053556.

Hwang, W-H., and R. Huggins. 2005. "An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data." *Biometrika* 92: 229–233. DOI: https://doi.org/10.1093/biomet/92.1.229.

Johnston, L.G., D. Prybylski, H.F. Raymond, A. Mirzazadeh, C. Manopaiboon, and W. McFarland. 2013. "Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: Case studies from around the world." *Sexually Transmitted Diseases* 40: 304–310. DOI: https://doi.org/10.1097/OLQ.0b013e31827fd650.

Kalton, G. 2009. "Methods for oversampling rare populations in social surveys." *Survey Methodology* 35: 125–141. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009002/article/11036-eng.pdf?st=PyvQkTH4 (accessed April 2020).

Khan, B., H.-W. Lee, I. Fellows, and K. Dombrowski. 2018. "One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity." *PLoS ONE* 13 (4): 1–39. DOI: https://doi.org/10.1371/journal.pone.0195959.

Killworth, P., E. Johnsen, C. McCarty, G. Shelley, and H. Bernard. 1998a. "A social network approach to estimating seroprevalence in the United States." *Social Networks* 20: 23–50. DOI: https://doi.org/10.1016/S0378-8733(96)00305-X.

Killworth, P., C. McCarty, H. Bernard, G. Shelley, and E. Johnsen. 1998b. "Estimation of seroprevalence, rape and homelessness in the United States using a social network approach." *Evaluation Review* 22: 289–308. DOI: https://doi.org/10.1177/0193841X9802200205.

Klovdahl, A.S. 1989. "Urban Social Networks: Some Methodological Problems and Possibilities." In *The Small World*, edited by M. Kochen, 176–210. Norwood, NJ: Ablex.

Korn, E.L., and B.I. Graubard. 1998. "Confidence intervals for proportions with small expected number of positive counts estimated from survey data." *Survey Methodology* 24: 193–201.

Lee, S., J. Wagner, R. Valliant, and S. Heeringa. 2014. "Recent developments of sampling hard-to-survey populations: An assessment." In *Hard-to-Survey Populations*, edited by R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, and N. Bates (Eds.), 424–444. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139381635.025.

MacKellar, D., L. Valleroy, J. Karon, G. Lemp, and R. Janssen. 1996. "The Young Men's Survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men." *Public Health Reports* 111(Suppl. 1): 138–44. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1382056/pdf/pubhealthrep00044-0140.pdf (accessed April 2020).

Magnani, R., K. Sabin, T. Saidel, and D. Heckathorn. 2005. "Review of sampling hard-to-reach populations for HIV surveillance." *AIDS* 19: S67–S72. DOI: https://doi.org/10.1097/01.aids.0000172879.20628.e1.

Maltiel, R., A.E. Raftery, T.H. McCormick, and A.J. Baraff. 2015. "Estimating population size using the network scale up method." *The Annals of Applied Statistics* 9: 1247–1277. DOI: https://doi.org/10.1214/15-AOAS827.

Marpsat, M., and N. Razafindratsima. 2010. "Survey methods for hard-to-reach populations: Introduction to the special issue." *Methodological Innovations Online* 5: 3–16. DOI: https://doi.org/10.4256/mio.2010.0014.

McCormick, T.H., M.J. Salganik, and T. Zheng. 2010. "How many people do you know? Efficiently estimating personal network size." *Journal of the American Statistical Association* 105: 59–70. DOI: https://doi.org/10.1198/jasa.2009.ap08518.

Meng, V.Y., and P. Gustafson. 2017. "Inferring population size: extending the multiplier method to incorporate multiple traits with a likelihood-based approach." *Stat* 6: 4–13. DOI: https://doi.org/10.1002/sta4.131.

Pledger, S. 2000. "Unified maximum likelihood estimates for closed capture-recapture models using mixtures." *Biometrics* 56: 434–442. DOI: https://doi.org/10.1111/j.0006-341X.2000.00434.x.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (accessed April 2020).

Reiser, M. 2019. "Goodness of fit testing in sparse contingency tables when the number of variables is large." *Wiley Interdisciplinary Reviews: Computational Statistics* 11(6): e1470. DOI: https://doi.org/10.1002/wics.1470.

Sanathanan, L. 1972. "Estimating the size of a multinomial population." *Annals of Mathematical Statistics* 43: 142–152. DOI: https://doi.org/10.1214/aoms/1177692709.

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

Spreen, M. 1992. "Rare populations, hidden populations and link-tracing designs: What and why?" *Bulletin de Méthodologie Sociologique* 36: 34–58. DOI: https://doi.org/10.1177/075910639203600103.

Spreen, M., and S. Bogaerts. 2015. "B-Graph Sampling to Estimate the Size of a Hidden Population." *Journal of Official Statistics* 31: 723–736. DOI: https://doi.org/10.1515/-jos-2015-0042.

Staudte, R.G., and S.J. Sheather. 1990. *Robust Estimation and Testing*. New York: Wiley.

St. Clair, K., and D. O'Connell. 2012. "A Bayesian model for estimating population means using a link-tracing sampling design." *Biometrics* 68: 165–173. DOI: https://doi.org/10.1111/j.1541-0420.2011.01631.x.

Thompson, S.K. 2012. *Sampling, Third edition*. New Jersey: Wiley.

Thompson, S.K., and O. Frank. 2000. "Model-based estimation with link-tracing sampling designs." *Survey Methodology* 26: 87–98. Available at https://www150.statcan.gc.ca/-n1/en/pub/12-001-x/2000001/article/5181-eng.pdf?st=dgk9U6pj (accessed April 2020).

Tourangeau, R. 2014. "Defining hard-to-survey populations." In *Hard-to-Survey Populations*, edited by R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, and N. Bates (Eds.), 3–20. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139381635.003.

UNAIDS/WHO (World Health Organization Working Group on Global HIV/AIDS, and STI Surveillance). 2010. *Guidelines on estimating the size of populations most at risk to HIV*. UNAIDS– Joint United Nations Programme on HIV/AIDS. Available at: https://data.unaids.org/pub/manual/2010/guidelines_popnestimationsize_en.pdf. (accessed June 2015).

Volz, E., and D. Heckathorn. 2008. "Probability based estimation theory for respondent driven sampling." *Journal of Official Statistics* 24: 79–97. Available at: http://www.sverigeisiffror.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/probability-based-estimation-theory-for-respondent-driven-sampling.pdf (accessed April 2020).

Williams, B.K., J.D. Nichols, and M.J. Conroy. 2002. *Analysis and Management of Animal Populations*. San Diego, California: Academic Press.

Received June 2020
Revised April 2021
Accepted June 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 907–930, http://dx.doi.org/10.2478/JOS-2021-0039

Comparing the Response Burden between Paper and Web Modes in Establishment Surveys

Georg-Christoph Haas[1], *Stephanie Eckman*[2], *and Ruben Bach*[3]

Previous research is inconclusive regarding the effects of paper and web surveys on response burdens. We conducted an establishment survey with random assignment to paper and web modes to examine this issue. We compare how the actual and perceived response burdens differ when respondents complete a survey in the paper mode, in the web mode and when they are allowed to choose between the two modes. Our results show that in the web mode, respondents have a lower estimated time to complete the questionnaire, while we do not find differences between paper and the web on the perceived response time and perceived burden. Even though the response burden in the web mode is lower, our study finds no evidence of an increased response burden when moving an establishment survey from paper to the web.

Key words: Perceived burden; experimental design; mode effects.

1. Introduction

Data on establishments are essential for monitoring national and international economies, for example, to help managers make decisions and enable politicians to craft informed policies (Jones et al. 2013). A large proportion of establishment data originates from surveys. However, for most establishments, responding to a survey is a task unrelated to business production, which potentially takes employee time away from other essential tasks (Willimack and Nichols 2010). This article is particularly concerned with response burden, which unfortunately is loosely defined in the literature (Yan et al. 2020). We define response burden as the strain experienced by respondents while they respond to a survey. Factors affecting response burden are multifaceted and include questionnaire design, content and length, question wording, and the data collection mode.

When the response burden is high, respondents have difficulties answering a questionnaire (Couper and Groves 1996). In establishment surveys, a high response burden is associated with low data quality and high data collection costs (e.g., Bavdaž et al. 2015; Jones 2012; Giesen 2012; Giesen et al. 2011; Hedlin et al. 2005; Haraldsen and Jones 2007). A high burden can also lead to more data editing and fewer timely responses (e.g., Haraldsen and Jones 2007; Berglund et al. 2013; Giesen 2013a) and may reduce respondents' motivation and efforts to answer correctly (Krosnick 1991).

[1] Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung), Statistical Methods, Regenburger Straße 100, Nürnberg, 90478, Germany. Email: georg-christoph.haas@iab.de
[2] RTI International (Statistical Methods), 701 13th St NW, Suite 750, Washington, D.C. 20005, U.S.A. Email: seckman@rti.org
[3] University of Mannheim, B5, Mannheim, 68159, Germany. Email: r.bach@uni-mannheim.de

One way to reduce the response burden in establishment surveys may be to change the survey mode from paper to the web. The web mode offers many advantages that can reduce the response burden. However, it can also introduce response burden if respondents are not comfortable with website navigation and forms. In practice, many surveys offer a choice of web or another mode, often paper. The choice of mode may allow respondents to choose their preferred mode, leading to a lower burden; or it may present respondents with another decision they must make, leading to a higher burden, as in Medway and Fulton (2012). Studies of the change in the response burden when moving establishment surveys from paper to the web have found that the introduction of the web mode reduces the response burden (Giesen 2013b; Gravem et al. 2011; Giesen et al. 2009) or has no effect (Snijkers et al. 2007). However, because the questionnaire content and structure in those studies also changed, we cannot draw a definitive conclusion on the effect of web on response burden (Gravem et al. 2011).

To address the shortcomings of previous studies, we conducted an establishment survey with an experimental assignment to the mode: *Paper-only*, *Web-only* or concurrent *Paper and Web mixed mode*. We examine the differences in response burden between modes, and we will answer the following two research questions:

1. Is response burden in an establishment survey lower in the web mode than it is in the paper mode?, and
2. Do respondents experience a lower burden if they can choose between the paper and the web mode?

To answer our research questions, we first define what type of response burden we evaluate. Second, by listing the benefits of the web mode, we explain why data collection agencies are interested in using the web mode for their establishment surveys. Third, we provide a literature overview on how response burden is measured. Fourth, we describe the possible effects of paper and the web on response burden, leading us to our hypotheses. Fifth, we describe our data, including our study and experimental design as well as key features of our web survey. Sixth, we describe the models we use to evaluate response burden differences. Seventh, we present our results. Finally, we summarize our results and the limitations of their scope.

2. Background

Establishment surveys can impose burden in three ways (Löfgren 2011; Haraldsen et al. 2013). First, each time an establishment is selected for a survey, the establishment is burdened with a response request, and large establishments are selected more often than medium-sized and small establishments (Jones 2012). Second, for those establishments that choose to participate, the participation costs are presumably greater than the benefits to the establishment (Verkruyssen and Moens 2011; Giesen 2011). As a result, establishments may have a low motivation to respond. Third, instrument design introduces burden through questionnaire content and length, the data collection mode (e.g., face-to-face, telephone, paper, and web), the wording of questions and other factors. This article focuses on burden introduced through instrument design. We refer to this type of burden as *response burden*. Specifically, we focus on the mode as part of the instrument design and

compare the difference in response burden between paper and web modes in establishment surveys.

In the remainder of this section, we provide a short overview of the benefits of web surveys compared to paper surveys. We explain how response burden is conceptualized and measured, and the possible effects of paper and web surveys on response burden. We then develop hypotheses regarding how response burden differs between paper and web surveys.

2.1. Benefits of Web Surveys

Although paper and web are both cost-efficient self-administered modes, web offers several advantages over paper. Web surveys reduce or eliminate mailing costs. Many establishment survey invitations can be sent via email; when mail invitations are used, only an invitation letter is sent rather than a large paper questionnaire and return envelope. Furthermore, web surveys reduce data entry costs. These savings usually more than offset potential increases in programming needed to set up the web survey.

The web mode can also increase data quality. Web questionnaires can provide feedback to respondents (Couper 2008; Conrad et al. 2007). If respondents submit an unlikely answer, plausibility checks can ask respondents to re-evaluate their answers, which could reduce the need for data editing. Furthermore, researchers can offer definitions and additional information on how to answer the question. Future web surveys may even include chatbots that can address respondents' questions during the response process (Lagerstøm 2018). Additionally, the web mode can manage calculation and counting tasks, which simplify responses (Giesen et al. 2009; Giesen 2007). Especially in establishment surveys, which often require responses from multiple respondents, web surveys may simplify the response process within the establishment as respondents can easily distribute a link for a web survey via email, while a paper questionnaire is more cumbersome to distribute to multiple respondents. Finally, web surveys enable a complex filter and skip pattern design while only showing items applicable to each respondent.

2.2. Conceptualizing and Measuring Response Burden

Bavdaž et al. (2015) summarize three reasons why National Statistical Institutes (NSIs) should consider response burden when designing data collection programs. The first is political: responding to a survey takes time away from an establishment's core business and may decrease competitiveness. The second is methodological: a high burden may reduce data quality and increase data collection costs. The third is strategic: burden can negatively affect the relationship between NSIs and the business community, reducing the motivation to respond to surveys. Therefore, NSIs should monitor and reduce burden to the fullest extent possible (e.g., see European Commission 2011), and burden management has become a key element for NSIs (e.g., see Giesen et al. 2018).

Response burden is a multifaceted concept influenced by motivation, task difficulty, survey effort and respondent perception (Yan et al. 2020). It is often "loosely defined", and Yan et al. argue for a unified concept of response burden. For our study, we follow the conceptualization of actual and perceived response burden, which we find is the most prominent within the establishment survey literature (e.g., Giesen 2013a; Berglund et al.

2013; Hedlin et al. 2005; Giesen et al. 2009; Giesen and Burger 2013; Haraldsen and Jones 2007). The literature suggests several indicators to measure actual response burden. Because respondents need time to read, think, and respond to a question, each item in the survey adds to the overall burden (Bradburn 1978). Therefore, questionnaire length is probably the most basic indicator for response burden (see, e.g., Groves et al. 1992; Van Loon et al. 2003). In our study, we asked respondents how much time they spent answering the questions (see, e.g., Dale et al. 2007; Giesen et al. 2011; Giesen 2013b). Additional indicators used by NSIs to track response burden imposed on establishments include the following: calls to the service number, requests for help, response rates, and average time for questionnaire completion (Downey et al. 2007; Snijkers et al. 2007; Sear 2011; Giesen et al. 2011).

Perceived response burden is a subjective measure of respondents' experiences responding to the survey, for example, as burdensome and time consuming (see, e.g., Haraldsen et al. 2013). It is not the actual time spent taking a survey but the perception of the time and effort of the survey that affects respondents' survey experience and response quality (e.g., Haraldsen and Jones 2007). Many factors can contribute to perceived response burden: structures within the establishment (who has the information needed to respond), the timing of a survey (during a firm's busy period or while a key informant is on vacation), question design, data collection mode, number of survey invitations, difficulty of the response task, and attitudes towards the data collector (Hedlin et al. 2005; Giesen 2013b).

Perceived response burden is often collected with two items. One item asks for the perception of time on a five-point scale, that is, if respondents perceive the survey as quick or time-consuming. The other item asks for the perception of burden on a five-point scale, for example, if respondents perceive the questionnaire as easy or burdensome to answer (see, e.g., Dale et al. 2007; Giesen et al. 2011; Giesen 2013b). We use the same perceived response burden indicators for our study.

Actual and perceived response burdens are conceptually different from each other but positively correlated (Giesen 2013a; Berglund et al. 2013). If respondents perceive a questionnaire as difficult, the actual response burden (time spent) is also likely high (Giesen 2013a). Giesen et al. (2011) found that 34 of 41 NSIs collect data on actual response burden, while 12 collect data on perceived response burden. We examine how assigned mode and mode choice affect both actual burden and perceived burden.

2.3. *Possible Effects of Paper and Web Surveys on Response Burden*

The impact of mode on actual and perceived burden is complex. Each page of a paper questionnaire introduces an additional workload, and respondents may perceive multipage questionnaires as burdensome. Even if not all questions apply to the respondent, the number of pages can make the survey seem overwhelming. Skip instructions in paper questionnaires may not be clear to respondents, and they may have a hard time navigating a paper survey. Web surveys, on the other hand, do not show all questions to the respondent but only those that apply. As a result, respondents never see the entire questionnaire and cannot immediately assess its total length. They also do not need to pay attention to filter instructions, which reduces the respondent's cognitive effort.

On the other hand, the web mode could increase response burden. Respondents with lower online skills may experience a greater burden (Gregory and Earp 2007). A poorly

designed instrument can be difficult or frustrating to fill out. Furthermore, even well-designed plausibility checks may increase response burden (Hedlin et al. 2005).

Most NSIs do not use web as a standalone mode but in combination with other modes of survey data collection, often a paper mode. Offering the web in addition to paper may reduce the perceived response burden: faced with a choice of mode, respondents should choose the mode they feel most comfortable responding to and the one that is lower burden for them (Erikson 2007). Lyly-Yrjänäinen and Van Houten (2011) propose offering multiple modes to reduce the respondent burden in Eurostat establishment surveys. However, offering multiple modes can overwhelm respondents and reduce response rates (Medway and Fulton 2012). Requiring respondents to choose a mode before they can begin the survey may also impose an additional burden on respondents.

2.4. Hypotheses

The above discussion leads us to several hypotheses regarding the relationship between the mode and response burden in establishment surveys. In accordance with the findings from earlier research (Gravem et al. 2011; Giesen et al. 2009; Snijkers et al. 2007), we hypothesize that burden will be lower for respondents assigned to the web mode than for those assigned to the paper mode (hypothesis 1).

Therefore, compared to the paper mode, we expect:

- a shorter time to complete the questionnaire in the web mode (hypothesis 1.1)
- a lower perceived time in the web mode (hypothesis 1.2)
- a lower perceived burden in the web mode (hypothesis 1.3)

Hypothesis 2 relates to mode choice: when respondents can choose their mode, they are likely to experience a lower burden than respondents who respond in the same mode but were not given a choice. We hypothesize that actual and perceived response burden among those who choose the web mode from a mixed-mode condition are lower than burden among those assigned to the web mode (hypothesis 2.1). Therefore, compared to the assigned web condition, we expect:

- a shorter time to complete the questionnaire by web respondents in the mixed-mode condition (hypothesis 2.1.1)
- a lower perceived time by web respondents in the mixed-mode condition (hypothesis 2.1.2)
- a lower perceived burden by web respondents in the mixed-mode condition (hypothesis 2.1.3)

Similarly, we expect lower actual and perceived response burden for respondents in the paper mode from a mixed-mode condition compared to respondents from the assigned paper condition (hypothesis 2.2). That is, compared to the assigned paper condition, we expect:

- a shorter time to complete the questionnaire by paper respondents in the mixed-mode condition (hypothesis 2.2.1)
- a lower perceived time by paper respondents in the mixed-mode condition (hypothesis 2.2.2)
- a lower burden by paper respondents in the mixed-mode condition (hypothesis 2.2.3)

Although respondents likely use their preferred mode when choosing between paper and web, we should still see differences in response burden between those choosing paper and those choosing web. The features of the web mode described earlier should reduce response burden. Therefore, we hypothesize that actual and perceived response burden will be lower for those who respond via the web in the mixed-mode condition than for those who respond via paper in the mixed-mode condition (hypothesis 3). Therefore, compared to those choosing paper, we expect:

- a shorter time to complete the questionnaire for those choosing web (hypothesis 3.1)
- a lower perceived time for those choosing web (hypothesis 3.2)
- a lower perceived burden for those choosing web (hypothesis 3.3)

3. Data

To examine our hypotheses regarding the differences in response burden between modes, we use data from a German establishment survey. The Institute for Employment Research (IAB) designed this survey to evaluate the effect of the mode on the data quality in establishment surveys.

Overall, 16,000 establishments were sampled from German administrative records. Sample selection was stratified by location (East and West Germany), establishment size class ($<$ 10 employees, 10–199 employees, and $\geq$ 200 employees) and industry class following the German Classification of Economic Activities (Destatis 2008). Establishments already selected for IAB surveys in 2015 were removed from the frame before selection to avoid causing any problems for those ongoing data collection efforts. The removed establishments were random selections from the frame and thus should not bias the sample. However, there are some strata where no unselected establishments remained on the frame. This issue particularly affected the largest size class in which there are few establishments. For this reason, the sample used in this study is not fully representative of the population of establishments, but efforts were made to be as complete as possible given the need to avoid overlap with ongoing surveys. Participation in the survey was voluntary, and the overall response rate was 10.2% (AAPOR RR1, according to AAPOR standard definitions, see AAPOR 2016) with 1,574 establishments responding.

All sampled establishments were randomly assigned to one of the three mode conditions (*Paper-only, Web-only, Choice*). To ensure we would have enough cases in all three mode groups and within the two modes in the *Choice group*, we assigned one-fourth of the establishments in our sample to *Paper-only*, one-fourth to *Web-only* and two-fourths to *Choice*.

We prepared two versions of the questionnaire with different topics, number of items and question formats. One version focused on the consequences of the introduction of the federal minimum wage in Germany in 2015. We refer to this version as *Minimum Wage*. Another version contains questions about the effect of increasing digitalization on labor markets. We refer to this version as *Digitalization*. We randomly assigned each sampled establishment to one of the two versions. Therefore, both versions are independent surveys with the same experimental mode design. However, our hypotheses should apply to both questionnaire versions. In fact, seeing similar results over both versions should increase the reliability of our results. All mode groups were invited to participate in the study via a

mailed letter. For the *Paper-only* group, we mailed establishments a cover letter with information about the study and a paper questionnaire. Depending on the assigned versions, the number of pages and questions differed slightly. The *Minimum Wage* questionnaire contained 74 questions on 20 pages. In contrast, the *Digitalization* questionnaire had 69 questions on 19 pages, printed in a 20-page booklet. Therefore, the difference in page volume between both versions was negligible.

For the *Web-only* group, we sent establishments a cover letter with information about the study, a link and the request to fill out our online questionnaire. To isolate mode effects, we took care to ensure that the paper and web questionnaire were visually similar to each other. However, the web mode offers functionalities that may reduce response burden, as discussed above. We implemented six web survey functionalities. First, the web survey presented questions in a paging design (one question on each page) so that respondents would not miss a question. Second, the web survey used automatic skips, that is, questions that did not apply to respondents were not shown. Third, the question about the number of different employment groups automatically summed and displayed the total number of employees. Fourth, we implemented plausibility checks. For instance, if the respondent stated that the regular weekly working hours were greater than the legal limit of 48 hours, the web survey prompted an error message in red that asked respondents to re-evaluate their answer. The number of plausibility checks differed by questionnaire version: *Minimum Wage* contained up to 13 plausibility checks, and *Digitalization* contained up to five plausibility checks. Fifth, at the end of each section, respondents were able to print the questionnaire section with their responses for their own documentation.

Sixth, the web survey contained an index that allowed respondents to navigate to specific sections. The index indicated the structure of the questionnaire and showed the headings for each section (see Figure 1). After finishing a section, the web survey redirected respondents to this index page. The index page gave respondents an understanding of what part of the questionnaire should be answered by whom in the establishment. In establishment surveys, respondents sometimes do not have the

Fig. 1. Index page for the web survey in the digitalization version.

information required to answer all questions. Therefore, they require help from colleagues to answer some questions.

For the establishments in the *Choice* groups, we sent a cover letter and the same paper questionnaire as in the *Paper-only* group. The cover letter offered a web link and presented the option to choose between the paper and web modes.

Figure 2 shows the response rates (RR) for our three mode groups and two versions. In the figure, *Choice-Paper* refers to cases that chose to respond via the paper mode in the mixed-mode condition. *Choice-Web* refers to those that responded on the web. In both versions, response rates in the *Web-only* and *Choice-Web* groups are smaller than those in the *Paper-only* and *Choice-Paper* groups. The response rates are also lower in all conditions for the *Digitalization* survey than the *Minimum Wage* survey (8.5% versus 11.9%). Furthermore, we find that compared to *Paper-only*, the *Choice* group is not different in terms of response rates (13.7% versus 13.9% in the *Minimum Wage* survey; 11.8% versus 11.7 % in the *Digitalization* survey). (To calculate the response rate for *Choice-Paper* and *Choice-Web*, we split the response rate of *Choice* into the proportion of *Choice-Paper* and *Choice-Web*, that is, $RR_{Choice} = RR_{Choice-Paper} + RR_{Choice-Web}$.) These results contradict findings from meta-analyses where offering a choice between modes is burdensome enough to not participate (Medway and Fulton 2012). However, the meta-analysis did not include establishment surveys.

To check whether respondents in each mode group differ from each other, a nonresponse analysis for the variables location (East and West Germany), establishment size class (< 10 employees, 10–199 employees, and ≥ 200 employees) and industry was conducted (see Haas et al. 2016). No systematic differences in nonresponse patterns between the mode groups were found.

Involving other people and managing the response process can be a burden to respondents. Overall, 16.2% of our respondents reported that they had help answering the questionnaire. Concerning the proportion of multiple respondents, a chi-squared test suggests no differences between the mode groups and questionnaire versions ($\chi^2_{7,\ N=1,663} = 5.6$, $p < 0.585$).

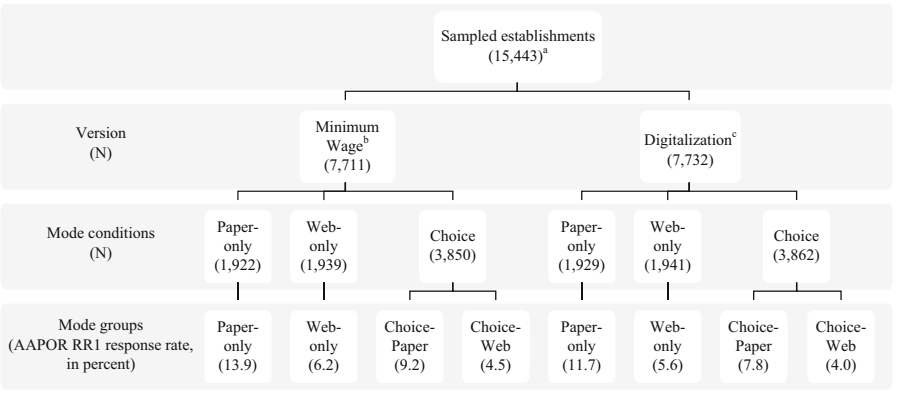

Fig. 2. Experimental assignment and response rates (RR).
[a]All Ns exclude 557 cases found to be ineligible.
[b]Overall AAPOR RR1 for Minimum Wage is 11.9%.
[c]Overall AAPOR RR1 for Digitalization is 8.5%.

4. Methods

To evaluate the differences in response burden between survey modes, we use median and ordered regression models. The dependent variables in all models are measures of response burden. The independent variables are the experimental conditions (mode and topic) and control variables about the establishments (size class, industry, and East versus West Germany).

4.1. Response Burden Variables

We measure actual and perceived burden with three questions (Dale et al. 2007) asked at the end of the questionnaire. First, we asked respondents to estimate the *time they needed to complete* the questionnaire. The question required an answer in hours and minutes and has been used as a measure of actual burden in earlier studies (e.g., Dale et al. 2007; Giesen 2013a; Berglund et al. 2013). However, as respondents retrospectively estimate time and do not actively measure it, our measure of actual burden is not as objective as the literature may suggest. Second, we asked respondents to rate the *perceived time* taken on a five-point scale from "very quick" to "very time consuming". Third, we asked respondents to rate the *burden* of *the* survey on a five-point scale ranging from "very easy" to "very burdensome". For the sake of simplicity, we will refer to these two variables as *perceived burden indicators*. Furthermore, we recode our scales from five points to three points (0, 1 and 2) by collapsing the two categories at each end. The results are not substantially different between the five- and three-point scales, but the three-point scale makes it easier for the reader to interpret the results. For the full wording of the three burden questions and response options, see Appendix (Section 7) Table 4.

4.2. Independent Variables

We have four mode groups, that is, *Paper-only, Web-only, Choice-Paper,* and *Choice-Web*, which are our independent variables of interest. We can test our three hypotheses by comparing the four groups. First, we compare *Paper-only* and *Web-only* to test whether response burden is lower for web in an establishment survey (hypothesis 1). Second, we compare *Paper-only* and *Choice-Paper* as well as *Web-only* and *Choice-Web* to test whether having the chance to choose a mode affects response burden (hypothesis 2). Third, we compare *Choice-Paper* and *Choice-Web* to test whether response burden is lower among respondents who opted for the web mode (hypothesis 3).

The models also control for the number of questions the respondent answered. Due to filters and skip patterns, the number of questions each respondent answered was not tightly controlled, even within the same questionnaire version. Therefore, we introduce the variable *number of applicable items* for each respondent. This variable counts the number of items respondents should have answered from the start of the interview until the response burden questions. In the last section, we asked respondents which questionnaire sections they answered themselves (as opposed to which ones a colleague answered). If they reported that more than one person answered the questionnaire, we consider only the number of items that the final respondent answered in the model because that respondent was the one who answered the burden questions. Furthermore, we include the indicator of more than one respondent in the model as a dummy variable.

The models also control for location, size and industry to account for possible selection bias between modes and to increase the precision of our estimates.

4.3. Models

To evaluate our hypotheses on response burden differences between modes, we use multivariate regression models. We ran a model for each of our three response burden variables: *time to complete the questionnaire*, *perceived time* and perceived *burden*. Furthermore, we ran our models for each questionnaire version separately. Therefore, we have six models. Because we do not claim to represent the population of establishments, all analyses are unweighted. Each model does include the three stratification variables as controls in all models; however, they are the only variables that influence the weights. Controlling for the components of sample weights is an alternative to the use of weights in regression analyses (Gelman 2007).

Because the dependent variables have different scales, we use different models. Our response burden variable *time to complete the questionnaire* has large outliers (see Table 1). For this reason, we use a median regression that is less susceptible to being influenced by very short and very long times than an ordinary least squares regression (e.g., Cameron and Trivedi 2005):

$$y_i = M_i^{'}\beta_M + X_i^{'}\beta_X + \varepsilon_i \tag{1}$$

where y_i is the time to complete the questionnaire for a questionnaire version, $M_i^{'}$ is the mode group, $X_i^{'}$ are the controls and ε_i are the unobserved variables or errors.

Using a median regression, we assume that $(\varepsilon_i | M_i^{'}, X_i^{'}) = 0$, which implies that:

$$MED(y_i \,|\, M_i^{'}, X_i^{'}) = M_i^{'}\beta_M + X_i^{'}\beta_X \tag{2}$$

The two *perceived burden* variables are ordinal scales, and we use ordinal logistic regression models with these variables (e.g., see Cameron and Trivedi 2005, 519 f.) and adapt our model as follows:

$$y_i^{*} = M_i^{'}\beta_M + X_i^{'}\beta_X + \varepsilon_i \tag{3}$$

Table 1. Summary statistics for the time to complete the questionnaire in minutes by questionnaire version and mode group.

	Minimum wage					Digitalization				
	N	Mean	Median	Min	Max	N	Mean	Median	Min	Max
Paper-only	272	34.4	30	5	180	190	48.9	35	10	240
Web-only	116	26.2	20	2	120	91	38.2	30	5	165
Choice-paper	355	37.5	30	5	210	245	55.4	30	5	1,440
Choice-web	171	32.1	20	1	210	123	44.9	30	1	1,200
Overall	914	34.2	30	1	210	649	49.1	30	1	1,440

$$Y_i = \begin{cases} 0 \;\; if \;\; y_i^* \leq \alpha_0 \\ 1 \;\; if \;\; \alpha_0 < y_i^* \leq \alpha_1 \\ 2 \;\; if \;\; \alpha_1 < y_i^* \end{cases} \tag{4}$$

where y_i^* is one of our perceived burden indicators and α_i the threshold parameters that are obtained by maximizing the log-likelihood. We calculate the marginal effects in the probabilities as follows:

$$\frac{\delta \Pr[y_i = j]}{\delta M_i'} = \left\{ F'\left(\alpha_{j-1} - (M_i'\beta_M + X_i'\beta_X\right) - F'\left(\alpha_j - (M_i'\beta_M + X_i'\beta_X\right) \right\} \beta_M \tag{5}$$

where F'denotes the derivative of the cumulative distribution function of ε_i.

The independent variables in all models are the same. Table 2 summarizes the six models. Because we focus on the differences between modes, we report only the linear prediction of the median time from the median regression ($\frac{\mathrm{MED}(y_i \mid M_i', X_i')}{\delta M_i'} = \beta_M$) and the predicted probabilities from the ordinal logistic regression (Equation (5)) for our mode groups.

The results of each model provide information supporting or rejecting our hypotheses. Running the models on the two questionnaire versions separately provides us with information about whether our results hold across both survey topics. Support for hypotheses 1.1 to 1.3 (response burden is lower in the web mode than in the paper mode) will be seen by comparing the coefficients of the mode indicators for Paper-only and Web-only. For the time to complete the questionnaire, we expect to see a lower estimated time for the Web-only group. For both perceived indicators, we expect to see higher predicted probabilities for the categories "quick" (perceived time) and "easy" (perceived burden) in the Web-only group. For hypotheses 2.1.1 to 2.1.3, we compare the coefficients of Choice-Web against Web-only; and for hypotheses 2.2.1 to 2.2., we compare Choice-Paper and Paper-only. We expect a lower burden in the Choice conditions than in the Only conditions. For hypotheses 3.1 to 3.3, we compare the coefficients of Choice-Web against

Table 2. Summary of the six models for evaluating the response burden.

Model	Dependent variable	Questionnaire version	Model type	Independent variables
1	Time to complete	Minimum wage	Median regression	• Mode • Number of applicable items • Establishment size • Industry • Region • Multiple respondents (Yes/No) • Interaction of mode with each of the above (except mode)
2	Time to complete	Digitalization		
3	Perceived time	Minimum wage	Ordinal logistic regression	
4	Perceived time	Digitalization		
5	Perceived burden	Minimum wage		
6	Perceived burden	Digitalization		

Choice-Paper. We expect all three models to indicate lower burden in Choice-Web than Choice-Paper.

5. Results

Before presenting the results of our hypothesis tests, we examine the burden within the two questionnaire versions with the three response burden indicators (*time to complete the questionnaire*, *perceived time* and *perceived burden*).

On average, respondents to the *Minimum Wage* version needed less time to *complete the questionnaire* (34 versus 49 minutes) than respondents in the *Digitalization* version. As the data for the *time to complete the questionnaire* is not normally distributed (see Table 1), we cannot conduct a two-sample t-test. However, a nonparametric equality-of-medians test (see Snedecor and Cochran 1989) shows that *complete time* ($\chi^2_{1,\ N=1,563} = 50.1$, $p < 0.001$) is different between the two versions.

Table 3 shows the descriptive results of our perceived time indicators for each questionnaire version independent of the mode. We use a chi-squared test to examine differences in the perceived time indicators between our questionnaire versions. Overall, the *Digitalization* version is perceived as more time consuming ($\chi^2_{2,\ N=1,668} = 32.0$, $p < 0.001$) and burdensome ($\chi^2_{2,\ N=1,660} = 54.6$, $p < 0.001$) than the *Minimum Wage* version (see Table 4). Because burden is very different in the two questionnaire versions, we run separate models for the two versions in the rest of the article.

5.1. *Hypothesis 1: The Response Burden in the Web-only Mode is Lower than that in the Paper-only Mode*

We hypothesized that the web mode leads to a lower response burden. We test this hypothesis using the six models described in the methods section. For the time to complete the questionnaire, we expect to see a lower estimated time for the *Web-only* group than for the *Paper-only* group. For both *perceived indicators*, we expect to see higher predicted probabilities for the categories "quick" (perceived time) and "easy" (perceived burden) in the *Web-only* group compared to the *Paper-only* group.

Figure 3 compares the marginal effects of the four mode conditions on the median *time to complete the questionnaire* for the *Minimum Wage* version. At the median, respondents

Table 3. Proportions of perceived time and burden by questionnaire version.

Perceived time*	Minimum wage (N = 967)	Digitalization (N = 701)
Quick	57.3	44.4
Neither	34.1	40.5
Time consuming	8.6	15.0
Perceived burden**	**Minimum wage (N = 962)**	**Digitalization (N = 698)**
Easy	66.6	49.0
Neither	29.0	42.1
Burdensome	4.4	8.9

*$\chi^2 = 32.0$, $p < 0.001$; **$\chi^2 = 54.6$, $p < 0.001$

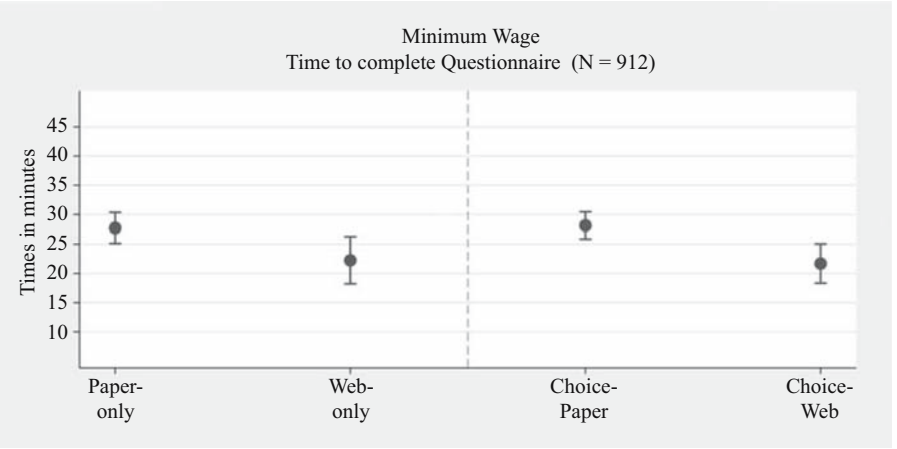

Fig. 3. Linear prediction of the estimated median time to complete the questionnaire in minutes for the minimum wage questionnaire (bars show 95% confidence intervals).

assigned to the *Web-only* group needed 5.5 fewer minutes to complete the questionnaire than respondents in the *Paper-only* group (based on self-reported completion time; $F_{1,\ 886} = 4.9$, $p < 0.013$). As the *time to complete the questionnaire* is lower in the web group, the results support hypothesis 1.1 that the web mode has a lower actual burden than the paper mode.

Figure 4 shows the average predicted probabilities from the ordinal logistic regression model for respondents' perceived time and burden over the four mode groups in the *Minimum Wage* version. The predicted probabilities provide us with a measure of how the respondents perceived responding to the survey mode while controlling for our independent variables (size, industry, number of applicable items and region). The left panel shows the results for *perceived time*, and the right panel shows the results for

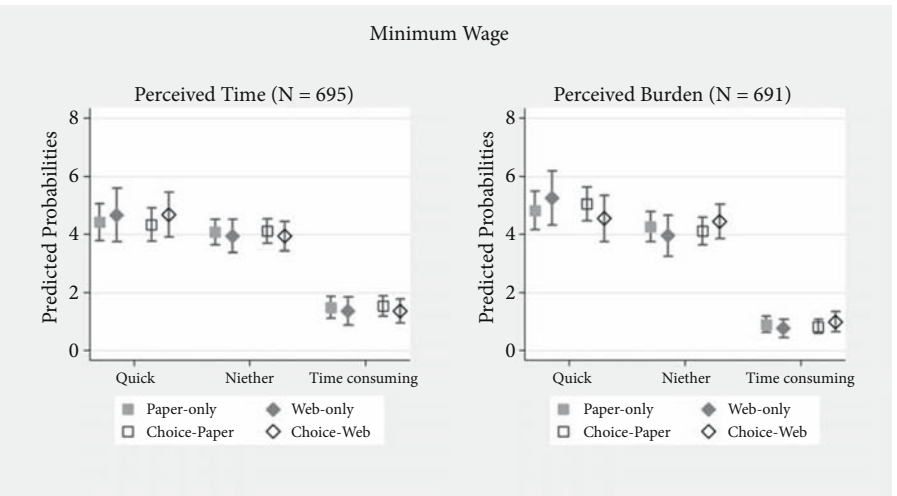

Fig. 4. Predicted probabilities from the ordinal logistic regression model for perceived time and perceived burden in the minimum wage version by mode group (bars show 95% confidence intervals).

perceived burden. Our model predicts similar probabilities for *Paper-* and *Web-only* for perceived time ($\hat{p}_{\text{quick}} = 0.54\text{-}0.61$, $\hat{p}_{\text{neither}} = 0.32\text{-}0.37$, and $\hat{p}_{\text{time consuming}} = 0.07\text{-}0.09$) and perceived burden ($\hat{p}_{\text{easy}} = 0.64\text{-}0.72$, $\hat{p}_{\text{neither}} = 0.25\text{-}0.31$, and $\hat{p}_{\text{burdensome}} = 0.03\text{-}0.05$). We see no variation between *Web-only* and *Paper-only* for either of our *perceived burden indicators* ($\chi^2_{4,\ N=409} = 3.2$, $p = 0.53$ for *perceived time* and $\chi^2_{4,\ N=409} = 0.7$, $p = 0.95$ for *perceived burden*). The results from the two *perceived burden indicators* do not support hypotheses 1.2 and 1.3.

The results for the *Digitalization* version are similar (see Figure 5 for the *time to complete the questionnaire* and see Figure 6 for *perceived time indicators*). Figure 5 shows that the estimated time for completing the questionnaire is ten minutes lower in *Web-only* ($F_{1,\ 615} = 6.0$, $p = 0.007$) and supports hypothesis 1.1 that response burden is lower for the web mode. In Figure 6, there is no variation in the marginal predicted probabilities between *Web-only* and *Paper-only* for *perceived time* ($\chi^2_{4,\ N=298} = 0.1$, $p = 1.0$) and *perceived burden* ($\chi^2_{4,\ N=295} = 0.5$, $p = 0.97$). Therefore, our results do not support hypotheses 1.2 and 1.3 that the perceived response burden is lower in the web mode.

5.2. *Hypothesis 2: The Burden is Lower when Respondents Choose a Mode than when that Mode is Assigned*

We hypothesized that the possibility of choosing one's preferred mode lowers the burden of respondents. Therefore, the estimated time should be smaller for the (2.1) *Choice-Paper* group than for the *Paper-only* and for the (2.2) *Choice-Web* group than the *Web-only*.

For the *Minimum Wage* version (see Figure 3), the difference between *Paper-only* and *Choice-Paper* in the *time to complete the questionnaire* is 0.4 minutes. In the group comparison of *Web-only* and *Choice-Web*, we find a 0.6-minute difference. For the *Digitalization* version (see Figure 5), there is no difference in the *time to complete the questionnaire* between the *Paper-only* and *Choice-Paper* groups and between the *Web-only* and *Choice-Web* groups.

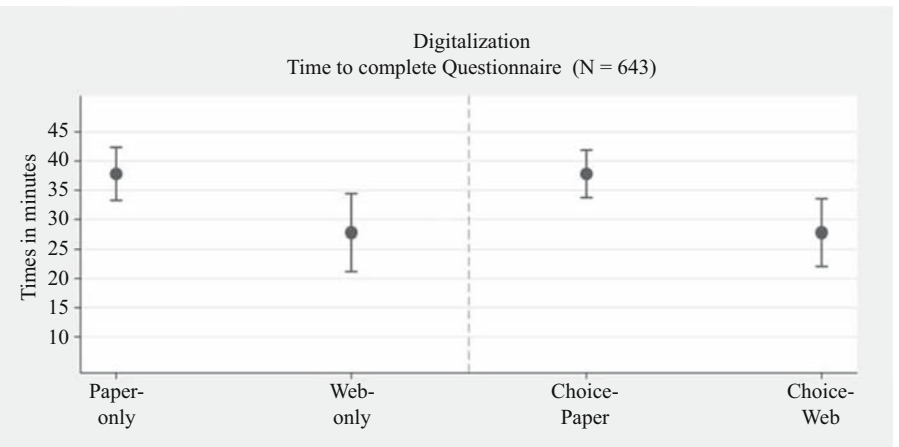

Fig. 5. Linear prediction of the estimated median time in minutes to complete the questionnaire for digitalization questionnaire (bars show 95% confidence intervals).

For our *perceived burden indicators* (see Figure 4 for the *Minimum Wage* version and Figure 6 for the *Digitalization* version), there is no variation in the predicted probabilities between our *Choice* and *Only* groups (see Appendix Table 5 for the joint χ^2 values). Overall, we find no support for our hypotheses 2.1.1 to 2.2.3.

5.3. *Hypothesis 3: The Response Burden in the Web Mode is Lower than that in the Paper Mode (mode Choice)*

Our third hypothesis is similar to the first, but it compares paper and the web when a choice is offered. We hypothesized that among those respondents given a choice, the web mode should have a lower response burden than the paper mode.

In the *Minimum Wage* questionnaire, *Choice-Web* respondents needed 6.4 fewer minutes *to complete the questionnaire* ($F_{1,\ 886} = 9.5$, $p = 0.001$) (see Figure 3) than *Choice-Paper* respondents. In the *Digitalization questionnaire*, the differences were larger: the estimated median *time for completing the questionnaire* is ten minutes lower in *Choice-Web* ($F_{1,\ 625} = 7.6$, $p = 0.003$) (see Figure 5).

Examining Figure 4 and Figure 6, we see no variation in the predicted probabilities between *Choice-Paper and Choice-Web* (see Appendix Table 5 for the joint χ^2 values). Therefore, we find mixed support for our hypothesis: when offered a choice, those choosing the web mode have a lower estimated time for completing the questionnaire (hypothesis 3.1), but there is no difference in the perceived burden (hypotheses 3.2 and 3.3).

6. Conclusion

We designed this study to determine the differences in response burden between paper and web modes in a German establishment survey. We designed two surveys with the same experimental mode groups. To evaluate response burden, we used three measures

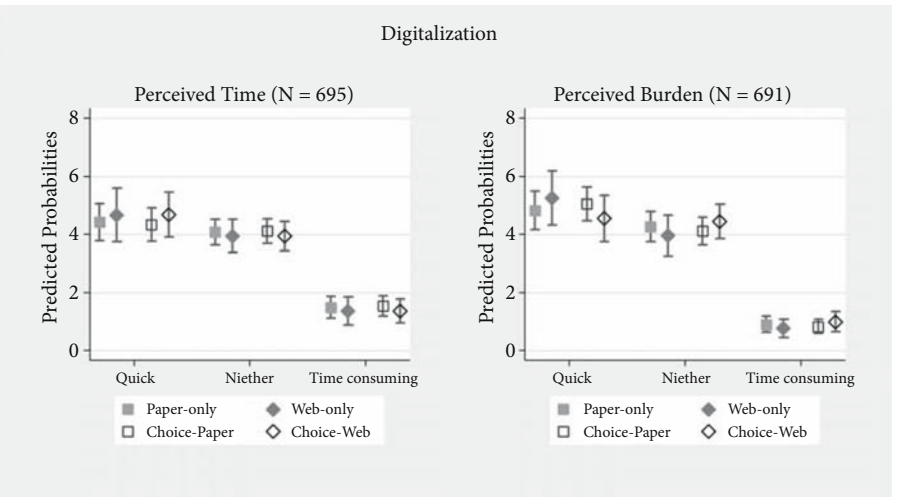

Fig. 6. Predicted probabilities from the ordinal logistic regression model for perceived time and perceived burden in the digitalization version by mode group (bars show 95% confidence intervals).

of burden (estimated time to complete the questionnaire, perceived time and burden) and four mode comparisons *(Paper-only* versus *Web-only*, *Choice-Paper* versus *Paper-only*, *Choice-Web* versus *Web-only*, *Choice-Paper* versus *Choice-Web)* to answer our research questions about whether response burden is lower in an establishment web survey and whether respondents feel less burdened if they can choose between paper and web modes.

This study has shown that web respondents, whether they were offered the web as a standalone mode or concurrently with a paper questionnaire, have a lower median time to complete the questionnaire compared to a paper questionnaire. These results held when respondents chose the web mode and when they were assigned to the web mode.

We found no evidence of a difference in either measure of perceived burden between the paper and web modes. As we have mentioned at the beginning of this article, response burden is a multifaceted concept. It is important to note that perceptions of burden could be affected by factors other than time. For instance, a questionnaire that seems relevant and straightforward to respondents might be less burdensome than a shorter but more difficult instrument.

Our results suggest that offering respondents the choice of their preferred mode has no effect on response burden compared to a single-mode setting. Therefore, concerning response burden, the web mode is a cost-effective alternative to the paper mode. Furthermore, the results of our study are consistent across two different topics: *Minimum Wage* and *Digitalization*. Therefore, our results may also be applicable to other surveys.

A reason why we find a lower estimated time for the time to complete the questionnaire may be that the web mode has an automatic questionnaire flow and does not show unnecessary questions to the respondents. However, the estimated time to complete could also indicate that web respondents are more satisfied than paper respondents. Future research needs to address this question.

One could argue that the lower response rate in the web survey is a sign that respondents find that mode more burdensome. However, there are several possible explanations for the lower response rate in the web mode. First, while the paper group received an invitation letter and a 20-page questionnaire, the web group received only a one-page invitation letter, which is easier to overlook. Second, the paper questionnaire may have served as a visible reminder to complete the survey in a way that the one-page letter did not. Third, we have anecdotal evidence from our pretest that some respondents had trouble entering the survey link in their web browser. Therefore, contact persons in the web group may have failed to participate because they could not access the web survey, a challenge that the contact persons in the paper group did not have to overcome. Response burden may not be the driving issue for lower response rates. However, researchers planning to use the web mode for their establishment survey should remember that administrating a web survey comes at a cost of lower response rates.

Finally, we need to consider a number of important limitations:

First, the generalizability of our results is limited to surveys with similar lengths and formats. Our results are especially limited to our web survey design. Web surveys with a different design may perform differently as they may have functions or design characteristics that impact the response burden. We designed our web survey to be visually very similar to the paper questionnaire. However, important design features may reduce

respondents' burden in a web survey. Further research might explore how to reduce the response burden in the web mode.

Second, participation in the surveys used in this study was not mandatory. However, for a large proportion of establishment surveys, participation is required by law. Voluntary establishment surveys are likely to exclude establishments that are not motivated to respond or that anticipate a high response burden. Unfortunately, we can only speculate about the relationship between the anticipated response burden, the response rate and our mode groups. Inviting establishments to participate a web survey may exclude respondents who are not very savvy in using digital technologies and therefore decide not to participate. Establishments in the paper group may have cross-read the questionnaire or even started to respond but decided to not respond. As the choice between several modes is likely to overwhelm respondents to not respond at all (Medway and Fulton 2012), the choice of mode in a mandatory survey may add a perceived burden to respond. In all three mode groups, we may have found a higher response burden if the establishment survey would have been mandatory.

Third, there may be establishments with no internet access or with internal security guidelines that block web surveys or render them poorly (Harrel et al. 2007). Furthermore, establishments may have problems logging in, finding the website or navigating the survey (Bremner 2011; Gregory and Earp 2007). Therefore, our web respondent sample may be biased by an unknown coverage error.

Fourth, offering a paper, web or paper/web survey may recruit different kinds of respondents. Therefore, our respondent sample may be biased by mode-introduced nonresponse not visible in the data. However, the fact that the differences between paper and the web in the *Only groups* and paper and the web in the *Choice groups* are similar and the fact that our findings are consistent over two questionnaire versions makes us somewhat confident in the validity and reliability of our results.

Fifth, our results only consider German establishments. The results may change in establishment populations with higher or lower digitalization rates or with higher or lower internet penetration rates. Furthermore, we can link our results only to establishments that finished the survey but not to all invited establishments.

Sixth, we only consider the effect of the paper and web modes and not any other mode; instrument design; or interaction between instrument designs, respondent characteristics and establishment structures such as size. Especially in relation to the mode, instrument design decisions, respondents' characteristics and establishment structures can interact with each other. As we know from surveys of individuals, younger, more affluent, and higher educated respondents prefer the web mode over the paper mode (Kaplowitz et al. 2004; Kwak and Radler 2002; Messer and Dillman 2010; Millar et al. 2009). Similar effects may occur in establishment surveys. Our sample does not allow testing for these interactions as the number of cases is insufficient. The interaction of the web mode with other survey properties, respondent characteristics and establishment structures should be evaluated in future research.

Seventh, independent of the mode, first-time respondents must become familiar with the survey instrument. Against this background, respondents will develop individual best practices on how to interact with the survey instrument, that is, they improve when responding to a mode each time they participate. Therefore, we may see a change in the response burden

over time. Future research should assess whether panel participation affects the response burden and whether the response burden decreases or increases over time in the web mode.

Eighth, we used a postal letter as the mode of contact to invite establishments in each mode group to participate. Using a different means for contact, for example, email, may affect respondents' perceived burden. To access a web survey, respondents usually use a link. If the link is provided within an email, respondents only need to click on that link to access the web survey. If the link is provided on a paper invitation letter, respondents should type the link into their browser search bar to access the web survey, which takes more effort than just clicking on a link. Therefore, in terms of the response burden, contacting establishments with postal letters may increase the burden.

Although these limitations seem numerous, our results provide important insights into the effect of the web mode on the response burden in establishment surveys. Moreover, we are convinced that the validity of our findings is very high due to our rigorous experimental manipulations. In addition, our findings are consistent across two different surveys, which increases the reliability of our results. Our study provides important findings for the development and design of establishment surveys in the online era. Even if the perceived response burden (for respondents) is not lower in the web mode, web surveys are cost effective and enable features that help to improve data quality. Our findings about response burden, combined with the lack of difference in the response rates between the Paper-only and the Choice conditions, lead us to recommend that surveys should offer establishments a choice of paper and web modes.

7. Appendix

Table 4. Wording and response options for the response burden indicators.

Dimension	Indicator	Question	Response options
Perceived burden	Perception of time	Did you find it quick or time consuming to fill in the questionnaire?	Very quick, Quite quick, Neither quick nor time consuming, Quite time consuming, Very time consuming
	Perception of burden	Did you find it easy or burdensome to fill in the questionnaire?	Very easy, Quite easy, Neither easy nor burdensome, Quite burdensome, Very burdensome
Actual burden	Time to complete **(if 1 + persons filled out the questionnaire)**	How much time did you spend on actually filling in the questionnaire **(sections)**?	Number of hours, Number of minutes

Table 5. Joint $\chi 2$ values from margin contrast for minimum wage and digitalization questionnaire versions and hypotheses 1–3.

	Minimum wage	**Digitalization**
H1: lower estimated time for web respondents		
Perceived time	$\chi^2_{4,\ N=409} = 3.2,\ p = 0.53$	$\chi^2_{4,\ N=298} = 0.1,\ p = 1.0$
Perceived burden	$\chi^2_{4,\ N=409} = 0.7,\ p = 0.96$	$\chi^2_{4,\ N=295} = 0.5,\ p = 1.0$
H2.1: lower estimated time for choice-paper respondents		
Perceived time	$\chi^2_{4,\ N=655} = 2.5,\ p = 0.64$	$\chi^2_{4,\ N=461} = 0.1,\ p = 1.0$
Perceived burden	$\chi^2_{4,\ N=652} = 0.7,\ p = 0.94$	$\chi^2_{4,\ N=457} = 0.1,\ p = 1.0$
H2.2: lower estimated time for choice-web respondents		
Perceived time	$\chi^2_{4,\ N=305} = 1.2,\ p = 0.88$	$\chi^2_{4,\ N=234} = 0.0,\ p = 1.0$
Perceived burden	$\chi^2_{4,\ N=304} = 1.5,\ p = 0.83$	$\chi^2_{4,\ N=234} = 1.0,\ p = 0.90$
H3: lower estimated time for choice-web respondents		
Perceived time	$\chi^2_{4,\ N=551} = 0.66,\ p = 0.96$	$\chi^2_{4,\ N=397} = 0.7,\ p = 0.95$
Perceived burden	$\chi^2_{4,\ N=547} = 0.0,\ p = 1.0$	$\chi^2_{4,\ N=396} = 0.6,\ p = 0.96$

$^{+}p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, and *** $p \leq 0.001$

8. References

AAPOR. 2016. *Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys*. Available at: http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed March 2019).

Bavdaž, M., D. Giesen, S.K. Černe, T. Löfgren, and V. Raymond-Blaess. 2015. "Response burden in official business surveys: Measurement and reduction practices of national statistical institutes." *Journal of Official Statistics* 31: 559–588. https://doi.org/10.1515/jos-2015-0035.

Berglund, F., G. Haraldsen, and Ø. Kleven. 2013. "Causes and consequences of actual and perceived response burden based on Norwegian data." In *Comparative report on integration of case study results related to reduction of response burden and motivation of business*, edited by D. Giesen, M. Bavdaž, and I. Bolko: 29–35.

Bradburn, N.M. 1978. "Respondent burden." In Proceedings of the Section on Survey Research Methods Section: American Statistical Association, August 14–17, 1978. 35–40. San Diego, USA. American Statistical Association: 35–40. Available at: http://www.asasrms.org/Proceedings/y1978f.html (accessed March 2021).

Bremner, C. 2011. "An investigation into the use of mixed mode data collection methods for UK business surveys." In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž. March 22–23, (pp. 217–220). Heerlen, The Netherlands: Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/23FD3DF5-6696-4A04-B8EF-1FAACE AD995C/0/2011proceedingsblueets.pdf (accessed March 2019).

Cameron, A.C., and P.K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Conrad, F.G., M.F. Schober, and T. Coiner. 2007. "Bringing Features of Human Dialogue to Web Surveys." *Applied Cognitive Psychology* 21: 165–187. DOI: https://doi.org/10.1002/acp. 1335.

Couper, M.P. 2008. *Designing Effective Web Surveys*. Cambridge: Cambridge University Press.

Couper, M.P., and R.M. Groves. 1996. "Household-level determinants of survey nonresponse." *New Directions for Evaluation* 70: 63–79. DOI: https://doi.org/10.1002/ev.1035.

Dale, T., J. Erikson, J. Fosen, G. Haraldsen, J. Jones, and O. Kleven. 2007. *Handbook for monitoring and evaluating business survey response burdens*. Luxemburg: Eurostat. Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/12-HANDBOOK-FOR-MONITORING-AND-EVALUATING-BUSINESS-SURVEY-RESONSE-BURDEN.pdf/600e3c6d-8e8d-44f7-a8f5-0931c71d9920 (accessed October 2021).

Destatis. 2008. *Klassifikation der Wirtschaftszweige*. Wiesbaden, Germany: Statistisches Bundesamt. Available at: https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/klassifikation-wz-2008.html (accessed June 2019).

Downey, K., D. Mccarthy, and W. Mccarthy. 2007. "Encouraging the Use of Alternative Modes of Electronic Data Collection: Results of Two Field Studies." In Proceedings of the Third International Conference on Establishment Surveys, June 18–21, 2007. Montréal, Canada: 517–524. Available at: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000168.PDF (accessed September 2019).

Erikson, J. 2007. "Effects of offering web questionnaires as an option in enterprise surveys." In Proceedings of the Third International Conference on Establishment Surveys, June 18–21, 2007. Montréal, Canada: 1431–1435. Available at: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000168.PDF (accessed September 2019)

European Commission. 2011. *European Statistics Code of Practice for the National and Community Statistical Authorities.* Adopted by the European Statistical System Committee, September 28, 2011. Available at: https://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15 (accessed September 2019).

Gelman, A. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22(2) : 153–164. DOI: https://doi.org/10.1214/088342306000000691.

Giesen, D. 2007. "Does mode matter? Comparing response burden and data quality of paper and an electronic business questionnaire." In Proceedings of the 6th Conference on Questionnaire Evaluation Standards (QUEST), April 24–26, 2007. Ottawa, Canada: 150–161. Available at: https://wwwn.cdc.gov/QBank/QUEST/2007/QUEST%202007%20Proceedings-all%20papers.pdf (accessed September 2019).

Giesen, D. 2011. "Burden reduction by communication." In *Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes*, edited by D. Giesen: 33–42.

Giesen, D. 2012. "Exploring Causes and Effects of Perceived Response Burden." In Proceedings of the Fourth International Conference on Establishment Surveys, June

11–14, 2012. Montréal, Canada. Available at: https://ww2.amstat.org/meetings/ices/2012/papers/302171.pdf (accessed September 2019).

Giesen, D. 2013a. "Causes and Consequences of Actual and Perceived Response Burden Based on Dutch Data." In *Comparative report on integration of case study results related to reduction of response burden and motivation of business*, edited by D. Giesen, M. Bavdaž, and I. Bolko: 33–39.

Giesen, D. 2013b. "Reducing Response Burden by Questionnaire Redesign." In *Comparative report on integration of case study results related to reduction of response burden and motivation of business*, edited by D. Giesen, M. Bavdaž, and I. Bolko: 63–68.

Giesen, D., M. Bavdaž, and G. Haraldsen. 2011. "Response burden measurement: Current diversity and proposal for moving towards standardisation." In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž. March 22–23: 125–134. Heerlen, The Netherlands: Statistics Netherlands. Available at: https://www.cbs.nl/-/media/imported/documents/2011/14/2011-4-4-4-2-giesen-et-al-presentation-blue-ets-2011.pdf (accessed March 2021).

Giesen, D., and J. Burger. 2013. "Measuring and understanding response quality in the Structural Business Survey questionnaires." In Proceedings of the European Establishment Statistics Workshop, September 9–11, 2013. Nuremberg, Germany. Available at: http://doku.iab.de/fdz/events/2013/Session5%20Giesen.pdf (accessed September 2019).

Giesen, D., M. Morren, and G. Snijkers. 2009. "The effect of survey redesign on response burden: An evaluation of the redesign of the SBS questionnaires." Paper presented at the European Survey Research Association Conference 2009, June 29–July 3, 2009. Warsaw, Poland. European Survey Research Association.

Giesen, D., M. Vella, and C. Brady. 2018. "Response Burden Management for Establishment Surveys at Four National Statistical Institutes." *Journal of Official Statistics* 34(2): 397–418. DOI: https://doi.org/10.2478/jos-2018-0018.

Groves, R.M., R.B. Cialdini, and M.P. Couper. 1992. "Understanding the decision to participate in a survey." *Public Opinion Quarterly* 56: 475–495. DOI: https://doi.org/10.1086/269338.

Gregory, G., and M. Earp. 2007. "Evolution of Web at USDA' National Agricultural Statistics Service." In Proceedings of the Third International Conference on Establishment Surveys (ICES-III), June 18–21, 2007. Montréal, Canada: 1442–1445. Available at: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000192.PDF (accessed September 2019).

Gravem, D. 2011. "Response burden trends and consequences." In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž. March 22–23: 221–236. Heerlen, The Netherlands: Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/23FD3DF5-6696-4A04-B8EF-1FAACEAD995C/0/2011proceedingsblueets.pdf (accessed March 2019).

Haas, G.-C., S. Eckman, R. Bach, and F. Kreuter. 2016. "Is Moving Establishment Surveys from Mail to Web a Good or Bad Decision in Terms of Performance and Data Quality?" In Proceedings of the International Conference for Establishment Surveys

2016 (ICES-V). June 21–23, 2016. Geneva, Switzerland. Available at: https://ww2.amstat.org/meetings/ices/2016/proceedings/ICESV_TOC.pdf (accessed February 2020).

Haraldsen, G., and J. Jones. 2007. "Paper and Web Questionnaires Seen from the Business Respondent's Perspective." In Proceedings of the Third International Conference on Establishment Surveys (ICES-III), June 18–21, 2007, Montreal, Canada: 1040–1047. Available at: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000259.PDF (accessed September 2019).

Haraldsen, G., J. Jones, D. Giesen, and L.C. Zhang. 2013. "Understanding and coping with response burden." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, G. Haraldsen, J. Jones, and D. Willimack: 219–252. Hoboken, NJ: John Wiley and Sons.

Harrell, L., H. Yu, and R. Rosen. 2007. "Respondent acceptance of web and E-mail data reporting for an establishment survey." In Proceedings of the Third International Conference on Establishment Surveys (ICES-III), June 18–21, 2007, Montreal, Canada: 1442–1445. Available at: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000230.PDF (accessed September 2019).

Hedlin, D., T. Dale, G. Haraldsen, and J. Jones. 2005. *Developing Methods for Assessing Perceived Response Burden*. Luxembourg: Eurostat. Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/10-DEVELOPING-METHODS-FOR-ASSESSING-PERCEIVED-RESPONSE-BURDEN.pdf/1900efc8-1a07-4482-b3c9-be88ee71df3b (accessed October 2021).

Jones, J. 2012. "Response Burden: Introductory Overview Lecture." In Proceedings of the Fourth International Conference on Establishment Surveys, June 11–14, 2012. Montréal, Canada. Available at: http://www.amstat.org/meetings/ices/2012/papers/302289.pdf (accessed September 2019).

Jones, J., G. Snijkers, and G. Haraldsen. 2013. "Surveys and Business Surveys." In *Designing and Conducting Business Surveys*, edited by G. Snijkers, G. Haraldsen, J. Jones, and D.K. Willimack: 1–33. Hoboken, NJ: John Wiley & Sons.

Kaplowitz, M.D., T.D. Hadlock, and R. Levine. 2004. "A Comparison of Web and Mail Survey Response Rates." *Public Opinion Quarterly* 68: 94–101. DOI: https://doi.org/10.1093/poq/nfh006.

Krosnick, J.A. 1991. "Response strategies for coping with the cognitive demands of attitude measures in surveys." *Applied Cognitive Psychology* 5: 213–236. DOI: https://doi.org/10.1002/acp. 2350050305.

Kwak, N., and B. Radler. 2002. "A Comparison between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality." *Journal of Official Statistics* 18: 257–273. DOI: https://doi.org/10.1177/1525822X08317085.

Lagerstøm, B. 2018. "Chatbots as digital interviewers." Paper presented at the International Household Nonresponse Workshop, August 22–24, 2018. Budapest, Hungary

Löfgren, T. 2011. "Burden reduction by instrument design." In *Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes*, edited by D. Giesen: 43–50.

Lyly-Yrjänäinen, M., and G. van Houten. 2011. "Reduce burden, increase motivation. Main findings of the quality assessment of the second European company survey." In

Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys March 22–23, 2011, Heerlen, Netherlands, edited by D. Giesen and M. Bavdaž: 107–118. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiFvfTm2OrzAhU8BWMBHYOlDv8QFnoECAgQAQ&url=https%3A%2F%2Fwww.cbs.nl%2F-%2Fmedia%2Fimported%2Fdocuments%2F2011%2F25%2F2011-06-20-lyly-yrjanainen-van-houten-2011-reduce-burden-increase-motivation.pdf&usg=AOvVaw0n4dUPG4Ectn1K-baV7hE0 (accessed October 2021).

Medway, R.L., and J. Fulton. 2012. "When More Gets You Less: A Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates." *Public Opinion Quarterly* 76: 733–746. DOI: https://doi.org/10.1093/poq/nfs047.

Messer, B.L., and D.A. Dillman. 2010. Using Address Based Sampling to Survey the General Public by Mail vs. "Web plus Mail". Technical Report 10–13. *Social and Economic Sciences Research Center*, Washington State University, Pullman. Available at: http://www.sesrc.wsu.edu/dillman/papers/2010/Messer%20Dillman%20WCSTechReport.pdf (accessed October 2021).

Millar, M.M., O.A.C. Neill, and D.A. Dillman. 2009. "Are Mode Preferences Real?" *Technical Report of the Social and Economic Sciences Research Center.* Pullman, Washington: Washington State University

Sear, J. 2011. "Response burden measurement and motivation at Statistics Canada." In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys* Heerlen, March 22–23, 2011, Heerlen, Netherlands. edited by D. Giesen and M. Bavdaž: 151–160. Available at: https://www.cbs.nl/-/media/imported/-documents/2011/25/2011-06-20-sear-2011-response-burden-measurement-and-motivation-at-statistics-canada.pdf (accessed October 2021).

Snedecor, G.W., and W.G. Cochran. 1989. *Statistical Methods*, (8th edition). Ames, IA: Iowa State University Press.

Snijkers, G., G. Haraldsen, A. Sundvoll, T. Vik, and H.P. Stax. 2011. "Utilizing web technology in business data collection: Some Norwegian, Dutch and Danish experiences." In *Proceeding of the European Conference on New Techniques and Technologies for Statistics (NTTS)*, Brussels, Belgium

Snijkers, G., E. Onat, and R. Vis-Visschers. 2007. "The Annual Structural Business Survey: Developing and Testing an Electronic Form." In Proceedings of the International Conference for Establishment Surveys 2007 (ICES-III) June 18–21, 2007, Montreal, Canada: 456–463. Available at: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000095.PDF (accessed October 2021).

Verkruyssen, F., and S. Moens. 2011. "Communication as a tool to reduce perceived response burden: Tips and tricks." In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, March 22–23, 2011, Heerlen, Netherlands edited by D. Giesen and M. Bavdaž. 237–242. Available at: https://www.cbs.nl/-/media/imported/documents/2011/14/2011-4-4-8-1-verkruyssen-moens-presentation-blue-ets-2011.ashx?la=en-gb (accessed October 2021).

Willimack, D., and E. Nichols. 2010. "A Hybrid Response Process Model for Business Surveys." *Journal of Official Statistics* 26: 3–24. Available at: https://www.scb.se/-

contentassets/ca21efb41fee47d293bbee5bf7be7fb3/a-hybrid-response-process-model-for-business-surveys.pdf (accessed September 2021).

Van Loon, A.J.M., M. Tijhuis, H.S. Picavet, P.G. Surtees, and J. Ormel. 2003. "Survey Non-response in the Netherlands: Effects on Prevalence Estimates and Associations." *Ann Epidemiol* 13: 105–110. DOI: https://doi.org/10.1016/S1047-2797(02)00257-0.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response Burden: What Is It and What Predicts It?" In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P.C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G.B. Willis, and A. Wilmot: 193–212.

Received September 2019
Revised February 2020
Accepted March 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 931–953, http://dx.doi.org/10.2478/JOS-2021-0040

Trends in Establishment Survey Nonresponse Rates and Nonresponse Bias: Evidence from the 2001-2017 IAB Establishment Panel

Corinna König[1], *Joseph W. Sakshaug*[1], *Jens Stegmaier*[1], *and Susanne Kohaut*[1]

Evidence from the household survey literature shows a declining response rate trend in recent decades, but whether a similar trend exists for voluntary establishment surveys is an understudied issue. This article examines trends in nonresponse rates and nonresponse bias over a period of 17 years in the annual cross-sectional refreshment samples of the IAB Establishment Panel in Germany. In addition, rich administrative data about the establishment and employee composition are used to examine changes in nonresponse bias and its two main components, refusal and noncontact, over time. Our findings show that response rates dropped by nearly a third: from 50.2% in 2001 to 34.5% in 2017. Simultaneously, nonresponse bias increased over this period, which was mainly driven by increasing refusal bias whereas noncontact bias fluctuated relatively evenly over the same period. Nonresponse biases for individual establishment and employee characteristics did not show a distinct pattern over time with few exceptions. Notably, larger establishments participated less frequently than smaller establishments over the entire period. This implies that survey organizations may need to put more effort into recruiting larger establishments to counteract nonresponse bias.

Key words: Survey participation; establishment characteristics; administrative data; unit nonresponse.

1. Introduction

Establishment surveys are indispensable tools for investigating economic relationships and providing up-to-date information about the labor market. Collecting information about the labor market while gaining deeper insights into the economic status of establishments and employee conditions are major goals of these surveys. By measuring a variety of topics, such as employment development, investments, and vocational education, economic research can be enhanced and new correlations and developments observed. This feeds into academic and non-academic discussions of the economic climate and informs important policy decisions. Although many establishment surveys are mandatory whereby participation is required by law, voluntary establishment surveys continue to play a key role in informing official statistics and policy development.

There are several prominent examples of voluntary establishment surveys. One example is the European Company Survey (ECS) (Eurofound 2015). The ECS provides European- and country-specific information on work organization, human resource management, and

[1] Institute for Employment Research, 104 Regensburger Straße, Nuremberg 90478, Germany. Emails: corinna.koenig@iab.de, joe.sakshaug@iab.de, jens.stegmaier@iab.de, and susanne.kohaut@iab.de

workplace innovations. Since 2004, telephone interviews are conducted every four to five years in up to 32 countries with establishments of all industries. The results contribute to policy discussions at both the employer- and employee-level as representatives of both levels are interviewed. In the United States, a large producer of establishment surveys is the Bureau of Labor Statistics (BLS), which conducts numerous voluntary and mandatory surveys that vary in their frequency (e.g., monthly, quarterly, annual) and design. The largest voluntary BLS survey is the Current Employment Statistics survey, which interviews about 145,000 businesses and government agencies monthly. The survey produces information on non-farm employment, hours, and earnings for employees in each state, which is used to generate monthly payroll estimates (Mullins 2016). A further example of a large, voluntary establishment survey – and the focus of the present study – is the IAB Establishment Panel in Germany, conducted by the Institute for Employment Research (IAB). About 16,000 establishments are interviewed annually via face-to-face for the purpose of studying the demand side of the labor market and collecting information about the establishment structure, as well as financial characteristics and employee attributes. Furthermore, the survey captures challenges and future assessments of establishments to inform policy debates on measures that facilitate economic growth (Ellguth et al. 2013). Collective bargaining coverage is another important topic of the survey, which is repeatedly discussed by politicians in Germany (Ellguth and Kohaut 2019).

Like all surveys, one of the largest threats to establishment survey data quality is unit nonresponse. For establishments, survey participation is largely a business decision. Unlike social surveys that might appeal to households for intrinsic, altruistic, or topical reasons, people in the establishments must evaluate whether they have the authority, capacity, and motivation to participate in a voluntary survey, which takes resources away from their primary business objectives (Tomaskovic-Devey et al. 1995; Willimack et al. 2002; Willimack and Snijkers 2013). This decision might be affected by the organizational structure and available staffing, which are also key characteristics that surveys attempt to measure. Thus, there is reason to believe that nonresponse is a non-random process and, as a consequence, may introduce nonresponse bias in establishment survey estimates.

Although response rates for household surveys have declined in recent decades (e.g., Luiten et al. 2020; Beullens et al. 2018; Brick and Williams 2013; De Leeuw and De Heer 2002; Groves and Couper 1998), response rate trends in voluntary establishment surveys and associated estimates of nonresponse bias are largely understudied. In this article, we investigated nonresponse trends in the IAB Establishment Panel over 17 years. The Panel is a unique data source for studying nonresponse trends over time as large cross-sectional refreshment samples are drawn each year to replenish the panel. In addition, the Panel can be directly linked to establishment-level characteristics derived from rich administrative data, which we exploited to study nonresponse bias trends and obtain a better understanding of the characteristics that correlate with nonresponse.

The remainder of the article is structured as follows. In Section 2, we briefly review theoretical frameworks of establishment survey participation, summarize the existing literature on response rates and correlates of participation, and present the research questions. In Section 3, the survey and administrative data sources used to study nonresponse are described. Section 4 details the analysis procedures and Section 5 presents outcome rates, nonresponse biases, and results from regression models of survey

participation for each of the 17 years. In Section 6, the main findings of the study are summarized and their implications for survey practice are discussed.

2. Background

2.1. Theoretical Frameworks of Establishment Survey Participation

In general, answering a questionnaire of a voluntary survey is a work task that does not contribute to the establishment's primary goal of maximizing revenue (Sudman et al. 2000). Hence, to obtain cooperation from establishments, the actual and perceived response burden should be sufficiently low. Actual response burden corresponds to the costs incurred by the establishment while responding to the survey, which are never completely removed. In this context, the expectations regarding the response process are important. Especially for establishments that are recruited for the first-time, participation implies a high cognitive burden and a large time expenditure as they have no prior knowledge about the survey process or questionnaire.

Willimack et al. (2002) developed a framework that classifies individual factors associated with the participation decision into two groups according to the survey organization's ability to influence them. The first group includes factors that are out of the survey organization's control: the external environment, the establishment, and the respondent delegated the response task. The second group of factors are under the control of the survey organization and mainly concern the survey design. Willimack et al. (2002) also noted the three key features of authority, capacity, and motivation to respond, as originally defined by Tomaskovic-Devey et al. (1995). Authority refers to both the formal and informal authorization to decide whether to participate. Capacity refers to the ability of the respondent to successfully complete the questionnaire in terms of cognition, time, and data access. Lastly, the motivation to respond is related to the willingness of the respondent to undertake the response task. A representative of an establishment and the establishment itself must possess these features in order to respond to a survey request.

An extension of the Willimack et al. (2002) model was proposed by Willimack and Snijkers (2013), which shows the causal order of the decision-making process. For example, the establishment's management may consider the external environment when deciding whether to participate, which happens before appointing the responding employee. The authors also argued that the dimensions of authority, capacity, and motivation to respond can be applied to each of the participation factors identified by Willimack et al. (2002).

2.2. Response Rate Trends in Establishment Surveys

When evaluating and comparing response rates of different establishment surveys, it is important to keep in mind that surveys differ in their design (e.g., mode of data collection, nonresponse follow-up), which can influence the response rate. In addition, mandatory surveys are expected to have higher response rates and different trends compared to voluntary ones (Petroni et al. 2004; Paxson et al. 1995). In the following review of response rate trends, we focus on voluntary surveys.

Christianson and Tortora (1995) interviewed 21 national statistical institutes in 16 countries to classify response rate trends for their establishment surveys and censuses as either increasing, decreasing, or unchanged over the past ten years of the respective studies. Information about 104 surveys was collected and anonymized, but no distinction was made between mandatory and voluntary studies. The authors found that most surveys and censuses (about 41%) observed no changes in response rates in the past ten years. In addition, similar proportions of surveys and censuses reported increasing (27%) and decreasing (26%) response rates, while the remainder (6%) provided no information. The authors mention possible reasons for increasing (or stable) response rates, including greater nonresponse follow-up efforts (e.g., use of reminders) and shortening of the questionnaire. Reasons for decreasing response rates included long questionnaires, sensitive questions about financial data, and diminishing survey budgets.

Among seven voluntary establishment surveys conducted by the BLS between 2010 and 2019, four showed decreasing response rate trends, while all others revealed no significant changes (https://www.bls.gov/osmr/response-rates/establishment-survey-response-rates.htm#BLStable_2020_4_27_14_11_footnotes). Among those with a declining trend, the average decline over the ten years was about 7% with a range between 4–10% (own calculations).

For three waves of the ECS, 14 out of 21 countries (e.g., Belgium, France, Germany, Hungary, Slovenia) observed increasing response rates between 2004 and 2013. The average increase was 20% with a range between 3–45% and dominated by increases in Hungary (45%), Slovenia (43%), and Luxembourg (33%). Five countries reported declining response rates over the same period, with a range between 8–25% and an average decline of 14%. Countries with the largest decline included Finland (25%) and Denmark (20%). The two remaining countries, Latvia and the Netherlands, showed no change in their response rates. These differences in response rate trends occurred despite efforts to standardize the survey design, as all countries used the same survey methods (e.g., use of advance letters/emails, and minimum number of contact attempts) whenever possible (Eurofound 2015).

In addition to descriptive response rate trends, a key question is whether such trends correlate with nonresponse bias, as there is no guarantee that decreasing response rates are accompanied by increasing nonresponse bias over time, or vice versa. As yet, extensive calculations of nonresponse bias in establishment surveys have been neglected. Borrowing from the social survey literature (Brick and Tourangeau 2017; Groves and Peytcheva 2008; Groves 2006), a rather weak correlation between nonresponse rates and (absolute relative) nonresponse bias across several surveys has been identified. However, it is unknown whether a similarly weak correlation exists for establishment surveys.

2.3. *Correlates of Establishment Survey Participation*

Several studies have identified correlates of participation in voluntary establishment surveys. We focus on establishment-level correlates. For a discussion of other correlates (e.g., interviewer characteristics, competing survey requests, statutory laws regarding vacation/working days), see Janik and Kohaut (2012), Janik (2011), Seiler (2010), and Davis and Pihama (2009).

Tomaskovic-Devey et al. (1995) analyzed businesses as part of the 1989 North Carolina Employment and Health Survey and found that larger establishments and establishments

in industries with higher average profits were less likely to respond. Likewise, Earp et al. (2018) ascertained with the 2012 Job Openings and Labor Turnover Survey of the BLS that small establishments with less than 50 employees respond at a higher rate in the first wave of a panel than larger establishments.

Phipps and Toth (2012) found in the BLS Occupational Employment Survey that the population size of the area where the establishment was located had a negative effect on the response rate. That is, establishments located in a city of one million or more inhabitants were less likely to respond. Additionally, they showed that large establishments had lower response likelihoods than smaller ones, and single-unit establishments were more likely to respond than multi-unit establishments. Overall, the lowest likelihood of response was observed for large multi-unit establishments in the information, finance, or professional/business services industries.

Janik (2011) and Janik and Kohaut (2012) investigated the correlation of establishment features and survey participation in the IAB Establishment Panel. Both studies reported that large establishments are less likely to participate, in line with the already-mentioned studies. Janik (2011) also found that establishments in East Germany are more likely to participate than those in West Germany.

While informative, the above studies are limited, in the sense that they analyze data of only one year or a short time period. What is missing from the literature is an extended analysis of temporal changes in establishment-level correlates over several years, which can inform whether compositional differences between respondents and nonrespondents are changing, and identify underrepresented groups that might require greater recruitment effort and/or targeted interventions going forward. We address this research gap by studying temporal changes in establishment-level correlates over 17 years of cross-sectional samples in the IAB Establishment Panel. In addition to establishment-level characteristics analyzed in previous studies, we also consider the demographic employee structure of the establishment (e.g., the share of women or employees with certain education levels). Such characteristics have not previously been examined in the nonresponse literature, but may provide a more detailed picture of the selectivity of establishment survey participation.

2.4. *Research Questions*

Although response rate trends are a simple, yet important descriptor of surveys, they are relatively understudied for voluntary establishment surveys, particularly in recent decades. In addition, a discussion of refusal and noncontact rates is missing and their consideration is relevant since overall response rate trends may be driven by differential trends in one or both factors (Groves and Couper 1998). We contribute to this research gap by answering the first research question:

(1) Have participation outcome rates (nonresponse, refusal, noncontact) in the IAB Establishment Panel's cross-sectional samples changed over time?

Even if overall response rate trends have changed over time, it reveals nothing about trends in the composition of the responding sample and the magnitude of potential nonresponse bias. Petroni et al. (2004, 15) emphasize that nonresponse bias may be a greater problem in establishment surveys than in household surveys as their "[. . .] underlying population is very skewed" in terms of their attributes. However, there is a lack of extended evaluations of

nonresponse bias in establishment surveys and their potentially dynamic behavior over time. Thus, it is unclear whether nonresponse biases are correlated with trends in the nonresponse rate. Furthermore, to the best of our knowledge, no study has separately analyzed refusal bias and noncontact bias trends for establishment surveys. Hence, the second research question is:

(2) How large are nonresponse biases in the IAB Establishment Panel's cross-sectional surveys and has their magnitude increased, decreased, or remained stable over time? Are nonresponse rate trends correlated with nonresponse bias?

Lastly, we analyze characteristics of establishments that may be related to survey participation and the potential changes in the magnitude of their relationship over time. Until now, almost all previous studies investigating correlates of establishment survey participation used only one year of data with a limited set of predictors. We extend this approach by using data for 17 years (from 2001 to 2017) to examine changes in associations, and consider a rich set of participation determinants, including general establishment characteristics and the employee structure of establishments. This leads to our third research question:

(3) To what extent are general establishment characteristics as well as employee characteristics of establishments associated with survey participation? Does the magnitude of these associations change over time?

3. Data

3.1. IAB Establishment Panel – Refreshment Samples

The IAB Establishment Panel is a voluntary annual longitudinal survey of establishments that gathers high quality data on labor demand in Germany. The questionnaire topics cover objective operational characteristics (e.g., employment indicators) as well as subjective assessments. The survey data influence government decisions at the federal and state-levels through consultation with the IAB and external researchers. The target population consists of every establishment in Germany with at least one employee who was liable for social security contributions on June 30 in the previous year.

Since 2001, approximately 16,000 establishments participated in the survey each year. The survey is composed of two samples. One sample consists of establishments who already participated in at least one of the last two waves and are approached for reinterview, and the second sample is a cross-sectional refreshment sample of establishments who are newly-recruited to join the panel (Fischer et al. 2008). We focus solely on the cross-sectional samples from 2001 to 2017. The panel is mainly carried out by face-to-face interviewing with a small proportion of data collection by mail until 2015 in two federal states (Ellguth et al. 2013). We excluded the mail cases and restricted the analysis to samples assigned to face-to-face interviewing only.

Across the 17 years, a total of 124,395 establishments were selected for the annual refreshment samples, an average of about 7,317 new establishments per year (range: 4,619 in 2002 to 9,812 in 2017). An advance letter was sent to all sampled establishments announcing the survey and the impending interviewer visit and included sponsorship letters by high-ranking authorities (Fischer et al. 2008). Interviews were sought with the owner or manager of the establishment. The survey organization, which pays its

interviewers per interview, determined how many contact attempts were made until a case was closed without an interview.

3.2. IAB Administrative Data

To evaluate nonresponse bias and identify correlates of participation, we utilized IAB administrative data. The data contain all establishments in Germany with at least one employee who was liable for social security contributions on June 30 in the previous year (Schmucker et al. 2018). The data also contain variables on numerous properties of the establishments and their workforce. For example, the number of employees by different education levels, age groups, and types of employment. Further information on this resource can be found in Schmucker et al. (2018). To use the administrative data, we performed a one-to-one linkage to the survey data, which is possible through a unique identifier. The current year of the IAB Establishment Panel is linked to the previous year of the administrative data since that is when the cross-sectional sample was drawn.

3.3. Variables of Interest

3.3.1. Participation Outcomes

Participation in the IAB Establishment Panel is defined as any establishment that completed the questionnaire with an interviewer in a face-to-face situation. If nonresponse occurred, interviewers were instructed to document reasons why a completed interview could not be obtained. Based on this paradata, the sample units were classified as respondents, refusals, and noncontacts. Online supplemental material Table S1 displays the categorization of the possible participation outcomes.

3.3.2. Administrative Establishment Variables

All administrative establishment variables used in the bias and regression analyses include the value observed at the time of sampling (i.e., approximately one year before the start of survey data collection). Table 1 shows the variables and their categories used in each analysis. The variables describing the establishments are summarized into two groups: general characteristics and employee structure. The general characteristics group contains the following variables: location, size, industry, year of foundation, change in the number of employees since the previous year, and the population size of the establishment's area. For the employee structure group, the variables include: shares of female employees, German employees, average age of employees, low-qualified employees, middle-qualified employees, and high-qualified employees. These variables were chosen based on their usage in previous substantive and methodological research on establishments, as well as their likely association with the survey topics (Sakshaug et al. 2019b; Brixy et al. 2007; Henze 2014; Wagner 2012), which make them suitable proxy indicators of nonresponse bias in the actual survey variables. Descriptive statistics for all variables are presented in online supplemental material Table S2. Interviewer characteristics were not analyzed as they are unavailable for the entire observation period from 2001 to 2017.

All variables were categorized to facilitate interpretation. In most cases, the categorization was performed arbitrarily based on uniform allocation of units into

Table 1. Administrative variables and categories used in the bias and regression analyses.

Variable	Categories	
	Bias analysis	Regression analysis
General characteristics		
Location	(0) East Germany (1) West Germany	(1) East Germany (REF) (2) South Germany (3) North Germany (4) West Germany
Establishment size (number of employees)	(1) 1–9 (2) 10–49 (3) 50+	(1) 1–4 (REF) (2) 5–9 (3) 10–19 (4) 20–49 (5) 50–99 (6) 100–199 (7) 200–499 (8) 500–999 (9) 1,000+
Industry	(1) Agriculture/production (2) Service (3) Public/educ/health/arts	(1) Agriculture/mining/energy/water (2) Manufacturing industry (REF) (3) Construction industry (4) Trade/repair (5) Transport/communication (6) Financial intermediation (7) Services mainly for companies (8) Other services (9) Public sector
Year of foundation	(1) 1970s/1980s (2) 1990s (3) 2000s (4) Unknown	(1) 1970s/1980s (2) 1990s (3) 2000s (REF) (4) Unknown
Change in the no. of employees since previous year	-	(1) Decrease (2) No change (REF) (3) Increase (4) Unknown
Area population size(number of inhabitants)	-	(1) $<$2,000 (2) 2,000–4,999 (3) 5,000–19,999 (4) 20,000–49,999 (5) 50,000–99,999 (6) 100,00–499,999 (7) $>$500,000 (REF)
Employee structure		
Pct. of female employees	(1) 0 – $\leq$15 (2) $>$15 – $\leq$45 (3) $>$45 – $\leq$75	(1) 0 – $\leq$15 (REF) (2) $>$15 – $\leq$45 (3) $>$45 – $\leq$75

Table 1. Continued

Variable	Categories	
	Bias analysis	Regression analysis
	(4) >75 – ≤100	(4) >75 – ≤100
Pct. of German employees	(1) 0 – <100 (2) 100	(1) 0 – <100 (REF) (2) 100
Average age of employees (years)	(1) 10.5 – ≤36 (2) >36 – ≤41 (3) >41 – ≤45 (4) >45 – ≤88	(1) 10.5 – ≤36 (REF) (2) >36 – ≤41 (3) >41 – ≤45 (4) >45 – ≤88
Pct. of low-qualified employees	(1) 0 (2) >0 – ≤100	(1) 0 (REF) (2) >0 – ≤100
Pct. of middle-qualified employees	(1) 0 – ≤50 (2) >50 – ≤75 (3) >75 – ≤90 (4) >90 – ≤100	(1) 0 – ≤50 (REF) (2) >50 – ≤75 (3) >75 – ≤90 (4) >90 – ≤100
Pct. of high-qualified employees	(1) 0 (2) >0 – ≤8 (3) >8 – ≤21 (4) >21 – ≤100	(1) 0 (REF) (2) >0 – ≤8 (3) >8 – ≤21 (4) >2 – ≤100

Note: (REF) specifies the reference category for every variable in the logistic regression.

approximately equal-sized groups, or inspection of the original distributions for natural cut-off values with sufficient cell sizes. Three variables (establishment size, location, industry) were categorized slightly differently depending on the type of analysis: bias or regression. Finer categorization was adopted for the regression models. For instance, establishment size, measured by the number of employees, was specified using three categories in the bias analysis (1–9, 10–49, and 50+ employees) and nine categories (from 1–4 to 1000+ employees) in the regression analysis.

Two additional general characteristics motivated by the literature were considered only in the regression analysis: change in the number of employees since the previous year and area population size (defined as the number of inhabitants in the city or metropolitan area where the establishment resides). The first variable was used as a proxy for the general economic situation by classifying the change in the number of employees from the previous to the current year in three categories: decrease, increase, or no change (Janik 2011). Some establishments had no information from the previous year and were allocated to a missing data category. The use of the second variable was motivated by Phipps and Toth (2012), who identified a correlation between area population size and participation in the aforementioned BLS Occupational Employment Survey.

4. Methods

4.1. Outcome Rate Definitions

Response and refusal rates were defined by Response Rate 1 and Refusal Rate 1 of the American Association for Public Opinion Research, respectively (AAPOR 2016). The

noncontact rate was also calculated and based on the same denominator (i.e., all sampled units). By definition, the sum of the refusal and noncontact rates is equal to the total nonresponse rate. By calculating the rates for each year of the IAB Establishment Panel, we analyzed the change in nonresponse over the 17 cross-sections. We note that each cross-section, by definition, excludes existing panel members who are likely to be more cooperative than the general establishment population. Thus, the response rates reported later might be considered as an upper bound compared to a repeated cross-sectional survey without a longitudinal component.

$$Response\ rate_{year} = \frac{respondents_{year}}{sample_{year}} \tag{1}$$

$$Refusal\ rate_{year} = \frac{refusals_{year}}{sample_{year}} \tag{2}$$

$$Noncontact\ rate_{year} = \frac{noncontacts_{year}}{sample_{year}} \tag{3}$$

4.2. Calculation of Nonresponse Biases

All three types of nonresponse bias (total, refusal, noncontact) were calculated by comparing the estimated percentages of each variable category based on the respondents (or contacts) to the estimate based on the full sample (or contacts) (D'Aurizio and Papadia 2019). For example, nonresponse bias for variable Y was calculated as the difference between the estimated percentage of respondents r belonging to variable category i in year y: $\bar{Y}_{r,i,y}$ and the corresponding percentage estimated for the total sample n: $\bar{Y}_{n,i,y}$. Similarly, for refusal bias, the respondent-based estimate was compared to the estimate based on the contacted cases c, and for noncontact bias, the estimates derived from the contacts and full sample were compared:

$$Nonresponse\ bias_y = \bar{Y}_{r,i,y} - \bar{Y}_{n,i,y} \tag{4}$$

$$Refusal\ bias_y = \bar{Y}_{r,i,y} - \bar{Y}_{c,i,y} \tag{5}$$

$$Noncontact\ bias_y = \bar{Y}_{c,i,y} - \bar{Y}_{n,i,y} \tag{6}$$

Another way of estimating bias is in relative terms (Sakshaug et al. 2019a; Sakshaug and Huber 2016; Groves 2006). Here, we adopted a measure of absolute relative bias, which assesses the magnitude of the bias relative to its reference estimate:

$$Absolute\ relative\ nonresponse\ bias_y = \left|\frac{\bar{Y}_{r,i,y} - \bar{Y}_{n,i,y}}{\bar{Y}_{n,i,y}}\right| \tag{7}$$

$$Absolute\ relative\ refusal\ bias_y = \left|\frac{\bar{Y}_{r,i,y} - \bar{Y}_{c,i,y}}{\bar{Y}_{c,i,y}}\right| \tag{8}$$

$$Absolute\ relative\ noncontact\ bias_y = \left|\frac{\bar{Y}_{c,i,y} - \bar{Y}_{n,i,y}}{\bar{Y}_{n,i,y}}\right| \tag{9}$$

To aid in pointing out particularly large biases, we adopted a subjective cut-off value of 10% absolute relative bias to define individual biases that might be considered "substantively

meaningful" (Sakshaug et al. 2019a). However, we acknowledge that such a cut-off is arbitrary and others are likely to have differing opinions regarding such a threshold.

To summarize the results across variable categories and variable groups, the average absolute relative bias is presented. This measure was calculated as the average of the absolute relative bias estimates across all K categories of a relevant variable group:

$$Average\ absolute\ relative\ nonresponse\ bias_y = \frac{\sum_{i=1}^{K} \left| \frac{\bar{Y}_{r,i,y} - \bar{Y}_{n,i,y}}{\bar{Y}_{n,i,y}} \right|}{K} \quad (10)$$

$$Average\ absolute\ relative\ refusal\ bias_y = \frac{\sum_{i=1}^{K} \left| \frac{\bar{Y}_{r,i,y} - \bar{Y}_{c,i,y}}{\bar{Y}_{c,i,y}} \right|}{K} \quad (11)$$

$$Average\ absolute\ relative\ noncontact\ bias_y = \frac{\sum_{i=1}^{K} \left| \frac{\bar{Y}_{c,i,y} - \bar{Y}_{n,i,y}}{\bar{Y}_{n,i,y}} \right|}{K} \quad (12)$$

4.3. *Modeling Survey Participation*

To model survey participation (1 = response; 0 = nonresponse), we run separate logistic regression models on the yearly cross-sectional samples over the entire observation period (i.e., 17 regressions) with the covariates shown in Table 1.

All analyses (outcome rates, nonresponse bias estimation, and regression modeling) were weighted to account for probabilities of selection as establishments with certain characteristics (e.g., larger establishments) were routinely oversampled in each cross-sectional sample. The analyses were performed using the "survey" commands in Stata 15 (StataCorp 2017).

5. Results

5.1. *Outcome Rate Trends*

Figure 1 shows the response rate, refusal rate, and noncontact rate for each of the corresponding years. Changes in the outcome rates were observed over the 17-year observation period. The response rate reduced by nearly a third from 50.2% (2001) to

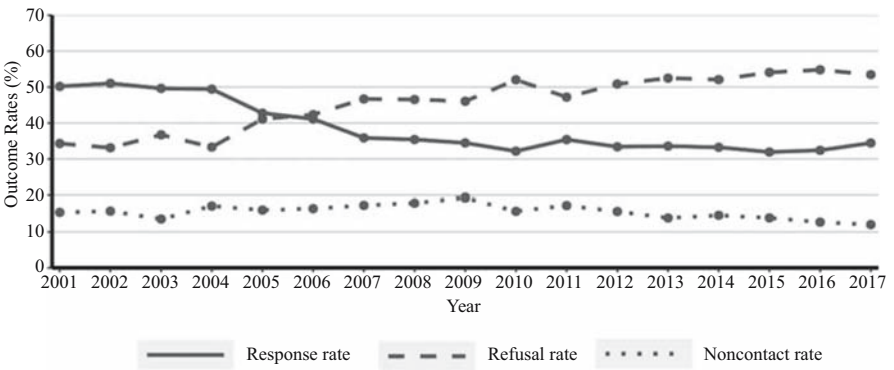

Fig. 1. Outcome rates in the cross-sectional samples of the IAB establishment panel by survey year.

34.5% (2017), an average yearly decline of about 1%. A closer look reveals that the largest decrease in the response rate of about 13.6% took place between 2004 and 2007, an average decrease of 4.5% per year. Since 2007, the response rate has been mostly stable and reached a low point of 31.9% in 2015. The declining response is mainly driven by refusals rather than noncontacts. While the noncontact rate fluctuated slightly around 15%, peaked at 19.4% in 2009, and decreased slightly since 2011, the share of refusals trended upward from 34.3% (2001) to 53.5% (2017). It is evident that the decrease in response is primarily explained by an increasing share of refusing establishments.

5.2. *Nonresponse Bias Trends*

Since a temporal change in the outcome rates was found, possible consequences in the form of compositional distortions in the respondent pool (i.e., nonresponse bias) may exist. We examined this possibility by first looking at nonresponse bias, followed by refusal and noncontact bias. Figure 2 depicts the average absolute relative nonresponse bias for the two summary variable groups (general characteristics and employee structure) and overall (for the tabular version, see online supplemental material Table S3). In general, there was an increasing average nonresponse bias overall and for both variable groups. The overall absolute relative nonresponse bias ranged from 5.23% in 2001 to 8.34% in 2017 – an overall increase of about 60% – with a low point of 2.95% in 2004.

A comparison of the variable groups revealed that the average biases were similar across groups, but showed some minor differences. General characteristics displayed the greatest range of 6.16% (2001) and 9.42% (2017), closely followed by employee structure with a range of 4.67% (2001) and 7.68% (2017). The percentage increases in the average bias of these variable groups were 53% and 65%, respectively, over the entire 17-year observation period. In summary, the aggregated nonresponse biases had a mostly increasing trend over the observation period.

The absolute relative nonresponse bias for each individual variable category is presented in online supplemental material Table S4. Among the general establishment characteristics, establishment size stood out the most. The largest absolute relative nonresponse bias, which ranged from 2.54% (2001) to 35.45% (2017), was observed for the category 50+ employees. Likewise, the relative bias for the 1–9 employees' category showed an increasing trend, but

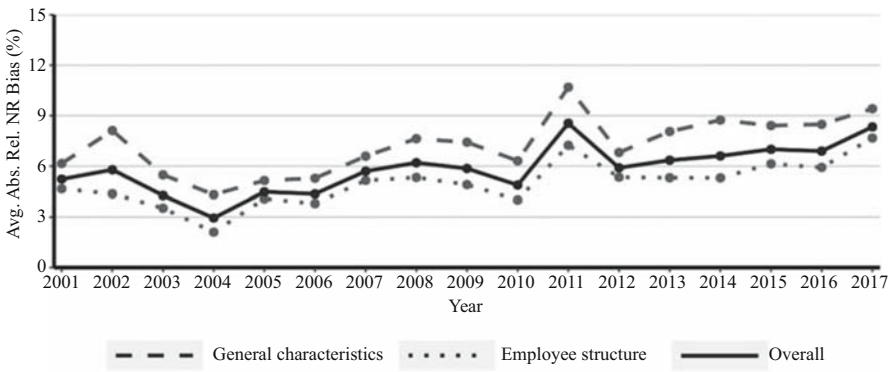

Fig. 2. Average absolute relative nonresponse bias of the summary variable groups and overall.

never exceeded 8%. Furthermore, the industry category agriculture/production displayed large relative bias values (often exceeding 10%), but no clear time trend was identified.

Regarding the employee structure group, establishments with 0-8% high-qualified employees had the largest growth in absolute relative nonresponse bias, ranging from 0.90% to 23.29% in 2001 and 2017, respectively. All other categories did not show a distinct pattern over the observation period and exceeded 10% relative bias only occasionally. The relative bias for the category of establishments with 100% German employees steadily rose over the observation period but never exceeded 10%.

Figure 3 displays the trends of average absolute relative refusal bias (for the tabular version, see online supplemental material Table S5). The results show that the pattern of refusal bias is similar to nonresponse bias with mostly slightly larger values. The overall average absolute relative refusal bias has more than doubled, with a range between 4.90% (2001) and 10.02% (2017). Thus, the increase in overall refusal bias was much larger than that of overall nonresponse bias.

Each variable group showed an increase in the refusal bias over the observation period, but a noteworthy difference between them is not readily apparent. However, it has to be highlighted that the employee structure group showed a larger overall increase in average refusal bias over the observation period than the general characteristics group. The trends for the single category refusal biases (online supplemental material Table S6) largely resembled their corresponding total relative nonresponse bias trends.

Figure 4 depicts the average absolute relative noncontact bias, which does not show any distinct trend (for the tabular version, see online supplemental material Table S7). The average absolute relative noncontact bias has declined from 5.31% in 2001 to 3.55% in 2017. Additionally, the average absolute relative noncontact biases for the two variable groups hardly differed. Only for the individual variable category 50+ employees was the relative noncontact bias noticeable (online supplemental material Table S8). The relative noncontact bias exceeded 10% in most years, but no increasing or decreasing trend was observed. In summary, noncontact bias was unremarkable compared to the refusal and overall nonresponse biases. The overall nonresponse bias, which increased especially in later years, was therefore mainly driven by an increasing refusal bias over time.

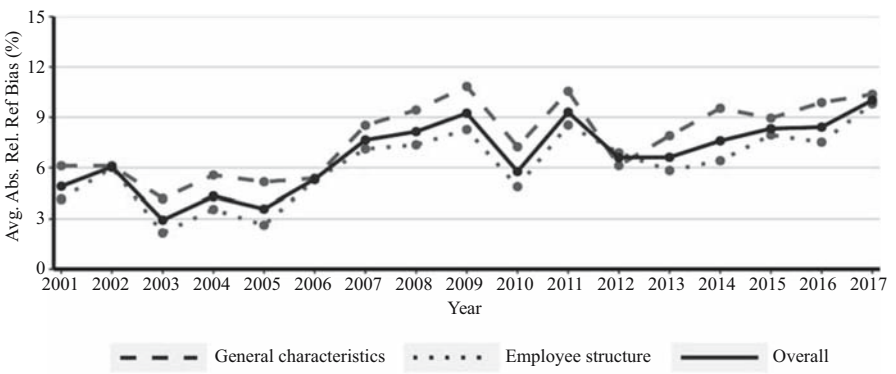

Fig. 3. Average absolute relative refusal bias of the summary variable groups and overall.

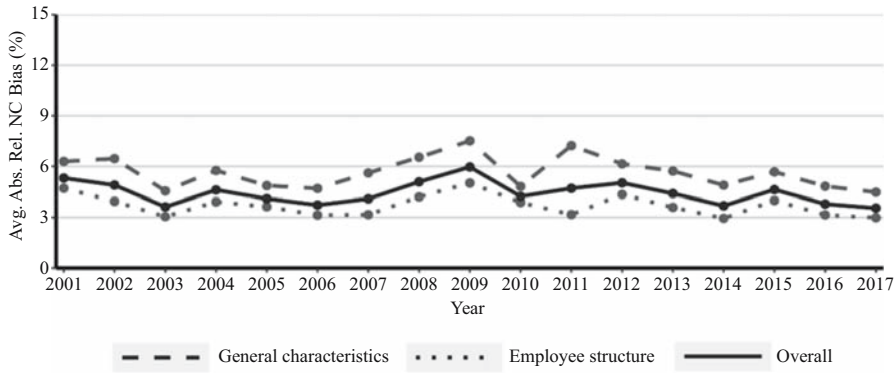

Fig. 4. Average absolute relative noncontact bias of the summary variable groups and overall.

To gain a deeper insight into the relationship between the nonresponse rate and nonresponse bias, we examined whether a correlation exists in the IAB Establishment Panel. We correlated the nonresponse rate to the individual bias values for each variable category of the same year. Figure 5 displays the nonresponse rate against the absolute relative nonresponse biases for all 29 variable categories. The overall correlation coefficient is 0.15 indicating a weak positive correlation. This correlation is consistent with studies from the household survey literature, which also found a small positive correlation, indicating that as the nonresponse rate increases so does the potential for nonresponse bias (Groves 2006; Groves and Peytcheva 2008).

5.3. *Trends in the Likelihood of Participation*

The third research question concerns the extent to which establishment characteristics are associated with survey participation and whether such associations change over time. Table 2 shows the averaged results of the 17 logistic regression models for year-specific survey participation conditional on the administrative predictor variables, and the number of times the predictor variables were statistically significant ($p < 0.05$) across all models. To simplify the presentation, average marginal effects (AMEs) are shown, which are

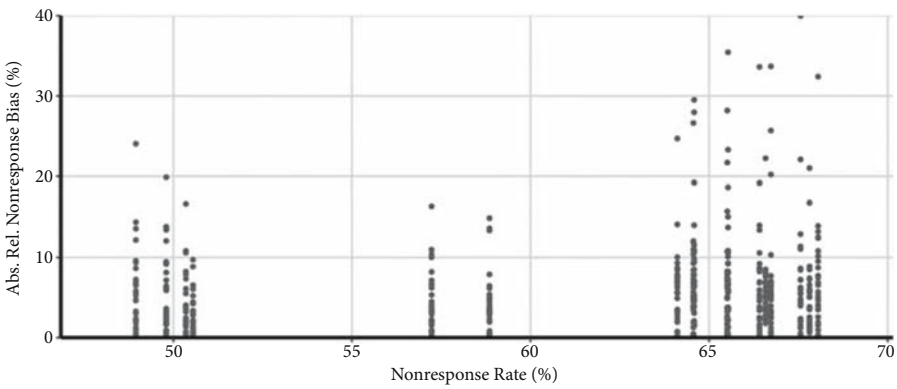

Fig. 5. Percentage absolute relative nonresponse bias of every variable category by the nonresponse rate of the same year.

interpreted as the average probability of response for each covariate compared to its reference group (Kohler and Kreuter 2012). The mean AMEs provide a summary impression of the association between establishment characteristics and survey participation across the 17 years. The range of all estimates is between -0.20 and 0.16. In general, most mean AME values are small and many are close to 0, which means that they did not have a strong influence on participation.

Out of all the predictors, establishment size had the largest negative influence on survey participation: establishments with more than 1,000 employees had an average probability of response that was 0.20 lower than establishments with 1–4 employees. The mean AME values show a consistent pattern that the likelihood to participate decreased with an increasing number of employees. The largest positive average AME value (0.16) was attributed to establishments located in areas with less than 20,000 inhabitants, suggesting that these establishments were more likely to participate than establishments located in areas with more than 500,000 inhabitants. Overall, area population showed a homogenous trend as the likelihood of participation decreased with increasing area population size.

The industry of an establishment also showed some significant effects across years. For example, establishments in the transport and communication industry were, on average, less likely (-0.08) to participate than those in the manufacturing industry (reference). Additionally, the services for companies industry showed more significant AME values than other industries, which resulted in a mean AME of -0.08. Furthermore, establishments in West and North Germany were less likely to participate (-0.05 and -0.03, respectively), on average, compared to those in East Germany. For establishments in South Germany the mean AME showed almost no difference in the response propensity compared to East Germany. The variable regarding changes in the number of employees from the previous to the current year, which is a proxy for the general economic conditions of an establishment's environment, was not strongly related to survey participation as the mean AME values were close to zero and rarely statistically significant in the year-specific models.

The employee structure predictors, which reflect the social demographics of the employees, showed mainly very small effects on participation, and rarely were these predictors statistically significant in the regression models. The strongest predictor that was most often significant in this group over the observation period was the percentage of German employees: establishments with only German employees were more likely (0.04) to participate than those that also employ non-Germans.

Lastly, we examined trends in the regression estimates over time. As the mean AMEs already indicated, the magnitudes of the coefficient estimates were small, and many predictors were not statistically significant in most years. Thus, for most variables, no reliable increasing or decreasing trend was observed over time. As an example, the year-specific AME values for the variable establishment size, which showed the largest average negative AME, are interpreted here (for full results, see online supplemental material Figure S1). In most years the AMEs for this variable were statistically significant, especially if the establishment had 20 or more employees; however, the trends were rather stable with minor fluctuations and did not show any increasing or decreasing shifts over time. One exception which showed a remarkable trend over time was location of establishment. As depicted in Figure 6, the impact of the location on survey participation changed over the years. In particular, establishments in South and North Germany showed

Table 2. Mean average marginal effects (AMEs) of logistic regression models of survey participation for years 2001 to 2017.

Variable	Category	Mean AME	Number of statistically significant AME values across all 17 models ($p < 0.05$)
General characteristics			
Location	East Germany	REF	
	South Germany	0.002	5
	North Germany	-0.03	10
	West Germany	-0.05	14
Establishment size (number of employees)	1–4	REF	
	5–9	-0.01	5
	10–19	-0.03	5
	20–49	-0.05	10
	50–99	-0.07	13
	100–199	-0.10	12
	200–499	-0.12	13
	500–999	-0.16	15
	1,000+	-0.20	13
Industry	Agriculture/mining/energy/water	0.03	4
	Manufacturing industry	REF	
	Construction industry	-0.03	4
	Trade/repair	-0.05	9
	Transport/communication	-0.08	13
	Financial intermediation	-0.05	7
	Services mainly for companies	-0.08	12
	Other services	-0.007	8
	Public sector	-0.02	2
Year of foundation	1970s/1980s	0.02	3
	1990s	0.003	3
	2000s	REF	
	Unknown	0.02	2
Change in the number of employees since previous year	Decrease	-0.02	2
	No change	REF	
	Increase	0.01	2
	Unknown	-0.01	3
Area population size (number of inhabitants)	<2,000	0.16	14
	2,000–4,999	0.16	16
	5,000–19,999	0.16	17
	20,000–49,999	0.13	16
	50,000–99,999	0.12	15
	100,00–499,999	0.09	17
	>500,000	REF	

Table 2. Continued

Variable	Category	Mean AME	Number of statistically significant AME values across all 17 models ($p < 0.05$)
Employee structure			
Pct. of female employees	0–15		REF
	>15–45	0.02	2
	>45–75	0.03	7
	>75–100	0.02	4
Pct. of German employees	0– < 100		REF
	100	0.04	12
Average age of employees (years)	10,5–36		REF
	>36–41	-0.01	2
	>41–45	-0.007	0
	>45–88	-0.005	1
Pct. of low-qualified employees	0		REF
	>0–100	0.001	1
Pct. of middle-qualified employees	0–50		REF
	>50–75	0.003	1
	>75–90	0.01	0
	>90–100	0.01	2
Pct. of high-qualified employees	0		REF
	>0–8	0.01	0
	>8–21	0.01	2
	>21–100	0.01	5

Note: REF specifies the reference category for every variable in the logistic regression.

increasingly positive AMEs for later years, that is, these establishments became increasingly more likely to respond over time compared to establishments in East Germany. For establishments in West Germany the trend was less consistent. All other predictors did not show any notable increasing or decreasing trends in their AMEs over time (see online supplemental material Figures S2–S11).

6. Discussion

This study examined trends in response rates and nonresponse bias in the IAB Establishment Panel's yearly cross-sectional samples from 2001 to 2017 and yielded three main findings. First, we found that yearly response rates decreased by almost a third from 50.2% in 2001 to 34.5% in 2017, with the largest decrease of 13.6% from 2004 to 2007. While the noncontact rate fluctuated almost evenly, the refusal rate steadily rose over this period and was the main driver of nonresponse. Second, the average absolute relative nonresponse bias, measured across 29 individual estimates, increased over the same period from 5.23% to 8.34%, an increase of about 60%. The largest increase in aggregate nonresponse bias was observed for estimates related to the establishments' employee structure (65% increase), followed by general characteristics of the establishments (53%).

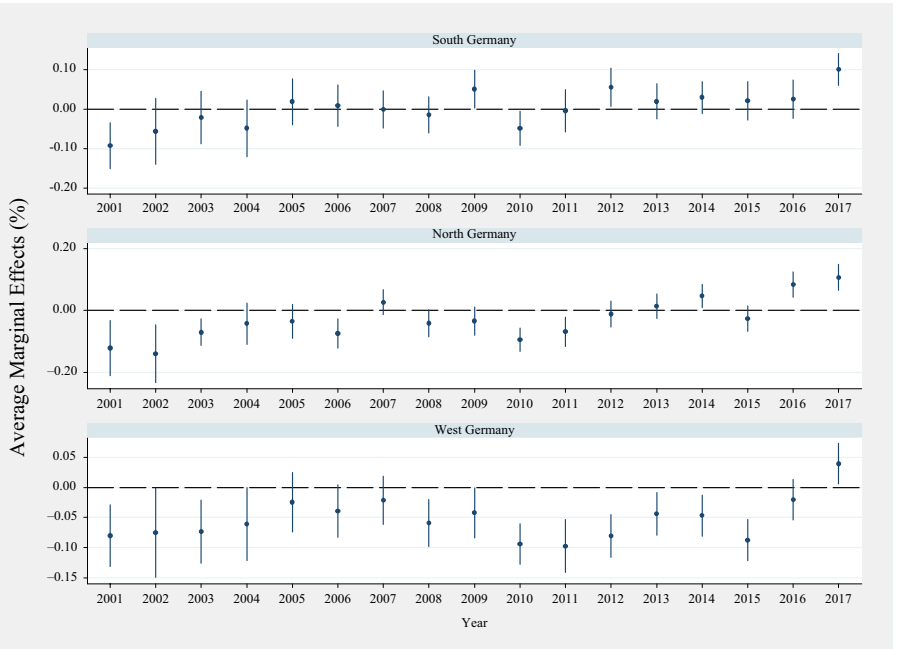

Fig. 6. Average marginal effects and 95% confidence intervals by year for the variable: Location (reference category: East Germany).

By separately considering noncontact and refusal bias, it became evident that the latter bias was the primary driver of the growth in nonresponse bias. A rather low positive correlation of 0.15 was found for the nonresponse rate and the absolute relative nonresponse biases for the same year, which corresponded to the small positive correlations found in the household survey literature (Groves 2006; Groves and Peytcheva 2008). Lastly, we found only few consistently strong predictors of survey participation across the 17-year observation period. Specifically, larger establishments with more than 1,000 employees were less likely to participate compared to smaller establishments. Furthermore, establishments located in smaller area population sizes were more likely to participate. Both relationships were relatively consistent over time, with minor fluctuations.

The declining response rate trend that we found in the repeated cross-sectional IAB Establishment Panel surveys is consistent with trends observed in household surveys, which generally showed declining response rates in recent years (e.g., Luiten et al. 2020; Beullens et al. 2018; Brick and Williams 2013). This is remarkable given that the survey participation decision among establishments is unique to the decision-making process among households. Our findings are also in line with Janik (2011) and Seiler (2010) who found that a change in the number of employees from the previous to the current year, which is a proxy for the general economic conditions of an establishment's environment, did not affect survey participation. Furthermore, we found evidence supporting the results of Phipps and Toth (2012), who ascertained that establishments located in large cities with more than one million inhabitants were less likely to participate compared to those located in smaller cities. Our results suggested this pattern also held for establishments in cities with more than 500,000 inhabitants. Finally, our results regarding establishment size

agreed with other findings showing that the likelihood of survey participation decreases with increasing number of employees (Earp et al. 2018; Phipps and Toth 2012; Janik 2011; Tomaskovic-Devey et al. 1995).

The present study further examined new predictors of survey participation, focusing not only on general establishment characteristics, but also on the composition of the establishments' employee structure. The results indicated that establishments with only German employees were slightly more likely to participate than establishments that also employ non-Germans. However, other predictors, such as the shares of female or middle-qualified employees, did not show consistently strong effects.

The strengths of the present study included the long observation period of 17 years and the rich administrative data available for analyzing nonresponse bias in each of the yearly cross-sectional samples. Nevertheless, some study limitations must be acknowledged. Namely, we considered only one data source, the IAB Establishment Panel, a large face-to-face survey based in Germany. The study results must therefore be interpreted with caution when generalizing them to other studies, countries, and data collection modes. Furthermore, it is important to note that the establishment characteristics considered in our participation models do not fully explain the internal decision-making processes that occur within an establishment. For example, larger establishments may have stricter participation policies for non-mandatory surveys compared to smaller establishments. In addition, larger establishments likely have more complex internal decision-making processes and more difficulties finding the best respondent and accessing the requested information. Including more detailed information about these internal processes would undoubtedly improve the explanatory power of establishment participation models (Bavdaz et al. 2019; Willimack and Nichols 2010; Fisher et al. 2003; Willimack et al. 2002).

In conclusion, the results provide evidence that large establishments are strongly underrepresented in voluntary surveys. Since these establishments have a larger impact on the resulting statistics compared to smaller ones, it is critical that the field of business survey methodology focus efforts on addressing this issue. One possible research direction in this context is the use of incentives to motivate participation (Dillman et al. 2014; Beckler and Ott 2006; Jobber et al. 1991). For example, survey sponsors might consider offering larger establishments detailed personalized reports of the study results showing how their establishment compares to other establishments in similar locations or industries (Luo and White 2005). Another approach is to tailor the recruitment procedure for the very large establishments, as they likely require special treatment. Specifically, voluntary surveys may benefit from borrowing from the recruitment strategies commonly used in mandatory surveys, such as providing personalized support and persistent follow-ups to the largest establishments through the use of a dedicated team of survey specialists and subject-matter experts who recruit and assist establishments throughout the entire survey process.

7. References

AAPOR (American Association for Public Opinion research). 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* AAPOR.

Bavdaz, M., D. Giesen, D.L. Moore, P.A. Smith, and J. Jones. 2019. "Qualitative Testing for Official Establishment Survey Questionnaires". *Survey Research Methods* 13(3): 267–288. DOI: https://doi.org/10.18148/srm/2019.v13i3.7366.

Beckler, D.G., and K. Ott. 2006. "Indirect Monetary Incentives with a Complex Agricultural Establishment Survey". In Proceedings of the American Statistical Association, Survey Research Methodology Section, 2741–2748. Alexandria, VA, USA. American Statistical Association. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.511.9782&rep = rep1&type = pdf (accessed November 2021).

Beullens, K., G. Loosveldt, C. Vandenplas, and I. Stoop. 2018. "Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts?" *Survey Methods: Insights from the Field*: 1–12. DOI: https://doi.org/10.13094/SMIF-2018-00003.

Brick, J.M., and R. Tourangeau. 2017. "Responsive Survey Designs for Reducing Nonresponse Bias." *Journal of Official Statistics* 33(3): 735–752. DOI: http://dx.doi.org/10.1515/JOS-2017-0034.

Brick, J.M., and D. Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-sectional Surveys." In *The Annals of the American Academy of Political and Social Science* 645: 36–59. DOI: https://doi.org/10.1177/0002716212456834.

Brixy, U., S. Kohaut, and C. Schnabel. 2007. "Do Newly Founded Firms Pay Lower Wages? First Evidence from Germany." *Small Business Economics* 29(1-2): 161–171. DOI: https://doi.org/10.1007/s11187-006-0015-x.

Christianson, A., and R.D. Tortora. 1995. "Issues in Surveying Businesses: An International Survey." In *Business Survey Methods,* 235–256. New York: John Wiley & Sons. DOI: https://doi.org/10.1002/9781118150504.ch14.

D'Aurizio, L., and G. Papadia. 2019. "Using Administrative Data to Evaluate Sampling Bias in a Business Panel Survey." *Journal of Official Statistics* 35(1): 67–92. DOI: https://doi.org/10.2478/jos-2019-0004.

Davis, W.R., and N. Pihama. 2009. "Survey Response as Organisational Behaviour: An Analysis of the Annual Enterprise Survey 2003-2007." New Zealand Association of Economists Conference 2009: 1–16. Wellington, New Zealand: New Zeland Association of Economists. Available at: https://ro.uow.edu.au/eispapers/826/ (accessed November 2021).

De Leeuw, E., and W. de Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse,* edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 41–54. New York: Wiley.

Dillman, D.A., J.D. Smyth, and L.M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* New York: John Wiley & Sons.

Earp, M., D. Toth, P. Phipps, and C. Oslund. 2018. "Assessing Nonresponse in a Longitudinal Establishment Survey Using Regression Trees." *Journal of Official Statistics* 34(2): 463–481. DOI: https://doi.org/10.2478/jos-2018-0021.

Ellguth, P., S. Kohaut, and I. Möller. 2013. "The IAB Establishment Panel – Methodological Essentials and Data Quality." *Journal for Labour Market Research* 47: 27–41. DOI: https://doi.org/10.1007/s12651-013-0151-0.

Ellguth, P., and S. Kohaut. 2019. "A Note on the Decline of Collective Bargaining Coverage: The Role of Structural Change." *Jahrbücher für Nationalökonomie und Statistik* 239(1): 39–66. DOI: https://doi.org/10.1515/jbnst-2017-0163.

Eurofound. 2015. *Third European Company Survey – Overview Report: Workplace Practices – Patterns, Performance and Well-Being*. Luxembourg: Publications Office of the European Union. Available at: https://www.eurofound.europa.eu/sites/default/files/ef_publication/field_ef_document/ef1502en_0.pdf (accessed November 2021).

Fischer, G., F. Janik, D. Müller, and A. Schmucker. 2008. "The IAB Establishment Panel – from Sample to Survey to Projection." FDZ Methodenreport 01/2008: Nuremberg. Available at: https://core.ac.uk/reader/6561380 (accessed November 2021).

Fisher, S., J. Bosley, K. Goldenberg, W. Mockovak, and C. Tucker. 2003. "A Qualitative Study of Nonresponse Factors Affecting BLS Establishment Surveys: Results." Presented at the Annual Joint Statistical Meetings, San Francisco, CA. Available at: https://www.bls.gov/osmr/research-papers/2003/pdf/st030230.pdf (accessed November 1st 2021).

Groves, R.M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5): 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M., and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons.

Groves, R.M., and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72(2): 167–189. DOI: https://doi.org/10.1093/poq/nfn011.

Henze, p. 2014. "Structural Change and Wage Inequality: Evidence from German Micro Data." *Center for European, Governance and Economic Development Research Working Paper* 204. DOI: http://dx.doi.org/10.2139/ssrn.2422471.

Janik, F. 2011. "Unit Non-response in Establishments Surveyed for the First Time in the IAB Establishment Panel." FDZ Methodenreport 04/2011: Nuremberg. Available at: https://www.researchgate.net/publication/228956272_Unit_non-response_in_establishments_surveyed_for_the_first_time_in_the_IAB_Establishment_Panel (accessed November 2021).

Janik, F., and S. Kohaut. 2012 "Why Don't They Answer? Unit Non-response in the IAB Establishment Panel." *Quality & Quantity* 46: 917–934. DOI: https://doi.org/10.1007/s11135-011-9436-y.

Jobber, D., H. Mirza, and K.H. Wee. 1991. "Incentives and Response Rates to Cross-national Business Surveys: A Logit Model Analysis." *Journal of International Business Studies* 22: 711–721. DOI: https://doi.org/10.1057/palgrave.jibs.8490852.

Kohler, U., and F. Kreuter. 2012. *Datenanalyse mit Stata: Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung.* München: Oldenbourg.

Luiten, A., J Hox, and E. de Leeuw. 2020. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys." *Journal of Official Statistics* 36(3): 469–487. DOI: http://dx.doi.org/10.2478/JOS-2020-0025.

Luo, A., and G.D. White. 2005. "Exploring a New Establishment Survey Incentive to Improve Response Rates." In Proceedings of the American Statistical Association, Survey Research Methodology Section, May 12-15, 2005. 3915–3918. Miami Beach,

Florida. American Statistical Association. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.528.9845&rep = rep1&type = pdf (accessed November 2021).

Mullins, J.P. 2016. "One Hundred Years of Current Employment Statistics – an Overview of Survey Advancements." *Monthly Labor Review*. DOI: https://doi.org/10.21916/mlr.2016.39.

Paxson, M.C., D.A. Dillman, and J. Tarnai. 1995. "Improving Response to Business Mail Surveys." In *Business Survey Methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, 303-316. New York: Wiley.

Petroni, R., R. Sigman, D. Willimack, S. Cohen, and C. Tucker. 2004 "Response Rates and Nonresponse in Establishment Surveys–BLS and Census Bureau." Presented to the *Federal Economic Statistics Advisory Committee*, 1–50. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.485.5983&rep = rep1&type = pdf (accessed November 2021).

Phipps, P., and D. Toth. 2012. "Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data." *The Annals of Applied Statistics* 6(2): 772–794. DOI: https://doi.org/10.1214/11-AOAS521.

Sakshaug, J.W., and M. Huber. 2016. "An Evaluation of Panel Nonresponse and Linkage Consent Bias in a Survey of Employees in Germany." *Journal of Survey Statistics and Methodology* 4(1): 71–93. DOI: https://doi.org/10.1093/jssam/smv034.

Sakshaug, J.W., S. Hülle, A. Schmucker, and S. Liebig. 2019a. "Panel Survey Recruitment With or Without Interviewers? – Implications for Nonresponse Bias, Panel Consent Bias, and Total Recruitment Bias." *Journal of Survey Statistics and Methodology* 8(3): 540–565. DOI: https://doi.org/10.1093/jssam/smz012.

Sakshaug, J.W., B. Vicari, and M.P. Couper. 2019b. "Paper, E-mail, or Both? Effects of Contact Mode on Participation in a Web Survey of Establishments." *Social Science Computer Review* 37(6): 750–765. DOI: https://doi.org/10.1177/0894439318805160.

Schmucker, A., A. Ganzer, J. Stegmaier, and S. Wolter. 2018. "Establishment History Panel 1975-2017." FDZ Datenreport 09/2018: Nuremberg. DOI: https://doi.org/10.5164/IAB.FDZD.1809.en.v1.

Seiler, C. 2010. "Dynamic Modelling of Nonresponse in Business Surveys." Ifo Working Paper No. 93. Munich: ifo Institute – Leibniz Institute for Economic Research at the University of Munich. Available at: https://www.econstor.eu/bitstream/10419/73706/1/IfoWorkingPaper-93.pdf (accessed November 2021).

StataCorp. 2017. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC.

Sudman, S., D.K. Willimack, E. Nichols, and T.L. Mesenbourg. 2000. "Exploratory Research at the U.S. Census Bureau on the Survey Response Process in Large Companies." In Proceedings of the Second International Conference on Establishment Surveys, June 17-21, 2000: 327–337. Buffalo, New York. American Statistical Association.

Tomaskovic-Devey, D., J. Leiter, and S. Thompson. 1995. "Item Nonresponse in Organizational Surveys." *Sociological Methodology* 25: 77–110. DOI: https://doi.org/10.2307/271062.

Wagner, J. 2012. "Average Wage, Qualification of the Workforce and Export Performance in German Enterprises: Evidence from KombiFiD Data." *Journal for Labour Market Research* 45: 161–170. DOI: https://doi.org/10.1007/s12651-012-0106-x.

Willimack, D.K., E. Nichols, and S. Sudman. 2002. "Understanding Unit and Item Nonresponse in Business Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little, 213–227. New York: Wiley.

Willimack, D.K., and E. Nichols. 2010. "A Hybrid Response Process Model for Business Surveys." *Journal of Official Statistics* 26 (1): 3–24. Available at: https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/a-hybrid-response-process-model-for-business-surveys.pdf%20 (accessed November 2021).

Willimack, D.K., and G. Snijkers. 2013. "The Business Context and Its Implications for the Survey Response Process." In *Designing and Conducting Business Surveys,* edited by G. Snijkers, G. Haraldsen, J. Jones, and D.K. Willimack, 39–82. Hoboken, NJ: John Wiley & Sons.

Received September 2020
Revised December 2020
Accepted March 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 955–979, http://dx.doi.org/10.2478/JOS-2021-0041

Robust Estimation of the Theil Index and the Gini Coeffient for Small Areas

Stefano Marchetti[1] *and Nikos Tzavidis*[2]

Small area estimation is receiving considerable attention due to the high demand for small area statistics. Small area estimators of means and totals have been widely studied in the literature. Moreover, in the last years also small area estimators of quantiles and poverty indicators have been studied. In contrast, small area estimators of inequality indicators, which are often used in socio-economic studies, have received less attention. In this article, we propose a robust method based on the M-quantile regression model for small area estimation of the Theil index and the Gini coefficient, two popular inequality measures. To estimate the mean squared error a non-parametric bootstrap is adopted. A robust approach is used because often inequality is measured using income or consumption data, which are often non-normal and affected by outliers. The proposed methodology is applied to income data to estimate the Theil index and the Gini coefficient for small domains in Tuscany (provinces by age groups), using survey and Census micro-data as auxiliary variables. In addition, a design-based simulation is carried out to study the behaviour of the proposed robust estimators. The performance of the bootstrap mean squared error estimator is also investigated in the simulation study.

Key words: Small area estimation; M-quantile models; inequality indicators.

1. Introduction

Formulating and implementing policies, and allocating funds requires timely, reliable and disaggregated estimates of a large set of parameters, such as means, quantiles, poverty and inequality indicators. Sample surveys provide an effective way of obtaining estimates for such population characteristics. Estimation, however, can become difficult when the focus is on domains (areas) with small sample sizes. The term "small areas" is typically used to describe domains whose sample sizes are not large enough to allow for reliable direct estimation, that is, estimation based only on the sample data from the domain (Rao and Molina 2015). When direct estimation leads to unreliable estimates, one has to rely upon alternative model-based methods for producing small area estimates. Two approaches for model-based small area estimation are based on the mixed effect models (Rao and Molina 2015) and the M-quantile models (Chambers and Tzavidis 2006).

Despite the fact that poverty indicators have been studied extensively under both approaches (Molina and Rao 2010; Marchetti et al. 2012), small area estimation of

[1] University of Pisa, Department of Economia e Management, Via C. Ridolfi 10, 56124 Pisa (PI), Italy. Email: stefano.marchetti@unipi.it
[2] University of Southampton, Social Statistics and Demography Social Sciences, Southampton SO17 1BJ, United Kingdom. Email: n.tzavidis@soton.ac.uk
Acknowledgments: This work has been developed under the support of the project Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy (InGRID-2), grant agreement 730998.

inequality indicators using the M-quantile approach has not been studied extensively. In this article, we study M-quantile small area estimators of the Theil index and the Gini coefficient. These are the two inequality indicators most commonly used by practitioners. The popularity of the Gini coefficient is mainly due to its simplicity, while the appeal of the Theil index lies in its decomposability into "between" and "within" domains. The estimation of these inequality measures is challenging because of their non-linear form. Model assumptions become even more important and departure from these assumptions may have a more noticeable effect on the estimates.

Often, inequality indicators are estimated from variables that are skewed and affected by outliers, such as consumption and income. Chambers and Tzavidis (2006) and Sinha and Rao (2009) proposed model-based outlier robust methods for small area estimation. Chambers and Tzavidis (2006) addressed the issue of outliers robustness in small area estimation using an approach based on fitting M-quantile models (Breckling and Chambers 1988) to the survey data, while Sinha and Rao (2009) addressed this issue from the perspective of linear mixed models. Chambers et al. (2014) defined such methods as robust projective, since they project the behavior of the robust working model of the sample onto the non-sampled part of the population. Tzavidis et al. (2010) and Chambers et al. (2014) proposed methods that allow for contributions from representative sample outliers. These methods are defined as robust predictive method, since they attempt to predict the contribution of the population outliers to target parameters. Other alternatives are possible, for example Gershunskaya and Lahiri (2010) include a modification of a classical linear mixed model assuming that the underlying distribution is a scale mixture of two normal distributions, where outliers are assumed to have a larger variance than regular observations. The proposed estimators can be classified as robust predictive. The ELL (or World Bank) proposed by Elbers et al. (2003) and the Empirical Best Predictor (EBP) proposed by Molina and Rao (2010) are among widely used methods for poverty mapping. These methods are based on linear mixed models, and assume normally distributed errors. When data are skewed, the log transformation is commonly used to obtain approximately normally distributed model residuals. However, in some cases a log transformation may not be appropriate. Recently, Tzavidis et al. (2018) and Rojas-Perilla et al. (2020) proposed the use of data-driven power transformations in small area estimation. An alternative is to specify a model with alternative distributional assumptions to deal with skew-data.

For instance Graf et al. (2019) discuss an EBP approach under a generalized beta distribution of the second kind for the errors terms and Elbers and Van der Weide (2014) propose a method for estimating distribution functions using a mixture of normal distributions for the model errors. Diallo and Rao (2018) derive an EB estimator by relaxing the normality assumptions, assuming skew-normal errors. The approach we propose in this article is based on the M-quantile model and is an alternative to estimators under the linear mixed model.

The remainder of the article is organized as follows. Section 2 introduces the quantities of interest, which are the Theil index and the Gini coefficient, Section 3 summarizes the M-quantile approach to small area estimation, Section 4 introduces the small area estimators of Theil index and Gini coefficient based on M-quantile models, using a Monte Carlo approximation and a bias correction technique, that is, the Chambers and Dunstan

(1986) correction. Moreover, we discuss mean squared error estimation. Section 5 is devoted to evaluating the performance of the proposed estimators by means of Monte Carl o design-based simulations. In Section 6 we present results on Gini and Theil estimates at provincial level in the Tuscany region in Italy. Section 7 summarizes the main results of the article and puts forward ideas for further research.

2. Direct Estimation of the Theil Index and the Gini Coefficient

Inequality measures are mainly based on non-linear statistics. The most popular of these is the Gini coefficient (Gini 1914). It has been shown to be inferior to more recently measures, such as the Zenga index (Zenga 2007), nevertheless, it has a number of advantages over other measures, such as its simplicity, and it is still widely proposed in empirical studies.

Let i be the index for domains (or areas), $i = 1, \ldots, m$ where m is the number of domains, and let j be the index for units within the domain. We denote the population size, sample size, sampled part of the population and non-sampled part of the population in area i respectively by N_i, n_i, s_i and r_i, We assume that the sum over the areas of N_i and n_i is equal to N and n respectively.

The Gini coefficient can be defined in many ways. Usually, it is defined by means of the Lorenz curve. A popular alternative is based on the absolute value of the difference between all pairs of the target variable:

$$G_i = \frac{\Delta_i}{2\mu_i}, \tag{1}$$

where $\mu_i = \int y dF_i(y)$, $\Delta_i = \int \int |y_1 - y_2| \, dF_i(y_1) \, dF_i(y_2)$, $y \geq 0$ and y_1, y_2 are random variables with a common distribution, that is $F_i(y_1) = F_i(y_2) = F_i(y)$. Usually y represent a measure of the income or consumption. In the rest of the article y is a continuous variable with support $(0, +\infty)$ and distributionfunction $F_i(y)$, where the subscript i indicates the domain.

The statistic G is equal to 1 when inequality is at its maximum and it is zero at its minimum (equal distribution).

Another popular inequality statistics is the Theil index (Theil 1967), which belong to the family of generalized entropy measures. It can be defined as (Bourguignon 1979; Shorrocks 1980; Cowell and Kuga 1981; Foster 1983; Maasoumi 1986)

$$T_i = \frac{\nu_i}{\mu_i} - \log(\mu_i), \tag{2}$$

where $\mu_i = \int y \, dF_i(y)$, $\nu_i = \int y \log(y) \, dF_i(y)$ and $y > 0$.

The statistic T is equal 0 when all the population units share the same amount of the total of y, that is, equal distribution, and it is equal $\log(N)$ (where N is the population size) under maximum inequality, that is, one unit holds the total amount of y and the other units hold 0. Its popularity is mainly due to its decomposability into "between" and "within" domains. Assuming T is the Theil index for the entire population that is divided into m domains, then

$$T = \sum_{i=1}^{m} f_i T_i + \sum_{i=1}^{n} f_i \log \frac{\mu_i}{\mu},$$

where $f_i = \frac{N_i \mu_i}{N_\mu}$ is the share of y in domain i, μ is the population mean of y and T_i is the Theil index in domain i. The first sum is the part that is due to inequality within domains, the second is the part that is due to differences between domains.

We now discuss direct estimation of inequality indicators for small areas (domains). Direct estimation for the Gini coefficient is not straightforward. Some popular direct estimators in the literature are known to be negatively biased in small samples (Deltas 2003; Alfons and Templ 2013), such as

$$\tilde{G}_i^{Dir} = \frac{2\sum_{j=1}^{n_i}\left(w_{ij}\sum_{h=1}^{j} w_{ih}\right) - \sum_{j=1}^{n_i} y_{ij} w_{ij}^2}{\sum_{j=1}^{n_i} w_{ij} \sum_{j=1}^{n_i} y_{ij} w_{ij}} - 1,$$

where the values $y_{ij}, j = 1, \ldots, n_i$ are assumed to be sorted in ascending order and w_{ij}, $j = 1, \ldots, n_i$ is the survey weight associated to y_{ij}.

Davidson(2009) notes that the main term in the bias of $\tilde{G}_i^{Dir}$ can be removed by a $n_i(n_i - 1)^{-1}$ multiplication, under simple random sampling design. However, as noted by Langel and Tillè (2013) under complex sample designs the correction of Davidson (2009) is not trivial. We decide to use the following direct estimator (Langel and Tillè 2013):

$$\hat{G}_i^{Dir} = \frac{\hat{\Delta}_i^{Dir}}{2\hat{\mu}_i^{Dir}} = \frac{\sum_{j=1}^{ni}\sum_{k=1}^{ni} w_{ij} w_{ik} |y_{ij} - y_{ik}|}{N_i^2} \frac{1}{2N_i^{-1}\sum_{j=1}^{ni} w_{ij} y_{ij}}, \tag{3}$$

where N_i is the populationsize in area i (assumed known).

For direct estimation of the Theil index we use the estimator proposed in Davidson and Flachaire (2007), here adapted to account for the use of a complex sampling design:

$$\hat{T}_i^{Dir} = \frac{\hat{\nu}_i^{Dir}}{\hat{\mu}_i^{Dir}} - \log\left(\hat{\mu}_i^{Dir}\right), \tag{4}$$

where $\hat{\mu}_i^{Dir} = N_i^{-1}\sum_{j=1}^{n_i} y_{ij} w_{ij}$, $\hat{\nu}_i^{Dir} = N_i^{-1}\sum_{j=1}^{n_i} y_{ij} \log\left(y_{ij}\right) wij$. The direct estimator we use is biased for small samples because $\hat{\nu}_i^{Dir}/\hat{\mu}_i^{Dir}$ is a biased ratio estimator of ν_i/μ_i, though it should be consistent for large samples. Nevertheless, we decided to use estimators (3) and (4) because their forms are suitable for applying the Chambers and Dunstan (1986) correction.

Although variance estimation of direct estimates is not of interest in this articles, it can be shown that an asymptotic variance estimator of $\hat{T}_i^{Dir}$ (under simple random sampling) can be derived using the Delta method. However, Davidson and Flachaire (2007) notes that this variance estimator leads to inference that is not accurate even in a large sample. The same result applies to standard bootstrap variance estimation. Variance estimation of the Theil index is also discussed, among others, in Mills and Zandvakili (1997).

Variance estimation of Equation (3) is not straightforward, even under assumption of log-normality of the target. Asymptotic estimators of the variance have been proposed by for example Battacharya (2007), while bootstrap techniques are discussed for example in Mills and Zandvakili (1997); Alfons and Templ (2013). A literature review about the variance estimation of the Gini coefficient can be found in Langel and Tillè (2013).

3. Outlier Robust Small Area Estimation Using M-Quantiles

3.1. M-Quantile Approach to Small Area Estimation

A robust approach to small area estimation is based on the use of the quantile/M-quantile regression model (Chambers and Tzavidis 2006).

In what follows, we assume that a vector of p auxiliary variable x_{ij} is known for each population unit j in small area $i = 1, \ldots, m$ and that values of the variable of interest y are available from a sample that includes units from all the small areas of interest. We further assume that the sampling design is ignorable conditional on the covariate information, for example conditional on the design variables.

The M-quantile of the order $q \in (0, 1)$ of the conditional density of y given the set of covariates x, $f(y|x)$, is defined as the solution $Q_y(q|x,\psi)$ of an estimating equation $\int \psi_q\{y - Q_y(q|x, \psi)\}f(y|x)dy = 0$, where ψ_q denotes an asymmetric influence function, which is the derivative of an asymmetric loss function ρ_q. In particular, a linear M-quantile regression model for y_{ij} given x_{ij} is one where we assume that

$$Q_y\left(q|x_{ij}, \psi\right) = x_{ij}^T \beta_\psi(q). \tag{5}$$

That is, we allow a different set of p regression parameters for each value of $q \in (0, 1)$. The estimator of $\beta_\psi(q)$ can be obtained by solving

$$\sum_{i=1}^{m} \sum_{j \in s_i} \psi_q\left(y_{ij} - x_{ij}^T \beta_\psi(q)\right) x_{ij} = 0$$

with respect to $\beta_\psi(q)$, assuming that

$$\psi_q\left(y_{ij} - x_{ij}^T \beta_\psi(q)\right) = 2\psi\left\{S^{-1}\left(y_{ij} - x_{ij}^T \beta_\psi(q)\right)\right\} \times \left\{qI\left(y_{ij} - x_{ij}^T \beta_\psi(q) > 0\right) + (1-q)I\left(y_{ij} - x_{ij}^T \beta_\psi(q) \le 0\right)\right\},$$

where s is a suitable robust estimate of scale, for example, the MAD estimate $s = median\,|y_{ij} - x_{ij}^T \beta_\psi(q)|/0.6745$. A popular choice for the influence function is the Huber, $\psi(u) = uI(|u| \le c) + c\ \mathrm{sgn}(u)I(|u| > c)$ (Chambers and Tzavidis 2006). However, alternative influence functions are also possible. Provided that the tuning constant c is strictly greater than zero, estimates of $\beta_\psi(q)$ are obtained using iterative weighted least squares (IWLS).

Chambers and Tzavidis (2006) extended the use of M-quantile regression models to small area estimation. They characterized the conditional variability across the population of interest by the M-quantile coefficients of the populationunits. For unit j in area i this coefficient is the value q_{ij} such that $Q_y(q_{ij}|x_{ij}, \psi) = y_{ij}$. The M-quantile coefficients are determined at the population level. Consequently, if a hierarchical (grouping/clustering) structure does explain part of the variability in the population data, then we expect units within clusters to have similar M-quantile coefficients.

When the conditional M-quantiles are assumed to follow the linear model (5), with $\beta_\psi(q)$ a sufficiently smooth function of q, Chambers and Tzavidis (2006) define a naive

estimator of the mean, that is, $\hat{\mu}_i^{naive} = N_i^{-1}\left\{\sum_{j\epsilon s_i} y_{ij} + \sum_{j\epsilon r_i} x_{ij}^T \hat{\beta}_\psi(\hat{\theta}_i)\right\}$, where $\hat{\theta}_i$ is an estimate of the average value of theM-quantile coefficients of the units in area i. See Chambers and Tzavidis (2006) for further details on the estimation of the M-quantile coefficients at unit level and for the computation of the small area M-quantile coefficients. Bianchi et al. (2018) proposed a test statistic for testing how close the domain-specific quantile coefficients are to 0.5, which is used in the application.

The M-quantile small area model can be more formally defined as follows:

$$y_{ij} = x_{ij}^T \beta_\psi(\theta_2) + \epsilon_{ij}, \tag{6}$$

where $\beta_\psi(\theta_i)$ is the unknown vector of M-quantile regression parameters for the unknown area-specific M-quantile co-efficient θ_i, and ϵ_{ij} is the unit level random error term with distribution function for which no explicit parametric assumptions are being made. The unknown parameters $\beta_\psi(\theta_i)$ and θ_i are estimated as mentioned from sample data, the model residuals are then $e_{ij} = y_{\mathrm{ij}} - x_{ij}^T \hat{\beta}_\psi(\hat{\theta}_{\mathrm{i}})$.

3.2. Bias Correction

A robust projective estimator (naive estimator, e.g., $\hat{\mu}_i^{naive}$) assumes that all the non-sampled units follow the (robustly fitted) working model. However, in practice we should expect that there will be outliers not only in the sample, but also among the non-sampled units. Hence, using the M-quantile predictions for the out-of-sample units directly leads to a biased estimator of the small area target parameter. This is linked to the idea of representative and non-representative outliers described in Chambers (1986) and Chambers et al. (2014). Using the ideas in Chambers (1986), Tzavidis et al. (2010) substitute a consistent estimator of the distribution function, using the approach of Chambers and Dunstan (1986), to derive a version of the M-quantile estimator adjusted for bias also referred to as a robust-predictive estimator. In particular, Tzavidis et al. (2010) define the Chambers-Dunstan (CD) estimator of the small area distribution function as

$$\hat{F}_i^{CD}(t) = N_i^{-1}\left[\sum_{j\in s_i} I(y_{ij} \le t) + n_i^{-1}\sum_{k\in r_i}\sum_{j\in s_i} I(x_{ik}^T \hat{\beta}_\psi(\hat{\theta}_i) + e_{ij} \le t)\right]. \tag{7}$$

Estimates of θ_i and $\beta_\psi(\theta_i)$ are obtained following Chambers and Tzavidis (2006).

By using the Chambers-Dunstan estimator of the small area distribution function, one can define a general framework for small area estimation that allows for the estimation of small area averages, quantiles, non-linear indicators for example, the Gini coefficient and the Theil index. For example the M-quantile CD-based estimator of the average of y in small area i is defined as

$$\hat{\mu}_i^{CD} = \int_{-\infty}^{+\infty} y\, \mathrm{d}\hat{F}_i^{CD}(y) = N_i^{-1}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij} + (1-f_i)\sum_{j\in s_i} e_{ij}\right]. \tag{8}$$

where $f_i = n_i N_i^{-1}$ is the sampling fraction in area i and $\hat{y}_{ij} = x_{ij}^T \hat{\beta}_\psi(\hat{\theta}_i), j \in r_i$ (Tzavidis et al. 2010). The bias correction is the third addend in Equation (8), and means that this estimator has higher variability than the naive M-quantile estimator. Nevertheless, because of its bias robust properties, Equation (8) is usually preferred, over the naive M-quantile estimator in practice.

Similarly, Tzavidis et al. (2010) use the CD estimator of the small area distribution function to propose an estimator of the small area quantiles, and Marchetti et al. (2012) discuss estimation of the Foster et al. (1984) poverty measures.

4. M-Quantile Model-Based Estimation of the Theil Index and the Gini Coefficient

In this section, we describe the methodology for estimating the Theil index and the Gini coefficient for small areas using the M-quantile approach. We derive these estimators using the bias correction introduced by Chambers and Dunstan (1986) and extended to the small area framework by Tzavidis et al. (2010). We start by describing the small area estimator of the Theil index and then the Gini coefficient. The Monte-Carlo version of these estimators is also considered at the end of the section.

4.1. Small Area Estimation for the Theil Index

To estimate T at the small area level we plug-in the CD estimator of the distribution function (7) in Equation (2). Therefore, the small area estimator of the Theil index can be written as

$$\hat{T}_i^{CD} = \frac{\hat{v}_i^{CD}}{\hat{\mu}_i^{CD}} - \log(\hat{\mu}_i^{CD}) \tag{9}$$

where $\hat{\mu}_i^{CD} = \int y d\hat{F}_i^{CD}(y), \hat{v}_i^{CD} = \int y \log(y) d\hat{F}_i^{CD}(y)$. As an alternative, v_i^{CD} can also be estimated using a transformed variable $z = y \log(y)$, therefore $\hat{v}_i^{CD} = \int z d\hat{F}_i^{CD}(z)$. Using first order Taylor expansion we can show that Equation (9) is unbiased, assuming model-unbiasedness of $\hat{\mu}_i^{CD}$ and $\hat{v}_i^{CD}$ (see Equations (17) and (18) in the Appendix, Section 7). The estimators $\hat{\mu}_i^{CD}$ and $\hat{v}_i^{CD}$ can be assumed model-unbiased because $\hat{F}_i^{CD}(t)$ is model-unbiased for $F_i(t)$ under some reasonable conditions specified in Chambers and Dunstan (1986), and Wu and Sitter (2001).

We already introduced the CD-based estimator of the small area mean $\hat{\mu}_i^{CD}$ in Equation (8). Noting that

$$\int_{-\infty}^{+\infty} g(t) d\hat{F}_i^{CD}(t) = N_i^{-1}\left\{\sum_{j\in s_i} g(y_{ij}) + n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i} g(\hat{y}_{ik} + e_{ij})\right\},$$

we can obtain the CD-based estimator of $\nu = g(y) = y \log(y)$ as follows,

$$\hat{v}_i^{CD} = N_i^{-1}\left\{\sum_{j\in s_i} y_{ij}\log(y_{ij}) + n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i}(\hat{y}_{ik} + e_{ij})\log(\hat{y}_{ik} + e_{ij})\right\}, \tag{10}$$

where $\hat{y}_{ik}$ is the predicted value of y_{ik} for the out of sample unit $k \in r_i$ using model in Equation (6) and $e_{ij}, j = 1, \ldots, n_i$ are the model residuals in area i. Alternatively, ν_i can also be estimated as

$$\hat{v}_i^{CD} = N_i^{-1}\left\{\sum_{j\in s_i} z_{ij} + \sum_{k\in r_i}\hat{z}_{ik} + (N_i + n_i - 1)\sum_{j\in s_i} e_{ij}^z\right\}, \tag{11}$$

where $z_{ij} = y_{ij}\log(y_{ij})$, $\hat{z}_{ik} = \hat{y}_{ik}\log(\hat{y}_{ik})$ and e_{ij}^zs are residuals obtained from the M-quantile small area model in Equations (6) where y_{ij} is replaced by z_{ij}, see Equations (15) and (16) in the Appendix for further details. Empirically, Equations (10) and (11) are equivalent and give the same results, however, it is difficult to show this algebraically.

However, Equation (11) is computationally faster because it doesn't involve the double summation present in Equation (10).

4.2. Small Area Estimation for the Gini Coecient

To estimate the Gini coefficient we adopt the same strategy used before for the Theil index. Therefore, we plug-in the distribution function estimator (7) in Equation (1) leading to the following small area estimator

$$\hat{G}_i^{CD} = \frac{\hat{\Delta}_i^{CD}}{2\hat{\mu}_i^{CD}}, \tag{12}$$

where $\hat{\Delta}_i^{CD} = \int\int |y_1 - y_2| d\hat{F}_i^{CD}(y_1) d\hat{F}_i^{CD}(y_2)$. Assuming $\hat{F}_i^{CD}$ is model-unbiased for F_i then using first-order Taylor expansion we can show that the estimator (12) is approximately model-unbiased (Equations (20) and (21) in the Appendix).

Estimator $\hat{\mu}_i^{CD}$ is that of Equation (8). Estimator $\hat{\Delta}_i^{CD}$ is obtained as follows (see Equation (19) in the Appendix)

$$\begin{aligned}\hat{\Delta}_i^{CD} &= \int\int |t_1 - t_2| d\hat{F}_i^{CD}(t_1) d\hat{F}_i^{CD}(t_2) \\ &= N_i^{-2} \left\{ \sum_{j \in s_i} \sum_{l \in s_i} |y_{ij} - y_{il}| + n_i^{-2} \sum_{j \in s_i} \sum_{k \in r_i} \sum_{l \in s_i} \sum_{h \in r_i} \left| \hat{y}_{ik} + e_{ij} - (\hat{y}_{ih} + e_{il}) \right\} \right|. \end{aligned} \tag{13}$$

Computing the quadruple summation in Equation (13) is computationally intensive when the population area size is large (for example greater than 5000 units). In the R language (R Development Core Team 2013) the use of arrays to speed up the computation is possible. As an alternative, we wrote a C function that can be called in R through a dynamic library, which uses a nested "for" to compute the quadruple summation in reasonable time also for large population domain sizes. The R-code is available in the supplementary materials. The required computational time is discussed in Section 6.

4.3. Small Area Estimation Based on Monte Carlo Approximation

It is important to mention that small area target parameters can alternatively be estimated by approximating the distribution of the unknown quantity $y_{ik,}\ k \in r$ by means of Monte-Carlo simulations. Let δ_i be a parameter of interest in area i that depends from a vector of known constants $c = \{c_1, c_2, \ldots\}$:

$$\delta_i = \delta_i(c) = h(y_{ij} \cup y_{ik,}c) \quad j \in s_i, k \in r_i,$$

where h is a function of the target variable y and the vector of known constants c. Let $y_s = \{y_j,\ j \in s\}$ be the vector of sample observations, which obey a superpopulation model, and let t be the vector of unknown parameters of the superpopulationmodel. A predictor of δ_i can be obtained by preserving the values corresponding to the sample units and predicting those corresponding to non sampled units:

$$\hat{\delta}_i = h(y_{ij} \cup E[y_{ik}|y_s; \hat{t}], c),$$

where $\hat{t}$ is a consistent estimator of t and $E[y_{ik}\,|y_s; \hat{t}] = \hat{y}_{ik}$ an unknown quantity that can be approximated by using the Monte Carlo simulation. It is important to note that if $E[y_{ik}|y_s; \hat{t}]$

depends on x_{ij} then the covariate values need to be known for all the units in the population. This is comparable to other methodologies that use unit-level models to estimate domain-specific non-linear indicators, for example the EBP and ELL methods.

When we use the M-quantile model to estimate δ_i the Monte Carlo approximation can be obtained asfollows:

1. Fit the M-quantile small area model using the sample values y_s and obtain estimates $\hat{t} = \{\hat{\theta}_i, \beta_\psi(\hat{\theta}_i)\}$,
2. Generate an out of sample vector of size $N_i - n_i$ using

$$y_{ik}^* = x_{ik}^T \hat{\beta}_\psi(\hat{\theta}_i) + e_{ik}^*, k \in ri, i = 1, \ldots, m,$$

   where $e_{ik}^*, k \in r_i, i = 1, \ldots, m$ is drawn from the empirical distribution function of the M-quantile model residuals (residuals can be drawn either from the domain (area) i residuals or from all the residuals).
3. Repeat the process L times. Each time, combine the sample data $y_{ij}, j \in s_i$ and out of sample data $y_{ik}^*, k \in r_i$ for computing $\hat{\delta}_i^{(l)}$.
4. Average the results over L simulations to obtain an estimate of δ_i, $\hat{\delta}_i = L^{-1} \sum_{l=1}^{L} \hat{\delta}_i^{(l)}$.

Further discussion on this Monte Carlo approach can be found in Marchetti et al. (2012). Usually, in real applications linkage between sampled units and population units is not possible, that is the set r is unknown. In this case, the prediction is carried out for all the units in the population $U_i = \{s_i, r_i\}$, then $\hat{\delta}_i = h(E[y_{ik}|y_s; \hat{t}], c), k \in U_i$. When the sampling fraction is very small $h(E[y_{ik}|y_s; \hat{t}], c), k \in U_i$ and $h(y_{ij} \cup E[y_{ik}|y_s; \hat{t}], c), k \in r_i$ are very similar.

Setting

$$h(y_1 \ldots y_{n_i}) = \frac{n_i^{-1}(n_i - 1)^{-1} \sum_{j=1}^{n_i} \sum_{l=1}^{n_i} |y_{ij} - y_{il}|}{n_i^{-1} \sum_{j=1}^{n_i} y_{ij}}$$

we obtain the Gini coefficient MC estimator, and setting

$$h(y_1 \ldots y_{n_i}) = \frac{n_i^{-1} \sum_{j=1}^{n_i} y_{ij} \log y_{ij}}{n_i^{-1} \sum_{j=1}^{n_i} y_{ij}} - \log n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$$

we obtain the Theil index MC estimator.

The M-quantile MC estimators mimic the Elbers et al. (2003) approach. However, it is challenging to theoretically justify this method, therefore, statistical properties are shown via simulations. In contrast, for the M-quantile CD estimators the theoretical background is better understood (Tzavidis et al. 2010).

4.4. MSE Estimation

MSE estimation for M-quantile small area estimators is widely discussed for linear statistics, such as means and totals (Chambers et al. 2014). Less research is available for non-linear statistics. An MSE estimator based on a non-parametric bootstrap scheme for small area estimators under the M-quantile model that can be used also with non-linear

statistics is extensively discussed in Marchetti et al. (2012). More details on the non-parametric bootstrap approach for finite population can also be found, among others, in Lombardía et al. (2003).

Starting from a random sample s selected from a finite population U without replacement, we fit the M-quantile small area model (6), and we obtain estimates $\hat{t} = \{\hat{\theta}, \hat{\beta}_\psi\ (\hat{\theta}_i)\}$ and residuals $e_{ij}, i = 1, \ldots, m;\ j \in s_i$. The bootstrap MSE estimates can be obtained as follows:

1. Given an estimator $\hat{G}(u)$ of the distribution of the residuals $G(u) = P(e \leq u)$, a bootstrap population, consistent with the M-quantile small area model can be generated by sampling from $\hat{G}(u)$ to obtain e^*_{ij}:

$$y^*_{ij} = x^T_{ij}\hat{\beta}_\psi(\hat{\theta}_i) + e^*_{ij}.$$

   For defining $\hat{G}(u)$ we consider two approaches: (a) sampling from the empirical distribution function of the model residuals or (b) sampling from a smoothed distribution function of the model residuals. For each of the two above mentioned approaches, sampling can be done in two ways: (1) by sampling from the distribution of all residuals without conditioning on the small area (unconditional approach) or (2) by sampling from the distribution of the residuals within small area i (conditional approach). These methods are described in detail in Tzavidis et al. (2010).
2. According to point 1, choose one approach from (a) or (b) and one from (1) or (2), and generate B bootstrap populations.
3. From *each* of the B bootstrap population draw L samples using simple random sample– of size n_i– within areas.
4. Using the L samples, compute the estimates of the Theil index and the Gini coefficient according to the methods proposed in Section 4.
5. Let $\hat{\tau}_i$ be the the estimated small area parameter (from the original sample), τ_i^{*b} value) of the bth bootstrap population, $\hat{\tau}_i^{*bl}$ be the small area parameter estimated by using the l sample from the b bootstrap population. The bootstrap estimator of the bias and the variance of $\hat{\tau}_i$ are defined respectively by

$$\hat{B}(\hat{\tau}_i) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left(\hat{\tau}_i^{*bl} - \tau_i^{*b}\right),$$

$$\hat{V}(\hat{\tau}_i) = B^{-1}L^{-1}\sum_{b=1}^{B}\sum_{l=1}^{L}\left(\hat{\tau}_i^{*bl} - \overline{\hat{\tau}}_i^{*b}\right)^2,$$

   where $\overline{\hat{\tau}}_i^{*b} = L^{-1}\sum_{l=1}^{L}\hat{\tau}_i^{*bl}$. The bootstrap MSE estimator of the estimated small area parameter is finally defined as

$$\widehat{MSE}(\hat{\tau}_i) = \hat{V}(\hat{\tau}_i) + \hat{B}(\hat{\tau}_i)^2. \tag{14}$$

Bootstrapping in the presence of outlier contamination is a challenging problem. The properties of the proposed bootstrap MSE are examined in Subsection 5.2. The issue of bootstrapping in the presence of outlier contamination is discussed in Schmid et al. (2016), but further research on bootstrap MSE estimation in the presence of contamination is

needed. A promising approach to tackling this problem is offered by the more recent work in Dongomo-Jiongo and Nguimkeu (2018). The authors propose to generate bootstrap populations by using the non-robust mixed model fit. Although this idea can be applied to the M-quantile predictors, this extension is not immediately applicable and will be considered in future work.

To estimate MSE of Equations (9) and (12) one can also attempt to use a Taylor linearization. However, using simulations, which are not reported here, we have noted that this approximation is not accurate to the desired order, and hence not reliable. The reason is that Taylor expansions are asymptotic results and depend on having a sufficient sample size to work well, while in the small area estimation framework a number of areas are expected to have small sample sizes. Moreover, the Taylor-linearized MSE for the Theil index is the same as the one obtained by the delta method in Davidson and Flachaire (2007), which they prove not to be accurate even for a large sample. It is worth noting that MSE estimation for such indicators is very difficult, in particular for small samples. Therefore, it may be reasonable to expect poor performance of MSE estimators. Future work will consider a bootstrap bias correction for the linearized MSE estimator.

5. Design-Based Evaluation of the Proposed Estimators

In this section we use design-based Monte-Carlo simulations to study the performance of the proposed small area robust estimators of the Theil index and the Gini coefficient. Moreover, we also evaluate the performance of the bootstrap MSE estimator of these.

The population underpinning the design-based simulation is based on the data used in the application in Section 6. Our target domains are the same as those used in the application. The population for the design-based simulation has been obtained by fitting a mixed effects model to the EU-SILC data, and then predicting the target using the Census data.

We fit a linear mixed model (random intercept) on the EU-SILC data using the household equivalized income as target variable and as auxiliary variables *owners* (proportion of households who hold their house), *work status* (a binary variable indicating if the head of the household works), *sex* (a binary variable indicating the sex of the head of the household), *education* (number of year of education of the head of the household), *household size* (number of household members), which are common between Census and EU-SILC.

Then, we generate the target value for all the population units using the Census auxiliary variables and the model estimates, adding variability by sampling from model-level one and two residuals. The resulting synthetic target value has a distribution similar to that observed in the EU-SILC data, as shown in Figure 1. We refer to the generated synthetic equivalized household income and the auxiliary variables as synthetic population.

From the synthetic population we draw 1,000 samples with a design similar to that of the EU-SILC survey in Italy in 2008. The survey design of the EU-SILC in Italy is a two-stage stratified sample with a rotating panel (for details see Istat Siqual). Applying this design to each sample leads to a different sample size, which varies between 1,277 to 1,704 households, with an average of 1,472 (the actual EU-SILC 2008 sample size is

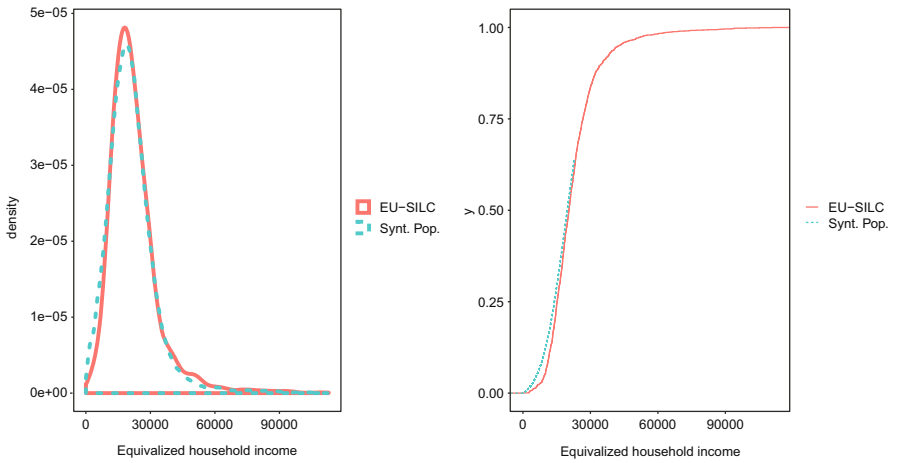

Fig. 1. Density estimates of the household equivalized income from the EU-SILC (solid line) and the synthetic population (dashed line).

1,495). The average sample size across the domains varies from a minimum of 4.9 to a maximum of 204.4, with a mean of 49.1 and a median of 40.

For each sample we estimate the Theil index and the Gini coefficient at the domain level (province by age class) using the M-quantile CD and MC estimators. To compare the results of the proposed estimators we use as a benchmark the Empirical Best Predictor (EBP) proposed by Molina and Rao (2010). This method is based on a linear mixed model and requires a transformation of the response variable to obtain an approximate normal distribution of the model error terms. We first tried to use the log scale, but the results were unsatisfactory. Therefore, we decided to use a data-driven Box-Cox transformation (Box and Cox 1964; Rojas-Perilla et al. 2020). We apply this data-driven transformation in each Monte Carlo replication using the R package emdi (Kreutzmann et al. 2019). For comparing the EBP and M-quantile estimators we also fit the M-quantile model using the same Box-Cox transformation as in the case of the EBP, even though we acknowledge that the best transformation for the EBP it is not necessarily the best transformation for the MQ model.

Usually, in applications to real cases it is not possible to link the sampled units with the population units, and then obtain the set r of the non sampled units. We replicate this situation in this design-based simulation. Estimators are then modified accordingly (see Equations (22), (23) and (24) in the Appendix.

5.1. *Discussion about Point Estimation*

In Table 1, we present results for comparing the M-quantile MC estimator and the EBP estimator. For the M-quantile MC estimator we produce results by using a model that is estimated both with the untransformed income data and the transformed income data. The EBP estimates are produced by using a mixed model fitted to the transformed income following the methodology described in Rojas-Perilla et al. (2020). At this point it is important to clarify the following points.

Although the EBP results on the untransformed scale have been produced, we have decided not to report these because the mixed model assumptions are not satisfied on this scale.

Table 1. Design-based simulation results. Average and median of the relative bias (%) and relative empirical root MSE.

	Transform	Theil		Gini	
		Relative bias %			
		Median	Average	Median	Average
M-quantile MC	No	-1.8	8.4	-1.9	1.7
M-quantile MC	Box-Cox	9.4	9.8	5.3	6.2
EBP	Box-Cox	3.9	5.2	7.0	6.4
		Relative root mean squared error %			
M-quantile MC	No	25.1	31.6	20.8	21.2
M-quantile MC	Box-Cox	20.8	20.7	10.7	13.2
EBP	Box-Cox	22.4	23.0	11.7	12.8

The results are available from the authors. Overall, the results from using the EBP on the untransformed scale show that estimates of the Theil index and the Gini coefficient have very large relative bias compared to M-quantile MC estimates on the same scale. This provides evidence for the robustness properties of the M-quantile estimators.

The results in Table 1 also show that the M-quantile MC estimator on the transformed scale (using the same transformation parameter as the one for the EBP) competes very well with the EBP on the same scale. Here, we acknowledge that using a transformation in conjunction with the M-quantile estimator is not done in an optimal way – as in the case with EBP – and should be used only for initial comparisons of the results on the transformed scale. More research is needed for developing data-driven transformations for the M-quantile methods.

Generally speaking, these results show that the M-quantile-based methods perform well both on the untransformed and transformed scales. Using a transformation appears to improve the results of the M-quantile MC further but as we mentioned above this requires additional research. The EBP method is only considered on the transformed scale for the reasons we described above. These results illustrate the robustness properties of the M-quantile-based methods.

Finally, the M-quantile CD (the results are available from the authors) estimator performs similarly in terms of relative bias to the M-quantile MC on the raw scale (6.2% average relative bias for the Theil index and 2.4% for the Gini coefficient). In terms of relative root MSE, the M-quantile CD shows more variability than the M-quantile MC for the Theil index (average relative MSE of 50.2%) and competes well with the M-quantile MC for the Gini coefficient (average relative root MSE of 22.1%). Moreover, further improvement of the M-quantile CD estimators could be obtained using an influence function for the residuals in Equations (9) and (7) as suggested in Chambers et al. (2014).

The M-quantile MC and CD both provide an alternative to the EBP in those cases where the mixed model assumptions are not met. Although the theory of the M-quantile CD is better understood, the M-quantile MC is computationally simpler and faster to implement. For these reasons practitioners may prefer this approach.

5.2. Empirical Evaluation of the Mean Squared Error Estimator

As concerns the estimation of the MSE, we evaluate the bootstrap estimator (14) using the same data as in the design-based simulation, but limited number of runs, equal to 250, given the high computational time required. We use 1 bootstrap population ($B = 1$) from which we draw 100 bootstrap samples. We draw residuals from the smooth error distribution function unconditionally to the areas (for further details on this technique see Marchetti et al. (2012)).

Due to the long computational time required, we select a sub-set of the population, namely, the provinces of Pisa, Lucca and Massa, which correspond to the North-West of Tuscany. Therefore, there are a total of nine domains, three age groups by three provinces. We study the performance of the bootstrap estimator (14) by computing the relative bias (RB)

$$RB(\widehat{MSE}(\hat{\tau}_i)) = H^{-1}\sum_{h=1}^{H}\frac{\widehat{MSE}(\hat{\tau}_{i,h}) - MSE(\hat{\tau}_i)}{MSE(\hat{\tau}_i)},$$

where $\widehat{MSE}(\hat{\tau}_{i,h})$ is the MSE bootstrap estimate of the target parameter $\hat{\tau}_{i,h}$ in area i and simulation h and $MSE(\hat{\tau}_{i,h})$ s the empirical MSE of estimator $\hat{\tau}_i$ (which we consider as the"true" MSE) computed over 1000 Monte Carlo simulations. We also show a summary of empirical MSEs and estimated MSEs for checking if the bootstrap estimator tracks well the empirical (true) MSE over domains.

The results are summarized in Table 2 and Figure 2. Table 2 shows the average and median across the nine small domains of the relative bias (RB) of the bootstrap MSE

Table 2. Design-based simulation bootstrap MSE estimator results. Average and median across domains of relative bias (%) of the bootstrap MSE estimator.

	Theil		Gini	
	Median	Average	Median	Average
M-quantile CD	-24.3	6.7	-19.0	-15.4
M-quantile MC	-22.3	7.6	-0.6	35.2

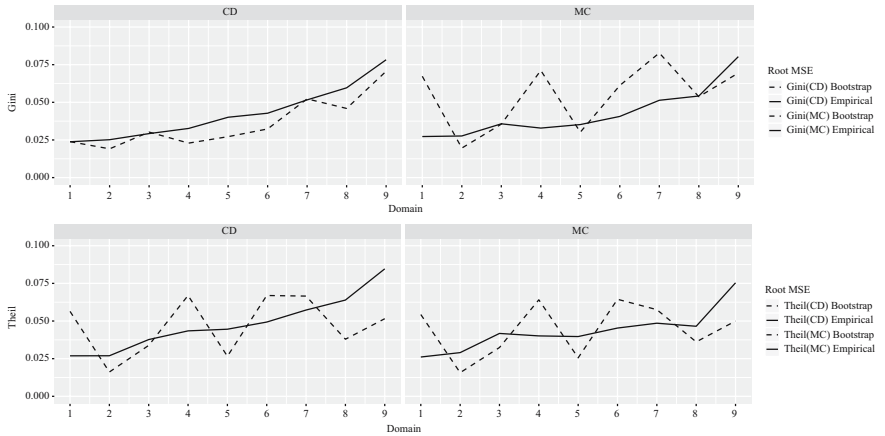

Fig. 2. Design-based simulation bootstrap MSE estimator results. Empirical (true) root MSE and estimated root MSE.

estimator for the Theil index and Gini coefficient M-quantile CD and MC estimators. The average RB is around 7% for the Theil index, while for the Gini coefficient M-quantile CD estimator is −15.4%. The average RB of the Gini coefficient M-quantile MC estimator is about 35%. This high value is mainly due to a high bias in three areas (indeed, the median RB is about 0%), where the presence of big outliers affects the MC method. Looking at both the median and the average of the relative bias (RB) of the M-quantile MC, we can see that the distribution of the RBs is skewed both for the Theil index and the Gini coefficient. The RB related to the M-quantile CD of the Theil index is also skewed, while it seems not to be skewed for the Gini coefficient. However, given the small number of areas used in the simulation due to computational time, it is hard to properly assess the quality of the proposed bootstrap MSE estimators. Considering the limited number of bootstrap populations generated the performance of the MSE estimator is judged to be acceptable for practical purposes. Moreover, since the values of the root MSE are small, a small difference has a big impact in relative terms. We also studied the convergence of the bootstrap MSE estimator for the M-quantile MC. More specifically, we computed the median of the difference between the estimated MSE and the "true" (empirical) MSE while increasing the number of bootstrap replications. The results seem to indicate a small negative biased value for the Theil index, which remains constant after 50 bootstrap replications and a bias that tends to zero for the Gini coefficient as the number of boostrap iterations increase. The results reported here are from a design-based simulation that uses real data. Model-based simulations assessing the properties of the bootstrap MSE estimator (not reported here but available upon request to the authors) show markedly better results.

From the results in Figure 2 we can see that the estimated root MSE tracks reasonably the empirical MSE both for the M-quantile CD and MC estimators of Theil index and Gini coefficient.

6. Estimating the Gini Coefficient and the Theil Index for Small Domains in Tuscany

In this section, we present an application of the proposed methodology, to EU-SILC (Statistics on Income and Living Conditions) data from Italy. A short description of the design was given in Section 5.

The aim is to study the differences in the inequality, if any, among age groups within provinces and provinces within age groups. The domains are defined by the cross-classification of provinces in Tuscany by the age class of the head of the household, leading to a total of 30 domains (ten provinces × three age categories). The age of the head of the household has been divided into three categories, "up to 34", "35–64", "65 and above". This classification comes from the age classes used by the Italian National Institute of Statistics (Istat) in some labor force statistics reports, for example Istat (2017). To evaluate inequality, we estimate both the Gini coefficient (1) and the Theil index (2) to see whether or not they result in estimates of inequality that point in the same direction.

Throughout the article, we refer to the age class "up to 34" as *Young*, "35–64" as *Worker* and "65 and above" as *Aged*. The domain-specific sample size varies between four households (Young in Grosseto) and 207 households (Worker in Firenze) with an average

sample size across domains of 46.9. The population size is about 1.39 million households, it varies between 7,329 (Young in Massa) and 201,019 (Worker in Firenze) with an average of 46,280 households per domain. The sampling fraction across domains is between 0.05% (Young in Grosseto) and 0.22% (Young in Pistoia), with an average of about 0.11%, which approximately correspond to the overall sampling fraction in the EU-SILC in Italy.

The outcome we model is the household equivalized disposable income which is available for each sampled household from the EU-SILC survey 2008. The household equivalized disposable income corresponds to the total household net income (the sum of households' member income after tax payments and social transfers, including pensions) divided by the equivalized household size, which gives a weight of 1.0 to the first adult, 0.5 to other persons aged 14 or over who are living in the household and 0.3 to each child aged less than 14. The explanatory variables are the marital status of the head of the household (four categories, single, married, divorced and widow), the employment status of the head of the household (working/not working), the years of education of the head of the household, the mean house surface (in square meters) at municipality level (LAU 2 level) and the number of household members. These covariates are available both from the EU-SILC and from the Population Census of Italy in 2001. Although the 2008 EU-SILC data were collected seven years after the Census, the 2001 – 2007 period (2008 EU-SILC data refers to 2007 income) was one of relatively slow growth and low inflation in Italy, Therefore, it is reasonable to assume that there was relatively little change in the considered period. It is also important to mention that EU-SILC and Census datasets are confidential. The datasets were provided by Istat to the researchers of the SAMPLE project and were analyzed by respecting the confidentiality restrictions.

Figure 3 shows box-plots of the household equivalized income in each of the 30 domains. The box-plots highlight the asymmetry of the income distribution. The box-plots are ordered (ascending) according to the estimated average of the equivalized household income. We can see that, in general and as expected, Young and Aged groups have a lower income than the Worker group, with some exceptions like the Young group in Lucca which has a rather high income while the Worker group in Massa has a low income.

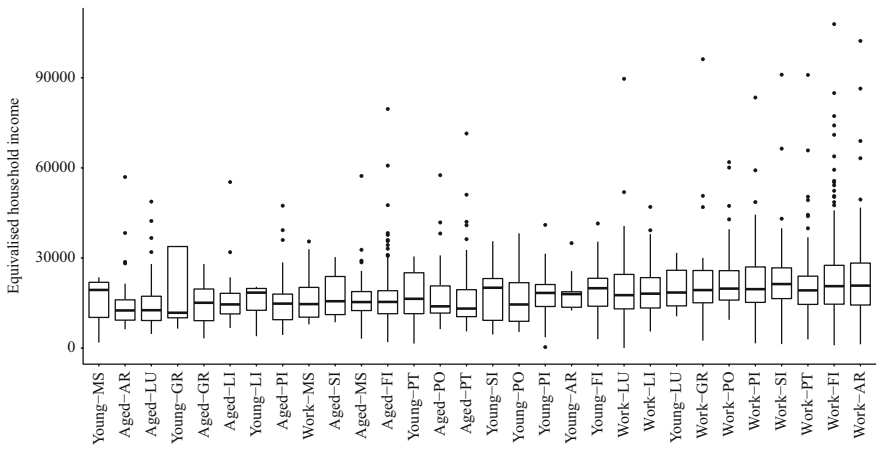

Fig. 3. Box-plots of equivalized income by province and age class. Domains are ordered by average income.

Figure 4 shows normal probability plots of level one and level two residuals obtained by fitting a two-level random effects model to the EU-SILC data both on the original scale outcomes (top) and log scale outcomes (bottom). Households are the level one units and the 30 domains define the level two units. Figure 4 suggests departures from the normality assumptions of level one errors, also for the log scale model. The use of the Shapiro and Wilk (1965) test statistic confirms that the hypothesis of normally distributed level one residuals, both when using the original and log-transformed income variable, is rejected. It may be appropriate in this case to use a small area estimation approach that imposes less strict parametric assumptions and it is robust to outliers.

Using the test statistic proposed by Bianchi et al. (2018), we test how close the domain-specific quantile coefficients are to 0.5. This test statistic is trying to emulate the test for the statistical significance of the random effects variance under the nested regression model. If the test statistic indicates statistically significant differences in the domain M-quantile coefficients, then the model that allows for domain-specific M-quantile coefficients should be preferred to a model that assumes a common M-quantile coefficient leading to a synthetic estimator. The Bianchi et al. (2018) test statistic has been applied to our data. The value of the test statistic is equal to 62.146 and the p-value is equal to 0.000331. The results show that for this application the domain M-quantile coefficients are statistically different from 0.5 and, as a result, using an M-quantile model with domain-specific M-quantile coefficients should be preferred in this case.

We estimate the Theil index and the Gini coefficient using direct, M-quantile CD and MC estimators (for M-quantile CD estimators we use Equations (22), (23) and (24) in the Appendix because it is not possible to link the sampled units with the population units). Comparing these three different point estimates within each domain, we observe that the M-quantile CD and MC estimates follow the same trend as the direct ones. The point estimates are shown in Figure 5.

Small area estimates of the Theil index and the Gini coefficient obtained by the M-quantile MC approach are summarized in Table 3. Both indices vary between provinces within each age group, and also vary between age groups within each province. In particular, the between province variation of the point estimates of the Theil index within the age groups is lower for the Aged group compared to the Young and Worker groups. The between province variation of the point estimates of the Gini coefficients is lower for the Aged group

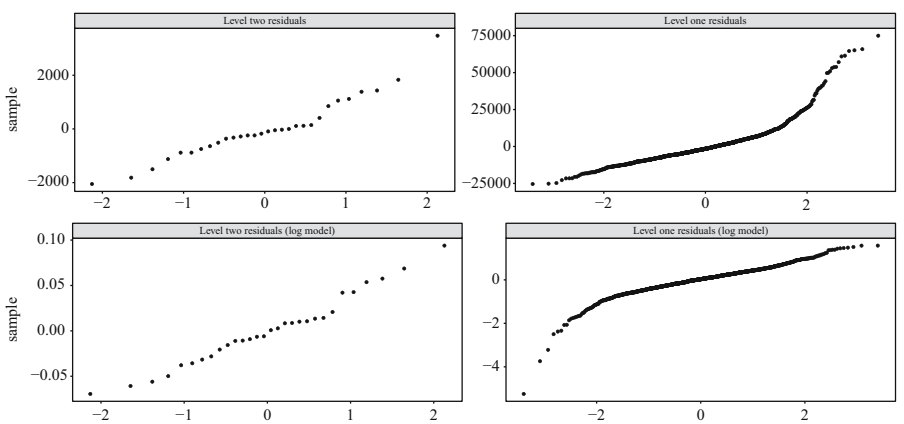

Fig. 4. Q-Q plots of level one and two residuals, row scale (top) and log scale (bottom).

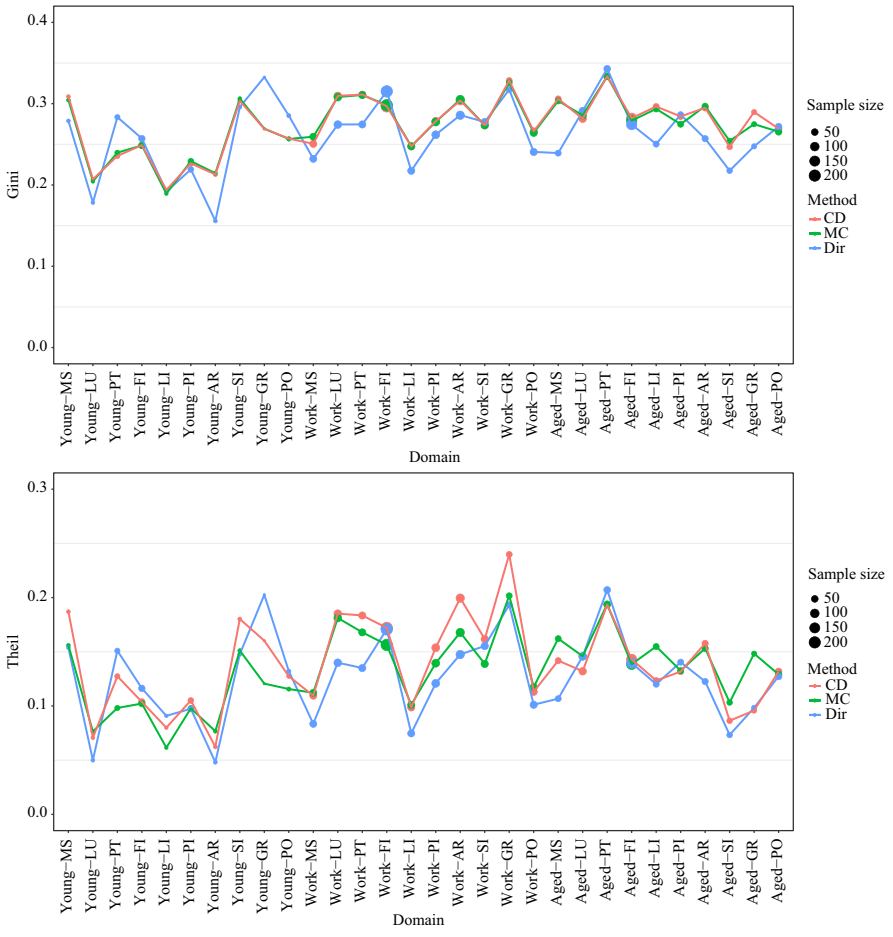

Fig. 5. Point estimates of the Gini coefficient estimates (upper plot) and Theil index estimates (lower plot).

Table 3. Small area estimates of Theil index and Gini coefficient (M-quantile MC approach) by provinces and age groups.

	Theil MC			Gini MC		
	Young	Work	Aged	Young	Work	Aged
MS	0.156	0.112	0.162	0.305	0.259	0.304
LU	0.076	0.181	0.146	0.205	0.308	0.286
PT	0.098	0.168	0.194	0.239	0.311	0.334
FI	0.102	0.156	0.138	0.249	0.298	0.279
LI	0.062	0.101	0.155	0.190	0.248	0.294
PI	0.098	0.139	0.132	0.229	0.277	0.274
AR	0.077	0.168	0.153	0.215	0.305	0.297
SI	0.151	0.139	0.103	0.306	0.273	0.254
GR	0.120	0.202	0.148	0.269	0.326	0.275
PO	0.116	0.117	0.129	0.257	0.264	0.266

compared to the Worker group, which is lower than the Young group. Moreover, according to both inequality indicators the Young group shows a lower inequality compared to Worker and Aged groups. The same conclusions are reached by looking at the M-quantile CD estimates. Finally, even though the two indices are not directly comparable, we can say that both the Gini coefficient and the Theil index show similar levels of inequality.

The results of Table 3 seem reasonable in the level and in the direction among the age groups. One result that can be highlighted is the remarkable difference of the level of inequality between the Work and the Aged group in the province of Grosseto (GR) and Livorno (LI). Somehow, these two results are unexpected. Indeed, we can accept a small reduction or increase of the inequality between Worker and Aged group, but not as big as for the Grosseto and Livorno cases.

Moreover, Grosseto and Livorno are quite similar provinces in terms of many aspects; from an economic point of view Grosseto and Livorno are among the medium-income provinces in Italy. Nevertheless, we observed an increase in the inequality of about 20 percentage points of the Gini coefficient and of about 50 percentage points of the Theil index in Livorno and a decrease of about 15 percentage points of the Gini coefficient and about 27 percentage points of the Theil index in Grosseto. We consider that these figures need to be further investigated, making use of other indicators – such as poverty indexes, income/consumption distributions, and GDP level. These estimates should help socio-economic analysts to better describe local phenomena.

Estimates of the MSE for the M-quantile CD and MC estimates have been obtained using $B = 50$ bootstrap populations and $L = 100$ bootstrap samples (from each population, for a total of 5000 samples). The residuals to generate the populations have been drawn from a smooth distribution unconditional to the areas both for the CD and MC estimators. The choice of the number of bootstrap populations and bootstrap samples has been discussed in Marchetti et al. (2012). The bootstrap resampling scheme we propose is time consuming, however, non-optimized R code run on 2.6GHz quad-core Intel Core i7 took about 260 minutes for the Gini M-quantile CD, 280 minutes for the Theil M-quantile CD and 1100 minutes for the Gini and Theil M-quantile MC. Therefore, we judge the method to be feasible for many applications. Estimates of the standard error of the direct estimates of the Gini coefficient and the Theil index have been obtained by bootstrap techniques. In particular, we obtained the standard error estimates of the Gini coefficient direct estimates using the bootstrap method proposed by Alfons and Templ (2013), available in the R package "laeken" (R Development Core Team, 2013; Alfons and Templ, 2013), and the standard error estimates of the Theil index direct estimates using the semiparametric bootstrap method proposed by Davidson and Flachaire (2007). The estimated variability of direct, M-quantile CD and MC estimates are summarized in Table 4. Both the proposed small area estimators show a gain in efficiency compared to the direct estimators.

7. Conclusions

In this article we presented robust small area estimators based on the M-quantile regression model for the Theil index and the Gini coefficient, two popular inequality measures. M-quantile based estimators are robust versus outliers, which occur frequently on income and consumption data that are often used in socio-economic studies to compute inequality

Table 4. Estimated MSE summarized across domains.

	Min	1st qu.	Median	Mean	3rd qu.	Max.
T^{CD}	0.008	0.016	0.024	0.027	0.034	0.072
T^{MC}	0.006	0.012	0.018	0.020	0.024	0.055
T^{Dir}	0.043	0.080	0.099	0.097	0.109	0.146
G^{CD}	0.010	0.017	0.022	0.027	0.032	0.069
G^{MC}	0.008	0.015	0.020	0.024	0.028	0.065
G^{Dir}	0.017	0.031	0.035	0.040	0.044	0.093

measures. For both the measures of interest we presented two estimating approaches: one based on the Monte Carlo approach and one based on the Chambers and Dunstan (1986) distribution function estimator extended for M-quantile models. The proposed estimators have been applied to EU-SILC data from Tuscany (an Italian region) combined with population Census micro data. The aim of the application was to compare the two inequality measures for provinces by age groups (30 domains in total). Results show that the two inequality indicators go to the same direction, pointing out different levels of inequality among provinces within age groups and vice versa. Moreover, we showed that the proposed methods succeed in improving the estimation efficiency compared to direct estimation. Finally, we evaluated the statistical properties of the proposed estimators as well as their bootstrap mean squared error estimators by means of a design-based Monte Carlo simulation. The proposed methodologies to estimate the Theil index and the Gini coefficient for small domains under a robust framework can be applied widely. The possibility to obtain sound estimates of inequality at a low aggregation level, breaking down domains and geographical areas, provides a valuable tool for socio-economic studies.

Future works may focus on analytic mean squared error estimation of the proposed estimators, and bootstrap based confidence intervals.

7. Appendix

7.1. Theil Index

The M-quantile CD estimator of the Theil index in area i is defined as $\hat{T}_i = \frac{\hat{\nu}_i^{CD}}{\hat{\mu}_i^{CD}} - \log \hat{\mu}_i^{CD}$, $\hat{\mu}_i^{CD}$ is derived in Equation (8). In what follows we show how to obtain $\hat{\nu}_i^{CD}$. First, an estimator of $E[g(y)]$ using the CD approach is:

$$
\begin{aligned}
E\left[g\left(y\right)\right] &= \int_{-\infty}^{+\infty} g(t) d\hat{F}_i^{CD}(t) \\
&= N_i^{-1} \int_{-\infty}^{+\infty} g(t) d\left\{ \sum_{j \in s_i} I(y_{ij} \leq t) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} I(\hat{y}_{ki} + e_{ij} \leq t) \right\} \\
&= N_i^{-1} \left\{ \sum_{j \in s_i} \int_{-\infty}^{+\infty} g(t) dI(yij \leq t) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} \int_{-\infty}^{+\infty} g(t) dI(\hat{y}_{ki} + e_{ij} \leq t) \right\} \\
&= N_i^{-1} \left\{ \sum_{j \in s_i} g(y_{ij}) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} g(\hat{y}_{ki} + e_{ij}) \right\}. \qquad (15)
\end{aligned}
$$

Then, the CD estimator of $v_i = E[y \log(y)]$ follows directly

$$\begin{aligned}
\hat{v}_i^{CD} &= \int_{-\infty}^{+\infty} t \log(t)\ d\hat{F}_i^{CD}(t) \\
&= N_i^{-1} \int_{-\infty}^{+\infty} t \log(t)\, d\Big\{ \sum_{j \in s_i} I(y_{ij} \le t) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} I(\hat{y}_{ki} + e_{ij} \le t) \Big\} \\
&= N_i^{-1} \Big\{ \sum_{j \in s_i} \int_{-\infty}^{+\infty} t \log(t)\, dI(y_{ij} \le t) \\
&+ n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} \int_{-\infty}^{+\infty} t \log t(t) dI(\hat{y}_{ki} + e_{ij} \le t) \Big\} \\
&= N_i^{-1} \Big\{ \sum_{j \in s_i} y_{ij} \log(y_{ij}) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} (\hat{y}_{ik} + e_{ij}) \log(\hat{y}_{ik} + e_{ij}) \Big\}. \qquad (16)
\end{aligned}$$

Let us show that Equation (9) is unbiased by using a first order Taylor expansion. Consider that $\hat{T}_i^{CD}$ is a function of the random variables (estimators) $\hat{\mu}_i^{CD}$ and $\hat{v}_i^{CD}$, and let us write $\hat{T}_i^{CD} = g(\hat{v}_i^{CD}, \hat{\mu}_i^{CD})$. Now let us expand function g using a first order Taylor series around point(v_i, μ_i)

$$g(\hat{v}_i^{CD}, \mu_i^{CD}) = \frac{v_i}{\mu_i} - \log(\mu_i) + \frac{1}{\mu_i}(\hat{v}_{i_i}^{CD} - v_i) - \frac{v_i}{\mu_i^2}(\hat{\mu}_i^{CD} - \mu_i) - \frac{1}{\mu_i}(\hat{\mu}_i^{CD} - \mu_i) + O\big(n^{-1}\big). \qquad (17)$$

If $\hat{v}_i^{CD}$ and $\hat{\mu}_i^{CD}$ are model-unbiased estimators of the parameters ν_i and μ_i the expectation of $g(\hat{v}_i^{CD}, \hat{\mu}_i^{CD})$ is

$$\begin{aligned}
E[\hat{T}_i^{CD}] &= E\big[g\big(\hat{v}_i^{CD}, \hat{\mu}_i^{CD}\big)\big] \\
&\approx E\left[\frac{v_i}{\mu_i} - \log\big(\mu_i\big) + \frac{1}{\mu_i}(\hat{v}_i^{CD} - v_i) - \frac{v_i}{\mu_i^2}\big(\hat{\mu}_i^{CD} - \mu_i\big) - \frac{1}{\mu_i}(\hat{\mu}_i^{CD} - \mu_i)\right] \\
&= \frac{v_i}{\mu_i} - \log\big(\mu_i\big) = T_i. \qquad (18)
\end{aligned}$$

7.2. Gini Coefficient

The estimator $\hat{\Delta}_i^{CD}$ used in Equation (12) is derived as follows

$$\begin{aligned}
\hat{\Delta}_i^{CD} &= \int \int |t_1 - t_2|\ d\hat{F}_i^{CD}(t_1)\ d\hat{F}_i^{CD}(t_2) \\
&= \int N_i^{-1} \int |t_1 - t_2| d\Big\{ \sum_{j \in s_i} I(y_{ij} \le t_1) \\
&+ n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_i} I(\hat{y}_{ik} + e_{ij} \le t_1) \Big\} d\hat{F}_i^{CD}(t_2)
\end{aligned}$$

$$= \int N^{-1}\Big\{\sum_{j\in s_i}\int |t_1 - t_2| dI(y_{ij} \le t_1)$$

$$+ n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i}\int |t_1 - t_2| dI(y_{ik} + e_{ij} \le t_1)\Big\}\, d\hat{F}_i^{CD}(t_2)$$

$$= \int N_i^{-1}\Big\{\sum_{j\in s_i}|y_{ij} - t_2| + n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i}|\hat{y}_{ik} + e_{ij} - t_2|\Big\} d\hat{F}_i^{CD}(t_2) \tag{19}$$

$$= N_i^{-2}\int \Big\{\sum_{j\in s_i}|y_{ij} - t_2| + n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i}|\hat{y}_{ik} + e_{ij} - t_2|\Big\}$$

$$\times\ d\Big\{\sum_{j\in s_i} I(y_{ij} \le t_2)| + n_i^{-1}\sum_{j\in s_i}\sum_{k\in r_i}(\hat{y}_{ik} + e_{ij} \le t_2)\Big\}$$

$$= N_i^{-2}\int \Big\{\sum_{j\in s_i}\sum_{l\in s_i}|y_{ij} - y_{il}| + n_i^{-2}\sum_{j\in s_i}\sum_{k\in r_i}\sum_{l\in s_i}\sum_{h\in r_i}|\hat{y}_{ik} + e_{ij} - (\hat{y}_{ih} + e_{il})|\Big\}.$$

Let us show that Equation (12) is unbiased by using a first order Taylor expansion. Consider that $\hat{G}_i^{CD}$ is a function of the random variables (estimators) $\hat{\mu}_i^{CD}$ and $\hat{\Delta}_i^{CD}$, and let us write $\hat{G}_i^{CD} = g(\hat{\Delta}_i^{CD}, \hat{\mu}_i^{CD})$. Now let us expand function g using a first order Taylor series around point(Δ_i, μ_i):

$$g(\hat{\Delta}_i, \hat{\mu}_i) = \frac{\Delta_i}{2\mu_i} + \frac{1}{2\mu_i}\left(\hat{\Delta}_i - \Delta_i\right) - \frac{\Delta_i}{2\mu_i^2}(\hat{\mu}_i - \mu_i) + o_n^{-1}, \tag{20}$$

then let compute the expectation of $\hat{G}_i^{CD} = g(\hat{\Delta}_i, \hat{\mu}_i)$ under the assumptions that $\hat{\Delta}_i$ and $\hat{\mu}_i$ are model-unbiased

$$E[\hat{G}_i^{CD}] = E[g(\hat{\Delta}_i, \hat{\mu}_i)] \approx E\left[\frac{\Delta_i}{2\mu_i} + \frac{1}{2\mu_i}\left(\hat{\Delta}_i - \Delta_i\right) - \frac{\Delta_i}{2\mu_i^2}(\hat{\mu}_i - \mu_i)\right] = \frac{\Delta_i}{2\mu_i} = G_i. \tag{21}$$

7.3. Estimator when Linkage Between Sampled Units and Population Units is Not Possible

When linkage between sampled units and population units is not possible, that is the set r is unidentifiable, then the prediction is carried out for all the units in the population $U_i = \{s_i \cup_{ri}\}$. Then the estimators of μ_i, ν_i and Δ_i are as follows

$$\hat{\mu}_i^{CD} = N_i^{-1}\left[\sum_{j\in U_i}\hat{y}_{ij} + (1 - f_i)\sum_{j\in s_i} e_{ij}\right] \tag{22}$$

$$\hat{\nu}_i^{CD} = N_i^{-1}\Big\{n_i^{-1}\sum_{j\in s_i}\sum_{k\in U_i}(\hat{y}_{ik} + e_{ij})\log(\hat{y}_{ik} + e_{ij})\Big\} \tag{23}$$

$$\hat{\Delta}_i^{CD} = N_i^{-2}\Big\{n_i^{-2}\sum_{j\in s_i}\sum_{k\in U_i}\sum_{l\in s_i}\sum_{h\in U_i}|\hat{y}_{ik} + e_{ij} - (\hat{y}_{ih} + e_{il})|\Big\}. \tag{24}$$

8. References

Alfons, A. and M. Templ. 2013. "Estimation of social exclusion indicators from complex surveys: The r package laeken." *Journalof Statistical Software* 54 (15): 1–25. DOI: https://doi.org/10.18637/jss.v054.i15.

Battacharya, D. 2007. "Inference on inequality from household survey data." *Journal of Econometrics* 137: 674–707. DOI: https://doi.org/10.1016/j.jeconom.2005.09.003.

Bianchi, A., E. Fabrizi, N. Salvati, and N. Tzavidis. 2018. "Estimation and testing in m-quantile regression with applications to small area estimation." *International Statistical Review* 86 (3): 541–570. DOI: https://doi.org/10.1111/insr.12267.

Bourguignon, F. 1979. "Decomposable income inequality measures." *Econometrica* 42: 27–41. DOI: https://doi.org/10.2307/1914138.

Box, G., and D. Cox. 1964. "An analysis of transformations." *Journal of the Royal Statistical Society Series B* 27 (2): 211–252. DOI: https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Breckling, J., and R. Chambers. 1988. "M-quantiles." *Biometrika* 75 (4): 761–771. DOI: https://doi.org/10.1093/biomet/75.4.761.

Chambers, R.L. 1986. "Outlier robust finite population estimation." *Journal of the American Statistical Associationtion* 81 (396): 1063–1069. DOI: https://doi.org/10.1111/rssb.12019.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis. 2014. "Outlier robust small area estimation." *Journal of the Royal Statistical Society Series B* 76 (1): 47–69. DOI: https://doi.org/10.1111/rssb.12019.

Chambers, R., and Dunstan. 1986. "Estimating distribution function from survey data." *Biometrika* 73: 597–604. DOI: https://doi.org/10.1093/biomet/73.3.597.

Chambers, R., and N. Tzavidis. 2006. "M-quantile models for small area estimation." *Biometrika* 93 (2): 255–268. DOI: https://doi.org/10.1093/biomet/93.2.255.

Cowell, F., and K. Kuga. 1981. "Inequality measurement: An axiomatic approach." *Journal of Economic Theory* 15: 287–305. DOI: https://doi.org/10.1016/S0014-2921(81)80003-7.

Davidson, R. 2009. "Reliable inference for the gini index." *Journal of Econometrics* 150: 30–40. DOI: https://doi.org/10.1016/j.jeconom.2008.11.004.

Davidson, R., and E. Flachaire. 2007. "Asymptotic and bootstrap inference for inequality and poverty measures." *Journal of Econometrics* 141 (1): 141–66. DOI: https://doi.org/10.1016/j.jeconom.2007.01.009.

Deltas, G. 2003. "The small-samples bias of the gini coefficient: results and implications for empirical research." *The Review of Economics and Statistics* 85: 226–34. DOI: https://doi.org/10.1162/rest.2003.85.1.226.

Diallo, M.S., and J.N.K. Rao. 2018. "Small area estimation of complex parameters under unit-level models with skew-normal errors." *Scandinavian Journal of Statistics* 45 (4): 1092–1116. DOI: https://doi.org/10.1111/sjos.12336.

Dongomo-Jiongo, V., and P. Nguimkeu. 2018. *Bootstrapping mean squared errors of robust small-area estimators: Application to the method-of-payments data.* Technical report, Staff Working Paper: 18–28, Bank of Canada. Available at: https://www.bankofcanada.ca/wp-content/uploads/2018/06/swp2018-28.pdf. (accessed November 2021).

Elbers, C., J.O. Lanjouw, and P. Lanjouw. 2003. "Micro-level estimation of poverty and inequality." *Econometrica* 71 (1): 355–364. DOI: https://www.jstor.org/stable/3082050.

Elbers, C., and R. van der Weide. 2014. *Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality.* The World Bank. Available at: https://openknowledge.worldbank.org/handle/10986/19362. (accessed November 2021).

Foster, J. 1983. "An axiomatic characteriazation of the Theil measure of income inequality." *Journal of Economic Theory* 31: 105–121. DOI: https://doi.org/10.1016/0022-0531(83)90023-6.

Foster, J., J. Greer, and E. Thorbecke. 1984. "A class of decomposable poverty measures." *Econometrica* 52: 761–766. DOI: https://doi.org/10.2307/1913475.

Gershunskaya, J., and P. Lahiri. 2010. "Robust small area estimation using a mixture model." In Proceedings of the Joint Statistical Meeting 2010, 1 July to 5 August 2010, Vancouver, British Columbia, Canada. Available at: https://ww2.amstat.org/meetings/jsm/2010/onlineprogram/AbstractDetails.cfm?abstractid=307425 (accessed November 2021).

Gini, C. 1914. "Sulla misura della concentrazione e della variabilita' dei caratteri." In *Atti del Regio Istituto Veneto di Scienze Lettere ed Arti*. Available at: https://www.hetwebsite.net/het/texts/gini/gini_1914.pdf.

Graf, M., J.M. Marín, and I. Molina. 2019. "A generalized mixed model for skewed distributions applied to small area estimation." *TEST* 28 (2): 565–597. DOI: https://doi.org/10.1007/s11749-018-0594-2.

Istat Siqual. 2008. "Information on EU-SILC survey." Available at: http://siqual.istat.it/SIQual/visualizza.do?id=5000170&refresh=true&language=IT.

Istat. 2017. "Occupati e disoccupati." Available at: https://www.istat.it/it/files/2017/07/CS_Occupati-e-disoccupati_giugno_2017.pdf.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis. 2019. "The R package emdi for estimating and mapping regionally disaggregated indicators." *Journal of Statistical Software* 91 (7): 1–33. DOI: https://doi.org/10.17169/refubium-25826.

Langel, M., and Y. Tillè. 2013. "Variance estimation of the gini index: revisiting a result several time published." *Journal of the Royal Statistical Society A* 7: 521–40. DOI: https://doi.org/10.1111/j.1467-985X.2012.01048.x.

Lombardía, M., W. González-Manteiga, and J. Prada-Sánchez 2003. "Bootstrapping the chambers-dunstan estimate of finite population distribution function." *Journal of Statistical Planning and Inference* 116: 367–388. DOI: https://doi.org/10.1016/S0378-3758(02)00240-9.

Maasoumi, E. 1986. "The measurement and decomposition of multi-dimensional inequality." *Econometrica* 54: 991–97. DOI: https://doi.org/10.2307/1912849.

Marchetti, S., N. Tzavidis, and M. Pratesi. 2012. "Non-parametric bootstrap mean squared error estimation for m-quantile estimators of small area averages, quantiles and poverty indicators." *Computational Statistics and Data Analysis* 56 (10): 2889–2902. DOI: https://doi.org/10.1016/j.csda.2012.01.023.

Mills, J., and S. Zandvakili. 1997. "Statistical inference via bootstrapping for measures of inequality." *Journal of Applied Econometrics* 12 (2): 133–50. DOI: https://doi.org/10.1002/(SICI)1099-1255(199703)12:2 < 133::AID-JAE433 > 3.0.CO;2-H.

Molina, I., and J. Rao. 2010. "Small area estimation of poverty indicators." *Canadian Journal of Statistics* 38 (3): 369–385. DOI: https://doi.org/10.1002/cjs.10051.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.yumpu.com/en/document/view/6853895/r-a-language-and-environment-for-statistical-computing. (accessed November 2021).

Rao, J., and I. Molina. 2015. *Small Area Estimation*. Wiley Series in Survey Methodology. Wiley.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis. 2020. "Data-driven transformations in small area estimation." *Journal of the Royal Statistical* Series A 183 (1): 121–148. DOI: https://doi.org/10.1111/rssa.12488.

SAMPLE (Small Area Methods for Poverty and Living Conditions). Project founded by the 7th Framwork Programme of the EU. Grant SSH - CT - 2007 – 217565. Available at: http://www.sample-project.eu.

Schmid, T., N. Tzavidis, R. Münnich, and R.L. Chambers. 2016. "Outlier robust small area estimation under spatial correlation." *Scandinavian Journal of Statistics* 43 (3): 806–826. DOI: https://doi.org/10.1111/sjos.12205.

Shapiro, S., and M. Wilk. 1965. "An analysis of variance test for normality (complete samples)." *Biometrika* 67: 215–216. DOI: https://doi.org/10.2307/2333709.

Shorrocks, A. 1980. "The class of additively decomposable inequality measures." *Econometrica* 48: 613–625. DOI: https://doi.org/10.2307/1913126.

Sinha, S., and J. Rao. 2009. "Robust small area estimation." *The Canadian Journal of Statistics* 37 (3): 381–399. DOI: https://doi.org/10.1002/cjs.10029.

Theil, H. 1967. *Economics and Information Theory*. Chicago: Rand McNally and Company.

Tzavidis, N., S. Marchetti, and R. Chambers. 2010. "Robust estimation of small area means and quantiles." *Australian and New Zeland Journal of Statistics* 52 (2): 167–186. DOI: https://doi.org/10.1111/j.1467-842X.2010.00572.x.

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla. 2018. "From start to finish: a framework for the production of small area official statistics." *Journal of the Royal Statistical Society Series A* 181 (4): 927–979. DOI: https://doi.org/10.1111/rssa.12364.

Wu, C., and R. Sitter. 2001. "Variance estimator for the finite population distribution function with complete auxiliary information." *The Canadian Journal of Statistics* 29. DOI: https://doi.org/10.2307/3316078.

Zenga, M. 2007. "Inequality curve and inequality index based on the ratios between lower and upper arithmetic means." *Statistica e Applicazioni* 4: 3–27. DOI: https://doi.org/10.1400/209575.

Received October 2019
Revised June 2020
Accepted January 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 981–1007, http://dx.doi.org/10.2478/JOS-2021-0042

Occupation Coding During the Interview in a Web-First Sequential Mixed-Mode Survey

Darina N. Peycheva[1], *Joseph W. Sakshaug*[2], *and Lisa Calderwood*[1]

Coding respondent occupation is one of the most challenging aspects of survey data collection. Traditionally performed manually by office coders post-interview, previous research has acknowledged the advantages of coding occupation during the interview, including reducing costs, processing time and coding uncertainties that are more difficult to address post-interview. However, a number of concerns have been raised as well, including the potential for interviewer effects, the challenge of implementing the coding system in a web survey, in which respondents perform the coding procedure themselves, or the feasibility of implementing the same standardized coding system in a mixed-mode self- and interviewer-administered survey. This study sheds light on these issues by presenting an evaluation of a new occupation coding method administered during the interview in a large-scale sequential mixed-mode (web, telephone, face-to-face) cohort study of young adults in the UK. Specifically, we assess the take-up rates of this new coding method across the different modes and report on several other performance measures thought to impact the quality of the collected occupation data. Furthermore, we identify factors that affect the coding of occupation during the interview, including interviewer effects. The results carry several implications for survey practice and directions for future research.

Key words: Occupational classification; self-administration; interviewer-administration; coding error.

1. Introduction

The collection and coding of respondent occupation has been one of the most important, yet challenging, tasks of social surveys for decades. Occupation coding, traditionally performed manually and post-interview, has been acknowledged as time consuming, costly and error-prone. The challenges of manual occupation coding have led to innovations in the use of computer-aided occupation coding, performed by coders or interviewers during or after the interview (Lyberg and Dean 1992). However, with the increasing use of online and mixed-mode surveys, still little is known regarding the feasibility of coding occupation during the interview and factors contributing to the performance of the occupation coding instrument in online and mixed-mode surveys.

[1] Centre for Longitudinal Studies, UCL Institute of Education, 55–59 Gordon Square, London WC1H 0NT, United Kingdom. Emails: d.peycheva@ucl.ac.uk and l.calderwood@ucl.ac.uk
[2] Institute for Employment Research, 104 Regensburger Straße, Nuremberg 90478, Germany. Email: joe.sakshaug@iab.de
Acknowledgments: This project was funded by a grant from the UK Economic and Social Research Council (ESRC) [ES/T00116X/1].

This article addresses these issues by presenting results from a large-scale sequential mixed-mode (web, followed by telephone, then face-to-face) cohort study of young adults in the UK, in which respondents were asked to self-code their occupation online in the initially offered web mode, and interviewers were tasked with identifying the relevant occupation code via telephone or face-to-face when an online interview was not possible.

To our knowledge, use of computer-aided occupation coding during the interview in large-scale probability-based surveys is still rare, and evidence on its performance is not readily accessible in the survey literature. Furthermore, the performance of occupation coding during the interview has not been investigated in a web-first sequential mixed-mode survey, nor the extent to which respondents and interviewers influence the use of the coding instrument. However, the importance of these issues has been acknowledged for future improvements of occupation coding instruments (Belloni et al. 2016; Conrad et al. 2016; Schierholz et al. 2018).

In the following sections, we provide a brief overview of the relevant literature (Section 2), outline the research questions (Section 3) and methods used to address them (Section 4), present the results (Section 5) and discuss the conclusions and practical implications of the findings (Section 6).

2. Literature Review

It is common practice in social surveys to collect information about occupation with a series of open-ended questions asking participants for their job title and to describe the kind of work they do. Such questions enable the collection of sufficient detail about respondents' occupation and assignment of a code at the most detailed level of the occupational classification. These questions may be administered in both interviewer- and self-administered survey settings, and are also asked in mixed-mode surveys. Less frequently, occupation is captured with closed-ended questions offering limited choice of occupational categories, thus resulting in highly aggregated occupation codes. Alternative formats have been offered in web-based surveys, including search tree navigation and semantic text matching techniques, and look-up methods to self-identify one's own occupation. Their application in computer-assisted personal and telephone interviews has also been documented (Tijdens 2014, 2015a). Open-ended occupation questions, however, still dominate the research practice. The verbatim responses collected from these questions are typically converted into occupation codes post-interview by specialist coders using manual or computer-aided (computer-assisted or automated) coding procedures. Today, the use of manual coding has been significantly reduced, but it still complements computer-aided coding methods.

In manual coding, coders assign an occupation code based on the open-ended responses, using a standardized classification scheme without any degree of computer assistance. While classification schemes differ, they all include hundreds of occupation codes nested within hierarchical groups with more specific occupation groups nested within more general groups. For example, the 2010 classification of occupations in the UK, known as the Standard Occupational Classification (SOC2010), and used for occupation coding in the current study, has 9 major (1-digit), 25 sub-major (2-digit), 90 minor (3-digit), and 369 unit (4-digit) (Office for National Statistics 2010a). Manually selecting a code at the most

detailed (4-digit) level of the classification scheme is a time-consuming and expensive process, but also error-prone as even professional coders, following detailed coding guidelines, might disagree on the occupation code for a given case (Lyberg and Dean 1992; Creecy et al. 1992; Campanelli et al. 1997). Manual coding has been found to be especially problematic by National Statistical Offices (NSOs) where coding is extensive (e.g., censuses and large-scale sample surveys), with manual coding error rates of 10% or greater (Lyberg and Dean 1992; Creecy et al. 1992).

Faced with this challenge, the endeavour of computerizing the coding of open-ended responses dates back to the early 1960s, with first applications of computer-aided occupation coding in the late 1970s (Lyberg and Dean 1992; Creecy et al. 1992). Two main forms of computer-aided coding emerged – computer assisted and computer automated. In the former, as in manual coding, the coding is performed by a coder who works interactively with the computer, which guides their decision. In the latter, occupations are coded automatically by software. However, fully automated coding is rare in practice. Rather, it is usually supplemented with manual or computer-assisted systems. Automated coding usually codes part of the occupation entries in which a desired level of certainty associated with the occupation code is set. That is, occupation descriptions with a high degree of certainty (i.e., above a certain threshold) are coded automatically, otherwise human intervention is sought. Even though coder involvement is significantly reduced, human coding is preferred for the so-called "hard-to-code" or "difficult" occupations. This is also known as semi-automated coding, in which automated coding is complemented by human coding for certain situations, compared to fully automated coding, in which 100% of the coding is performed by the software. Computer-aided coding can be administered during or post-data collection, although post-data collection coding is more prevalent.

Computer-aided coding involves computer-stored dictionaries that can be built from coding manuals (e.g., classification schemes) or on empirical patterns of responses provided by respondents in earlier surveys (i.e., previously coded occupation information from previous studies, pilots, etc.), or a combination of manuals and empirical patterns. However, dictionaries constructed from manuals have been considered a disadvantage in that they are strongly dependent on the experience of the coder (e.g., respondents may use words or phrases not included in the manual). In contrast, it has been considered more efficient to base the dictionary on empirical response patterns, in which matching can benefit from the use of words and phrases given by previous respondents (Lyberg and Dean 1992; Creecy et al. 1992).

Various forms of matching have been applied, generally classified as rule-based and data-driven (e.g., statistical or machine learning) (Gweon et al. 2017; Schierholz and Schonlau 2020). For example, if the open-ended answer meets a prespecified logical condition (e.g., presence of a certain word), then a specific code is assigned. Such 'if-then' statements are called 'rules'. Rules are written by experts or can be based on previous data analysis. More recently, statistical learning or machine learning approaches have been employed, whereby a model is trained on manually coded training data and used to predict the most probable code for new data (Gweon et al. 2017; Schierholz et al. 2018; Schierholz and Schonlau 2020).

Evaluations of occupation coding performance between coding methods are scarce, mostly based on comparisons with manual coding used as a 'gold standard', despite the

challenges it poses. Some of these studies have shown that automated coding works reasonably well in reducing the number of cases to be manually coded post-interview, but that it is not yet ready to replace human coders (Ossiander and Milham 2006; Burstyn et al. 2014; Helppie-McFall and Sonnega 2018). Belloni et al. (2016), using the automated coding tool as a benchmark, also stressed the benefits of using automated and human coding jointly.

Campanelli et al. (1997) compared manual coding with computer-assisted and automated coding. They found only a modest gain in performance using computer-assisted compared to manual coding. Automated coding was particularly sensitive to the amount and type of input which was entered, and sometimes scored significantly lower than manual coding with respect to the plausibility of the occupation code. For example, when both the job title and job description were used in the search algorithm, automated coding was comparable to manual coding. However, when the description input was limited and only the job title was used, the quality of coding was substantially lower than manual coding. In terms of time-saving, manual coding and computer-assisted coding did not differ – a result that the authors suggested was due to the fact that the coders were new to the coding software. As expected, automated coding yielded the largest time (and cost) savings, assigning a code in nearly all (99%) of the cases.

The level of detail of the verbal descriptions used in automated coding was also acknowledged by Belloni et al. (2016). The authors used data from the Dutch sample of the Survey of Health, Ageing and Retirement in Europe (SHARE) and compared manually coded verbatim responses on current and last occupation with the codes assigned by the Computer Assisted Structured Coding Tool (CASCOT) occupation coding software. The authors found that automated coding was significantly improved when additional auxiliary information, such as training and qualifications needed for the job and the industry in which the respondent is working, was included in the search algorithm. In contrast, Helppie-McFall and Sonnega (2018) found that the NIOSH Industry and Occupation Computerized Coding System (NIOCCS), employed in the Health and Retirement Survey (HRS) to code occupation history data, and compared to coding results from a highly trained human coder, worked well only with short descriptions, one to three words each, of job title or job description (and industry) as inputs – a finding in line with Conrad et al. (2016).

Conrad et al. (2016) found the length of the occupation description to be a factor strongly related to the reliability of post-interview coded occupations. However, the observed relationship was dependent on the particular occupation terms. For example, for 'easy' occupation terms, longer descriptions were less reliably coded than shorter descriptions, but for 'difficult' occupation terms, longer descriptions were slightly more reliably coded than shorter descriptions. The authors argued that the occupation descriptions do not necessarily need to be long or overly specific, particularly for 'easy' occupations, and that interviewers should rather be trained on the logic and rationale behind the coding structure so that they have a better sense of the kinds of decisions coders need to make. That longer descriptions do not necessarily result in more accurate occupation coding was supported by Massing et al. (2019) who found that reliability decreased as descriptions became longer. Bergmann and Joye (2005) also suggested that the more detailed the information to be coded, the less reliably individual cases are

assigned to categories. Cantor and Esposito (1992) reported that coders who were asked to comment on recordings of interviewers' questioning strategy only rarely indicated that more detailed information would be useful, and some even criticized the fact that interviewers had collected too much information as most of the information needed to code a case comes from the job title, and that interviewers should focus their efforts on obtaining good information there.

Given the potential disconnect between interviewers and coders regarding what constitutes a useful occupation description, as well as the additional costs of post-interview coding, several studies have stressed the potential advantages of computer-assisted occupation coding during the interview, which would eliminate or at least minimize the need for post-interview coding (Campanelli et al. 1997; Conrad et al. 2016; Belloni et al. 2016; Helppie-McFall and Sonnega 2018). These studies cite the potential for a reduction in coding errors as the uncertainties likely to arise in a post-interview coding, due to insufficient or contradictory information provided by respondents, can be resolved by the interviewer. This, in turn, may yield a more parsimonious list of best matching occupations to choose from, and even allow the coding decision to be confirmed by the respondents themselves. If the coding instrument produces a lengthy list (or conversely, an empty list) of likely occupations to choose from, then it is much easier for the interviewer to probe for additional information during the interview than it is for any post-interview intervention to be performed. Although it is typically assumed that interviewers do not achieve the same levels of accuracy as specialist coders, with increasing experience the interviewer may develop a better idea of what constitutes a good occupational description and probe accordingly (Lyberg and Dean 1992; Campanelli et al. 1997; Conrad et al. 2016). Occupation coding during the interview is also expected to reduce costs and processing time since a smaller number of occupation descriptions will require post-interview coding. The method also has low maintenance costs as the code frame and search algorithms can be constructed and updated automatically (Hacking et al. 2006). However, a number of concerns have been raised as well, including the potential for interviewer effects and the challenge of implementing the coding instrument in a web survey, in which respondents perform the entire occupation coding process themselves without the assistance of an interviewer. Furthermore, the potential risks of mode effects when applying the coding method in a mixed-mode survey have been acknowledged elsewhere (Conrad et al. 2016; Tijdens 2014, 2015a; Tijdens and Visintin 2017).

Occupation coding of open-ended descriptions during the interview is typically implemented as a special form of computer-assisted coding in which the computer suggests the most relevant occupation code(s) to the interviewer or an online respondent. It usually follows a two-step approach. In the first step, the interviewer (or respondent) types into the open text field the job title and/or description of the occupational activity. On the basis of this verbatim text, and sometimes other input from the interview, the search engine then shows a list of best matching occupations from the code frame, from which the interviewer (or respondent) selects the most appropriate occupation. As interviewers (or respondents) enter more inputs, the search engine adapts the list to find the best matching occupations and the list of occupations becomes smaller. If the search results do not yield any likely matches, then respondents may be asked to provide further details or the case might be referred to a specialist coder post-interview. As mentioned, various matching

algorithms could be employed. The classical algorithm consults a coding index or look-up table and produces a list of appropriate categories that are identical or similar to the job title and job description information provided by the respondent using rule-based techniques (Hacking et al. 2006; Elias et al. 2014; Tijdens 2015; Tijdens and Visintin 2017; Brugiavini et al. 2017; Belloni et al. 2016; Schierholz et al. 2018; Gweon et al. 2017; Schierholz and Schonlau 2020). A more sophisticated approach uses machine-learning algorithms to identify possible occupation codes from previously coded data, known as training data (Schierholz et al. 2018; Gweon et al. 2017; Schierholz and Schonlau 2020). Benefits of combining algorithms that rely on job titles from a coding index with statistical learning algorithms trained on data from previous surveys have also been documented (Schierholz and Schonlau 2020).

Assessments of occupation coding of open-ended questions during the interview in probability-based sample surveys have been mostly positive. For example, around 80% of all occupations collected in the sixth wave of the Survey of Health Aging and Retirement in Europe (SHARE), conducted in 2015, were coded during face-to-face interviews (Brugiavini et al. 2017). A coding rate of 72% was observed by Schierholz et al. (2018) in a telephone survey in Germany in 2014, commissioned by the Institute of Employment Research (IAB). A field experiment by Statistics Netherlands in September 2003 with the Dutch Labour Force Survey (LFS) and the actual LFS in January 2004, administered face-to-face, achieved interview coding rates of 79% and 75%, respectively (Hacking et al. 2006).

Schierholz et al. (2018) also evaluated data quality by comparing telephone interview coding with office coding performed by two independent professional coders. This resulted in a high level of agreement between the office coding and the interview-coded data, which provided some reassurance that interview coding yields comparable data quality to manual coding. However, higher disagreement rates between office coders were observed for more complex occupation descriptions, to which an occupation code could not be assigned during the interview and the occupation description had to be manually coded. The authors suggest that 'simpler' occupations are more easily codable during the interview, and that more difficult descriptions are more appropriate for office coding.

However, in terms of the time needed to code occupation, concerns have been expressed that occupation coding during the interview may significantly extend interview time. Hacking et al. (2006) reported that the average duration of coding occupation using the look-up table method was 47 seconds compared to 36 seconds when only the open text information (to be coded post-interview) was collected. Schierholz et al. (2018) noted that, for respondents whose occupations were coded successfully during the interview, the duration of the interview was shortened by a few seconds compared to those who did not select one of the suggested categories and were presented with an additional follow-up question. Tijdens (2016) reported a mean time of 48 seconds to code one's occupation, using a semantic matching tool in the Wageindicator web survey, a non-probability survey on work and wages, which was a few seconds longer compared to a search tree navigation also offered on the web platform.

The self-coding of occupation based on open-text descriptions in an online setting is still rare. Insights on its feasibility have been documented for the aforementioned Wageindicator web survey, in which semantic matching, using look-up tables, has been

used to code respondents' occupations since 2015 (Tijdens 2015a). Respondents type their occupation and word-matches in the look-up table are instantly shown to them to select the most relevant match. The semantic matching tool has been noted as being preferred by respondents over search tree navigation offered on the platform alongside. Furthermore, it has been noted as being most suitable for self-coding of occupations provided that the occupational look-up database is sufficiently large (Tijdens 2015b).

3. Research Gaps and Study Questions

Given that the literature on interview-based coding of open-ended occupation descriptions comes primarily from interviewer-administered settings, mainly telephone and face-to-face surveys, the feasibility of coding occupations in probability-based online and mixed-mode surveys involving both self- and interviewer-administered modes remains unclear. To our knowledge some NSOs perform occupation coding for their own labour force surveys online during the interview (e.g., Statistics Denmark and Netherlands), however, evidence on their performance is not readily available.

In web surveys, occupation coding during the interview is challenging as there is no interviewer to provide assistance or probe for additional information. Thus, an easy-to-use interface is needed to facilitate respondent self-coding of occupation. In mixed-mode survey designs, this becomes even more challenging as the interface should be standardized across modes, so that the measurement of occupation is comparable across all respondents.

In this study, we investigate the feasibility of coding open-ended occupation descriptions during the interview by implementing a computer-assisted look-up system in a web-first sequential mixed-mode survey, in which web respondents were asked to self-code their occupation and interviewers in the telephone and face-to-face follow-up modes identified and recorded the relevant occupation code. Following the entry of key words, the coding tool searched for relevant job titles in the Standard Occupational Classification 2010 (SOC2010) coding index and offered a list of corresponding codes. If the look-up method was not successful in identifying a relevant occupation code, then the traditional post-interview occupation coding procedure using an open-ended question to describe the respondent's job tasks was employed.

In addition to the take-up rate of the look-up system, we examine indicators of its performance, such as the time to code occupation during the interview, the specificity of the allocated look-up codes, and the length of the occupational description (only asked as a follow-up question when an occupation code could not be assigned using the look-up system), which are thought to impact the quality of the collected occupation data. This study does not directly assess the quality of the collected occupation information, typically measured by reliability (the extent to which the same occupation code will be repeatedly assigned to the same case) or validity (the accuracy of the assigned code), which is a limitation to be considered in future work. Finally, we assess the extent to which respondents and interviewers influence the use of the new coding method. To our knowledge, no other study has previously investigated these issues in a web-first sequential mixed-mode survey.

Using data from the Next Steps Age 25 (wave 8) cohort study in the UK, we address the following research questions:

1. To what extent do web respondents use the look-up method to self-code their occupation, and how does this compare to interviewer administration of the look-up method in the telephone and face-to-face follow-up modes? Does the rate of self-coding in the web mode vary by device type (PC, laptop, tablet)?

Here, we expect that respondents interviewed via web will use the look-up system at a lower rate than those interviewed by the telephone or face-to-face follow-up modes. The absence of an interviewer is a key disadvantage in this case, as there is no one to motivate the respondent to engage with the coding instrument and provide guidance and/or probe for relevant details to make the look-up task more manageable. Interviewers undergo specific training with the coding system and likely have relevant experience in collecting occupation information; thus, they are more likely to be aware of what constitutes a valid occupation description than respondents. Without the assistance of an interviewer, the look-up task may become more burdensome for respondents, who may expend less effort than an interviewer would to select an occupation code, especially if the occupation is difficult to code.

The burden of self-coding is expected to be correlated with device type and, specifically, the presentation size of the look-up interface. Larger screens (e.g., desktop PCs) are likely to better handle longer look-up lists, improve visibility and limit the amount of burdensome scrolling necessary to identify the most relevant occupation, compared to smaller screens (e.g., laptop and tablet). Thus, we expect the coding rate of the look-up method to be proportional to the relative screen size and, thus, higher with desktops followed by laptops and tablets. We do not assess the use of the look-up method for smartphones, as smartphone participation was strongly discouraged in Next Steps and very few of such cases occurred.

2. Is the performance of the look-up method and characteristics of open-ended occupational descriptions – which are both linked to occupational data quality in previous research – comparable between the three sequential modes (web, telephone, and face-to-face)?

The absence of an interviewer to motivate respondents and probe for relevant details could compromise the quality of occupational coding in web surveys (Conrad et al. 2016). For example, given the cognitively demanding and time-consuming task of self-coding (Tijdens and Visintin 2017), it has been suggested that respondents (or interviewers) will select more generic occupation titles presented in classification lists, which appear to be correct but are suboptimal, rather than exert the necessary effort to choose a more specific occupation code (Schierholz et al. 2018). We thus expect a higher prevalence of generic codes for web respondents compared to respondents who are guided in the interviewer-administered modes. Generic codes refer to suboptimal descriptions or descriptions that are too abstract to be assigned to a more specific category. Such codes have a last digit of '0' or '9' at the 4-digit (unit group) level of the SOC. A last digit of '0' indicates that the 4-digit (unit) group is equivalent to the broader 3-digit (minor) group, or that there is only one unit group within the minor group and the coding could not be achieved at a more detailed level. A last digit of '9' indicates occupations "not elsewhere classified – n.e.c.", thus containing a mix of occupations which are not in sufficient numbers to merit their own unit group.

Furthermore, we expect shorter occupational descriptions to the standard open-ended question (only used as a follow-up if coding was not successful using the look-up method) by respondents on the web, compared to the interviewer-administered modes, and that the look-up procedure and the standard open-ended question will each take longer to administer for web respondents compared to respondents assisted by a trained interviewer in the telephone and face-to-face modes.

3. Do respondent attributes (e.g., sex, ethnicity, education, cohabitation status) and interviewer characteristics (e.g., sex, age, years of interviewing experience) influence the performance of the look-up method during the interview?

Based on the literature, we expect that study members' and interviewers' characteristics will affect whether the look-up method is successful in assigning an occupation code during the interview, or if post-interview coding is required. For example, Belloni et al. (2016) showed that coding errors (namely coding disagreement between manually coded verbatim responses on current and last occupation and codes assigned with the CASCOT software, while the automated coding was taken as a benchmark) were more common for male than for female respondents. For coding 'last job', errors were more likely to occur for the most educated individuals and for the self-employed. Cognitive abilities were found to play an important role in explaining coding errors for 'current job'.

It has been suggested in previous research that coder experience might affect agreement and thus coder information should be included in the analysis of coding quality (Conrad et al. 2016). Schierholz et al. (2018) also acknowledged the potential for interviewer effects on the selection of an occupation code and analyzed the extent to which interviewers correctly applied standardized interviewing techniques that were prescribed for coding occupation using behavioural coding. The authors found that many interviewers did not closely follow the rules for standardized interviews and it was rather an exception that the interviewer read out loud the exact question text and all answer options, including the last option for 'other occupation'. However, the authors noted that when the script was not followed, it was often because the interviewer already had a good understanding of the respondent's job and thus good reasons for departures from it. Nevertheless, the interplay between the interviewer and respondent was acknowledged as an important issue for future improvements of the occupation coding instrument. We therefore consider the inclusion of interviewer covariates as a strength of the present study. However, the direction of their expected impact is less clear.

4. Methods and Data

4.1. Next Steps and the Age 25 Survey

Next Steps follows the lives of 16,000 people in England born in 1989/90, sampled from state and independent schools. The sample design considered schools the primary sampling unit, with deprived schools being over-sampled by 50%. A total of 647 state and independent secondary schools as well as pupil referral units participated in the study out of 892 selected schools. Pupils from minority ethnic groups (Indian, Pakistani, Bangladeshi, Black African, Black Caribbean, and Mixed) were over-sampled to provide

sufficient base sizes for analysis. The school and pupil selection approach ensured that, within a deprivation band and ethnic group, pupils had an equal probability of selection (Department for Education 2011).

The study began in 2004, when the cohort members were aged 14. They were surveyed annually until 2010 (waves 1–7), and then in 2015–2016 when they were aged 25 (wave 8). The interviews for the first four waves were conducted face-to-face, and from wave 5 onwards a sequential mixed mode approach – online, followed by telephone, and then face-to-face interviews – was used. Next Steps has collected information about cohort members' education and employment, economic circumstances, family life, physical and emotional health and wellbeing, social participation and attitudes. A total of 15,531 cohort members were issued to field in the most recent age 25 survey, achieving a response rate of 51% with 7,707 completed interviews (4,797 online, 690 telephone, and 2,220 face-to-face) (Centre for Longitudinal Studies 2017).

4.2. *Occupation Coding System*

Economic activity data has been collected in the study since its initial wave at age 14 and occupation data in particular since study members were aged 16 – that is, when they reached the compulsory school leaving age and were eligible to start an apprenticeship or traineeship, or spend 20 or more hours per week working or volunteering while in part-time education or training. Occupation has since been captured with open-ended questions asking about the title of their job and a description of what they mainly do in their job. This information has been subsequently coded post-interview by professional office coders.

Labor market entry is a key milestone in the cohort's transition from adolescence to young adulthood, and thus work participation was a key theme in the age 25 survey. In this survey, occupation was captured by a text-based search and coding method during the interview. Following an open-ended question about their job title ('What is your current job title?'), respondents in the web survey were asked to enter key words into a search box describing what they mainly do in their job, then select the most appropriate response option from a list of occupations generated by the search system (Figure 1). The input string of words was matched against a concatenated string of the fields describing the job

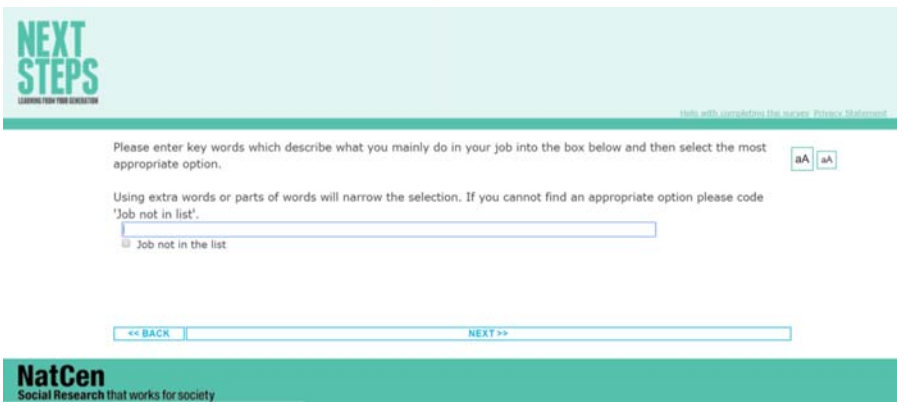

Fig. 1. Screenshot of the look-up question in the web survey.

title in the Standard Occupational Classification 2010 (SOC2010) coding index (containing over 27,000 job titles) as a look-up. This concatenated string included the indexing word (usually the word describing the core set of tasks that characterize a job), occupational, industrial, and additional qualifying terms. The occupational qualifying term is separated from the indexing word by a comma (e.g. teacher, head). Industrial qualifying terms are shown within brackets and can take the form of an industry or branch of industry in which the occupation lies (e.g. teacher, head, (secondary school)). Additional qualifying terms usually indicate the type of material worked with, the machinery used or the processes involved, or can take the form of professional qualifications (Office for National Statistics 2010b). As the indexing word is rarely enough to enable the job title to be correctly coded, the additional qualifying terms aimed to make the search more specific, and, in turn, the coding more accurate. The look-up method used a "word-chunk" search system processing the input string as a string of chunks (comprised of at least three characters) and searching for each chunk in the SOC2010 job title index (i.e., a simple lookup of the job titles containing all of the word chunks). There was no pre-processing (standardization) of the key words entered in preparation for matching, and no amendments were made to the underlying job title list.

Following the entry of the key words, occupation codes at the most detailed 4-digit unit group level of the classification were displayed in alphabetical order. As the SOC2010 index of job titles includes occupation codes with a '0'-ending digit (indicating a single unit group within the minor group) and a '9'-ending digit (indicating 'not else classified'), these respective job titles were displayed if relevant to the search. The procedure was similar in the interviewer-administered telephone and face-to-face modes, except interviewers entered the key words into the search box, read out the list of occupation search results to the respondent, and selected the most appropriate occupation from the list.

A further instruction, aimed at handling long lists of occupations, stated that using extra words or parts of words will narrow down the list of displayed options. For example, just typing in "teacher" would bring up a long list of possible occupations, but entering additional search terms, for example, "teacher secondary", would narrow down the list of options (Figure 2).

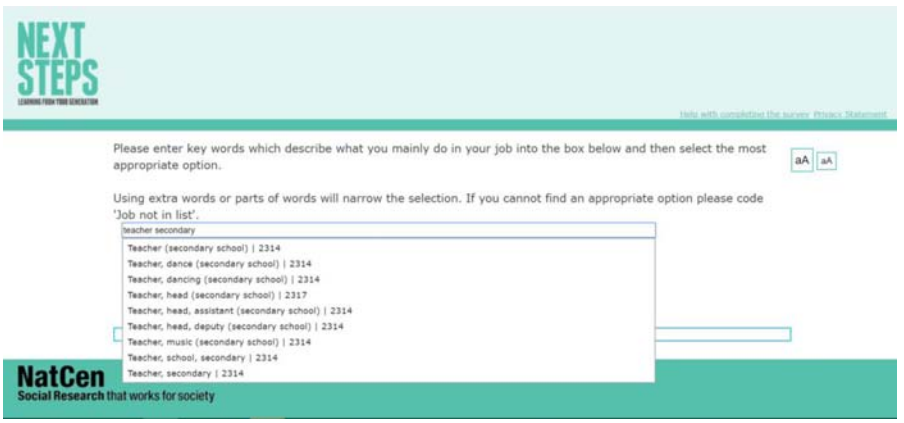

Fig. 2. Screenshot of the look-up question and search results for "teacher secondary" in the web survey.

If an appropriate option was selected during the interview, the associated code from the SOC2010 coding index was automatically assigned, replacing the respondent's entry of key words (required to perform the search). After trying different search terms, if an appropriate option was not found (e.g., no search results were presented, or none of the presented search results were considered appropriate by the respondent or interviewer), then respondents (or interviewers) were asked to select the 'Job not in the list' option and answer a follow-up standard open-ended question describing their job ("Please describe in your own words what you mainly do in your job."). For this question, respondents (or interviewers) were encouraged to provide (or probe for) full details (for example, the type of work) to allow office coders to accurately identify the correct SOC code after the interview.

There was no difference in the design or presentation of the look-up question or standard open-ended job description question between the self- and interviewer-administered modes other than a minimal altering of the question wording and accompanying instructions, so that they were appropriate for either the respondent or interviewer. Telephone and face-to-face interviewer scripts included an additional instruction 'to probe as required'.

4.4. Statistical Analysis

To address the first research question, we report percentages of respondents who were assigned an occupation code during the interview using the look-up coding method and the complementary set that were referred to traditional office coding post-interview. Using chi-squared tests, we show how the look-up rates varied by the sequentially offered modes (online, telephone and face-to-face). We also use chi-squared tests to evaluate differences in the coding rates by self-reported device type: desktop computer, laptop and tablet in the web interviews. A very small number of web respondents ($n = 25$) reported completing the survey on a smartphone, even though this practice was discouraged. These few cases are excluded from all analyses.

To address the second research question, we perform three separate analyses, which we compare across the three modes. First, using chi-squared tests we assess the difference in the prevalence of (suboptimal) generic codes (i.e., codes with a last digit of '0' or '9' at the 4-digit level) compared to more specific occupation codes collected using the look-up method. The second analysis evaluates the length of the open-text descriptions (measured by the total number of characters) provided by respondents to the third open-ended question asked only if an occupation code was not assigned during the interview using the look-up method. We use a Wald test to evaluate differences in the mean description length between the sequentially offered modes. The third analysis assesses the time that was required to code occupation using the look-up method and provide an open-text description in the case of inability to assign an occupation code during the interview. Using a Wald test, we evaluate if the mean time (measured in seconds) differed across the web, telephone, and face-to-face modes.

To address the third research question, we first fit a multivariable logistic regression model on whether or not an occupation code could be assigned during the interview (using the look-up method). We report crude (unadjusted for other characteristics) and adjusted (for all characteristics) odds ratios (OR) and 95% confidence intervals (CI) for the effects of the following respondent characteristics: sex, ethnicity (white/non-white), if ever attended

university and if in a cohabiting relationship by age 25, and survey mode. We further include interactions between each of the respondent characteristics and response mode.

To more fully account for mode-specific effects and differing levels of information available for each mode, including interviewer characteristics in the face-to-face mode, we fit separate logistic regression models on the assignment of an occupation code during the interview for each of the three response modes. A two-level random intercepts model is used for the face-to-face respondents to account for respondents nested within interviewers and to assess the effects of the following interviewer characteristics: sex, age (recoded: 49 years or younger, 50–59, 60–69, 70+ years) and years of experience (recoded: 1 year or less, 2–5 years, 6–9 years, 10 or more years). The intraclass correlation coefficient (ICC) is also reported for the face-to-face model as an approximate measure of the "interviewer effect" or the proportion of variation in the coding outcome attributable to the interviewer level. This analysis is only possible for the face-to-face interviews as interviewer IDs were not recorded for telephone interviews.

We note that Next Steps did not randomly assign cohort members to interviewers. Thus, the reported ICC may reflect both area and interviewer effects which are confounded. The lack of an interpenetrated design is a limitation which we attempt to address by including respondent and interviewer characteristics in the model. Still, it is plausible that further controls are needed to isolate the pure interviewer effect. Thus, we interpret the ICC with caution.

All analyses addressing the three research questions account for the complex sample design of the Next Steps study using the SVY commands in Stata 16.0 and control for selection into each sequential mode using weights, which we describe next. Descriptive statistics for all variables used in the analysis are supplied in online supplemental material Table S1.

4.5. Accounting for Sequential Mode Selection

To evaluate and compare the performance of the look-up method across the three survey modes, it is useful to control for differential nonresponse at each stage of the sequential mixed-mode design. Several methods have been proposed to control for differential nonresponse in mixed-mode surveys (Vannieuwenhuyze and Loosveldt 2013; Vannieuwenhuyze et al. 2014; Klausch et al. 2017). One of the most common methods is to use selection weights that are based on the (estimated) propensity of a respondent to participate in each mode (Hox et al. 2015). The mode propensities are typically estimated from a generalized linear model (e.g., probit, logit), with each mode treated as a possible outcome, conditioning on available covariate information which, in longitudinal studies, is often limited to fixed baseline characteristics (e.g., demographics) and/or data collected from the previous wave.

We adopt and extend this method by applying a data-driven nonresponse weighting procedure. Instead of using only baseline covariates or covariates collected only in the previous wave of Next Steps, we use covariates selected from all (seven) previous waves to adjust for mode selection in the Next Steps age 25 (wave 8) survey. A two-step approach was implemented separately and sequentially for each of the three response modes, starting with web followed by telephone and then face-to-face. In the first step, seven

multivariable log-binomial regressions predicting nonresponse at wave 8 were fitted, each regression containing only predictor variables collected from one of the seven prior waves of Next Steps. All statistically significant ($p < 0.05$) predictors were retained for the second step of the procedure.

In the second step, all of the variables retained from the first step were imputed to produce a complete dataset of predictors. These wave-specific predictors then entered into a series of log-binomial regression models predicting nonresponse at wave 8, each model building on the previous one by incorporating additional variables from the subsequent wave. For example, the first model of nonresponse at wave 8 included only predictors from wave 1, then wave 2 predictors were added into the next model, and so on. After introducing a given set of wave-specific predictors, these predictors were checked for statistical significance ($p < 0.05$). If a current-wave predictor was no longer statistically significant after controlling for the predictors from past waves, it was dropped from the model. Only predictors which remained significant after controlling for predictors from the past waves were retained. This was done to maintain the temporal sequence of the predictors available in the longitudinal data.

All retained variables (shown in online supplemental material Tables S2–S4) were then used to create propensity score adjustment weights for mode-specific unit nonresponse. The propensity to respond at each stage of the sequential mixed-mode design (web, telephone, and face-to-face) was calculated separately for each sample unit. The estimated propensity scores were then sorted into quintiles. The nonresponse adjustment weight was then calculated as the inverse of the average propensity score in each quintile. This process was performed separately for each response mode in their sequential order, yielding three nonresponse adjustment weights. The final analysis weights were then computed as the product of the Next Steps base weight and the nonresponse adjustment weights generated from the above procedure.

5. Results

5.1. Look-Up Method and Office Coding Rates

The occupation coding rates are presented in Table 1. First, we report the percentage of respondents who were assigned an occupation code during the interview using the look-up coding method. Overall, across the three modes, 82.0% of respondents were successfully assigned an occupation code using the look-up method. The remaining respondents who were not assigned an occupation code during the interview were either assigned one by an office

Table 1. Occupation coding rates by survey mode.

Occupation coding method	Web % (n)	Telephone % (n)	Face-to-face % (n)	Total % (n)
Look-up coding	90.3 (3,580)	90.6 (493)	69.2 (1,180)	82.0 (5,253)
'Successful' office coding	8.7 (356)	8.1 (50)	30.5 (483)	17.4 (889)
'Unsuccessful' office coding	0.3 (9)	0.8 (3)	0.0 (1)	0.2 (13)
Refused	0.7 (30)	0.5 (3)	0.2 (7)	0.5 (40)
Total	100 (3,975)	100 (549)	100 (1,671)	100 (6,195)

Notes: Occupation coding outcome vs. survey mode: $\chi^2_6 = 508.8212$, $p < 0.000$. Results are weighted to account for selection into each of the respective sequentially administered modes.

coder after the interview (17.4%), or lacked sufficient information to be coded due to failure to answer the occupation questions (0.5%) or inability of the office coder to identify an appropriate code (0.2%).

The percentage of respondents who successfully used the look-up method varied significantly by mode (Table 1). Contrary to our expectation, look-up rates in the web mode were among the highest with 90.3% of respondents able to self-code their occupation during the interview. This was comparable to the telephone mode, in which 90.6% of respondents were coded during the interview (web versus telephone: $\chi^2_3 = 5.1816$, $p = 0.3679$). The look-up rate was significantly lower among face-to-face respondents, who were assigned an occupation code in only 69.2% of interviews (web vs. face-to-face: $\chi^2_3 = 450.4548$, $p < 0.000$; telephone versus face-to-face: $\chi^2_3 = 82.3264$, $p < 0.000$).

The assignment of an occupation code with the look-up method or with standard office coding varied slightly by (major 1-digit) occupation groups. Namely, occupations in Major Group 8 (Process, plant and machine operatives) appeared to be coded at a higher rate post-interview than during the interview, whereas occupations in Major Group 2 (Professional occupations) were more often coded using the look-up method (results not shown). All other occupation groups were assigned at similar rates by coding method.

Coding rates by device type are reported for the web respondents in Table 2. Look-up coding rates were slightly higher for desktop computers (93.1%) compared to devices with likely smaller screens, including laptops (92.1%) and tablets (90.9%). This pattern followed our expectations; however, the overall differences are not statistically significant ($\chi^2_2 = 3.06$, $p = 0.541$).

5.2. Other Performance Measures

Next, we assess several other indicators of performance for the look-up method and the standard occupation question (used in post-interview coding) likely to impact the collected occupation data. First, we examine the prevalence of generic vs. specific codes, then we look at the length of open-text descriptions to the open-ended occupation question (if the look-up method was unsuccessful). Lastly, we evaluate the mean time it took respondents and interviewers to select an appropriate occupation using the look-up method and answering the open-ended occupation question (if applicable).

5.2.1. Specific Versus Generic Codes

The allocation of an occupation code does not necessarily mean that an optimal code has been found, since some occupation codes refer to more general or abstract occupations

Table 2. Occupation coding rates by self-reported device type in web interviews.

Coding method	Desktop % (n)	Laptop % (n)	Tablet % (n)	Overall % (n)
Look-up coding	93.1 (691)	92.1 (1,818)	90.9 (954)	91.9 (3,463)
Office coding	6.9 (54)	7.9 (156)	9.1 (116)	8.1 (326)
Overall	100 (745)	100 (1,974)	100 (1,070)	100 (3,789)

Notes: Refusals (n = 30) and non-coded post-interview cases (n = 9) are excluded. Cases with missing information on device type (n = 4) and cases completed on a smartphone or other device (n = 25) are also excluded. A further 118 cases assigned a non-applicable code for self-reported device type were also excluded. Results are weighted to account for selection into the initially administered web mode.

compared to those with more detailed or specific occupation meanings. If a general occupation code is selected, when a specific code exists, this may result in misclassification and contribute to bias in the analysis of occupation data. Of course, there are legitimate reasons why a general occupation code may be chosen rather than a more specific one, not driven by lack of respondent engagement; for example, if a specific code doesn't exist in the SOC for a person's occupation, or if the occupation is not prevalent enough in the population to warrant its own category. However, assessing the validity of the assigned codes is beyond the scope of this analysis and here we focus on the prevalence of specific and generic occupation codes across response modes and coding methods.

Percentages of respondents assigned specific and generic (4-digit) occupation codes during the interview (using the look-up method) and via post-interview office coding, by response mode, are presented in Table 3. Overall, we find no statistically significant differences in the assignment of specific or generic occupation codes by response mode for either look-up coding ($\chi^2_4 = 5.63$, $p = 0.571$) or office coding ($\chi^2_4 = 6.50$, $p = 0.157$). About 85% of web respondents selected a 'specific' code (not a '0' or '9' last digit) using the look-up method. Contrary to our expectation, this was comparable to the percentages in the telephone (82.8%) and face-to-face (84.6%) modes. About 84.2% of web respondents referred to post-interview office coding were assigned a specific code, which was also comparable to the telephone (88.9%) and face-to-face (79.5%) respondents.

5.2.2. Length of Open-Text Occupation Descriptions for Post-Interview Coding

Previous research on the detail of the information collected via standard open-ended occupation questions and their usefulness for successful office coding is mixed. For example, longer descriptions were found by Massing et al. (2019), Helppie-Mcfall and Sonnega (2018), Conrad et al. (2016), and Bergmann and Joye (2005) to be less reliably coded than shorter ones, and the additional information (either in longer descriptions or through additional questions or probes) was found to be associated with lower levels of coder agreement. However, it was also observed that this effect was stronger for particular occupation terms (Conrad et al. 2016). Belloni et al. (2016), on the other hand, highlighted the importance of auxiliary information, which substantially increased the level of detail (i.e., number of digits) at which the observations were coded. Similarly, Campanelli et al. (1997) noted that the combination of both the job title and job description used in automated coding led to results that were comparable to manual coding, which was used as a benchmark.

Our analysis does not provide evidence about the accuracy or reliability of coding based on the length of the occupation description collected from respondents who were not assigned an occupation code during the interview, other than the fact that almost all descriptions collected with the standard open-text question, irrespective of mode, were successfully coded post-interview. Rather, our analysis compares the length of the occupation descriptions provided in the self-administered web mode to the two interviewer-administered modes. It has been previously hypothesized that obtaining detailed occupation descriptions from respondents is more challenging in passive self-administered modes because there is no interviewer to probe for more information. Thus, there is a higher risk that occupation descriptions supplied in self-administered modes will be shorter and less sufficient in terms of detail (Conrad et al. 2016).

Table 3. Percentages and 95% confidence intervals (CI) of specific and generic occupation (4-digit) codes by coding method and survey mode.

Occupation code	Coding method							
	Look-up % (95% CI)				Office coded % (95% CI)			
	Overall	Web	Telephone	Face-to-face	Overall	Web	Telephone	Face-to-face
0-ending	5.3 (4.4-6.3)	4.7 (3.8-5.9)	6.2 (3.9-9.9)	6.0 (4.4-8.1)	5.7 (3.9-8.5)	3.3 (1.9-5.6)	0.4 (0.1-3.0)	6.9 (4.4-10.9)
9-ending	10.0 (9.0-11.1)	10.3 (9.0-11.6)	11.0 (7.4-16.2)	9.4 (7.6-11.5)	13.2 (10.3-16.6)	12.5 (0.8-19.2)	10.6 (4.4-23.5)	13.53 (10.1-17.9)
Specific (not ending with '0' or '9')	84.7 (83.3-86.0)	85.0 (83.4-86.6)	82.8 (76.9-87.5)	84.6 (82.0-86.9)	81.1 (77.1-84.5)	84.2 (77.7-89.2)	88.9 (76.2-95.3)	79.5 (74.3-83.9)

Note: Results are weighted to account for selection into each of the respective sequentially administered modes.

On the contrary, we find that in fact web respondents provided, on average, longer descriptions (62 characters) to the open-text occupational question than their telephone and face-to-face counterparts (both 45 characters). The differences between the web mode and the interviewer-administered modes were statistically significant at the 0.05 level (web vs. telephone: $F_{1,397} = 3.92$, $p = 0.048$; web vs. face-to-face: $F_{1,397} = 10.11$, $p = 0.002$). No significant difference was found between the telephone and face-to-face modes ($F_{1,397} = 0.00$, $p = 0.981$).

5.2.3. Time to Select An Occupation or Provide An Occupation Description

Although the primary aim of the look-up method is to increase the accuracy and cost efficiency of occupation coding, it has been emphasized in previous research that, if interview coding is to replace office coding, then it is important that the procedure does not significantly increase interview duration, as longer interviews are more expensive and burdensome for participants (Schierholz et al. 2018). Moreover, gains in cost efficiency due to the reduction of manual coding may be partly or completely offset if interview duration increases significantly (Hacking et al. 2006).

Overall, across the three modes, the look-up method used in Next Steps took, on average, 43 seconds regardless of whether an occupation category was chosen. This varied significantly by mode and took longer in web (47 seconds) compared to telephone (35 seconds) and face-to-face (34 seconds) (web vs. telephone: $F_{1,616} = 42.10$, $p < 0.000$; web vs. face-to-face: $F_{1,616} = 79.81$, $p < 0.000$). The difference between telephone and face-to-face was not statistically significant (telephone vs. face-to-face: $F_{1,616} = 0.36$, $p = 0.550$).

For those respondents who could not be assigned an occupation code using the look-up method and who had to describe their occupations in the follow-up open-text question, it took on average an additional 38 seconds to answer this question. The time to write an occupation description in the open-text question also varied by mode and took longer in web (41 seconds), followed by face-to-face (35 seconds) and telephone (27 seconds). The difference between web and telephone is statistically significant ($F_{1,334} = 7.80$, $p = 0.001$), whereas the differences between web and face-to-face ($F_{1,334} = 3.38$, $p = 0.067$) and telephone and face-to-face ($F_{1,334} = 2.86$, $p = 0.092$) are both marginally significant.

We note that our timing analysis is performed on a reduced sample due to loss of timing information in 11% of the Next Steps face-to-face interviews (Centre for Longitudinal Studies 2017). It also excludes outliers, defined as observations above the 99th percentile.

5.3. Correlates of Occupation Coding During the Interview

We next explore the extent to which study members' characteristics and interviewer attributes (observed only for the face-to-face interviews) influenced whether or not an occupation code could be assigned during the interview (via the look-up method). We look at the effects of these characteristics separately for each of the three sequential response modes (i.e., by fitting mode-specific models) to allow for the different numbers of respondent- and interviewer-level characteristics to vary across the different modes, which would otherwise be obscured in a single combined model restricted to attributes observed in all three modes. Nevertheless, a combined model with these restrictions is presented in

online supplemental material, Table S5, for the interested reader. However, only the mode-specific models are presented and interpreted below.

Table 4 shows the crude (unadjusted for other characteristics) and adjusted odds ratios reflecting the association between each of the respondents' (and, if available, interviewers') characteristics and successful use of the look-up method to assign an occupation code during the interview, presented separately for each response mode.

5.3.1. Online Model

Starting with the online model (Table 4, column A), we find that successful coding of occupation during the web interview (via the look-up method) is related to respondents' ethnic background, whether or not they attended university by age 25, and whether or not they are in a cohabiting relationship at age 25. White study members had 2.2 times higher odds of being coded during the interview compared to non-white study members (OR = 2.20, 95% CI: 1.44-3.36). Those who attended university by age 25 were 1.7 times more likely to be assigned an occupation code during the interview compared to those who have not attended university by that age (OR = 1.70, 95% CI: 1.16-2.50). And those living with a partner, at age 25, were 1.6 times more likely to be assigned an occupation code during the interview compared to those not living with a partner. There was no evidence that study members' sex or the device type used to complete the web survey were related to successful use of the look-up method, after accounting for all other characteristics of interest.

5.3.2. Telephone Model

Continuing with the telephone model (Table 4, column B), we find no strong evidence that respondents' demographic characteristics influenced the likelihood of being assigned an occupation code during the interview. Respondents' sex, ethnic background, cohabitation status and whether or not they have attended university by age 25 were not associated with receiving an occupation code during the telephone interview.

5.3.3. Face-to-Face Model

For the face-to-face model (Table 4, column C), there was also no evidence that study members' demographics influenced their likelihood of being assigned an occupation code during the face-to-face interview, apart from their cohabitation status. As in the web interviews, participants in a cohabiting relationship were more likely to receive an occupation code during the interview (OR = 1.67, 95% CI: 1.08-2.57). However, as expected, there was strong evidence that interviewers influenced whether an occupation code was assigned during the face-to-face interviews. In particular, males, older and less experienced interviewers were less likely to successfully assign an occupation code to study members during the interview using the look-up method. Compared to males, female interviewers had over three times higher odds of assigning an occupation code during the interview (OR = 3.19, 95% CI: 1.92-5.31). The odds of assigning an occupation code notably decreased with increasing age of the interviewer. Compared to interviewers with less than a year of experience in the fieldwork agency, those with experience between two and ten years were considerably more likely to successfully use the look-up method to assign an occupation code. Interviewers with over ten years of

Table 4. Odds ratios (OR) and 95% confidence intervals (CI) of successful occupation coding during the interview (via the look-up method) on respondent and interviewer characteristics, by survey mode

	(A) Web (n = 3,789)		(B) Telephone (n = 543)		(C) Face-to-face (n = 1,576)	
	Crude OR (95% CI)	Adjusted OR (95% CI)	Crude OR (95% CI)	Adjusted OR (95% CI)	Crude OR (95% CI)	Adjusted OR (95% CI)
Respondent characteristics						
Female (REF: Male)	0.87 (0.60-1.24)	0.93 (0.65-1.34)	0.66 (0.34-1.28)	0.68 (0.36-1.30)	0.92 (0.0.69-1.23)	0.73 (0.45-1.19)
White (REF: Non-white)	2.09*** (1.43-3.08)	2.20*** (1.44-3.36)	0.83 (0.42-1.64)	0.67 (0.32-1.42)	0.89 (0.64-1.22)	0.84 (0.48-1.48)
Ever attended university (REF: No)	1.53* (1.08-2.17)	1.70** (1.16-2.50)	0.60 (0.30-1.19)	0.61 (0.31-1.21)	0.84 (0.62-1.15)	0.67 (0.40-1.13)
Cohabiting (REF: No)	1.76** (1.21-2.54)	1.55* (1.06-2.25)	1.07 (0.52-2.20)	1.17 (0.56-2.41)	1.34 (1.02-1.77)	1.67* (1.08-2.57)
Laptop (REF: Desktop)	0.91 (0.66-1.26)	0.81 (0.45-1.46)	–	–	–	–
Tablet (REF: Desktop)	0.65* (0.46-0.91)	0.71 (0.41-1.24)	–	–	–	–
Interviewer characteristics						
Female (REF: Male)	–	–	–	–	4.08*** (2.46-6.78)	3.19*** (1.92-5.31)
Age: 50-59 years (REF: 49 or younger)	–	–	–	–	0.10*** (0.04-0.29)	0.12*** (0.04-0.36)
Age: 60-69 years (REF: 49 or younger)	–	–	–	–	0.05*** (0.02-0.14)	0.08*** (0.03-0.24)
Age: 70+ years (REF: 49 or younger)	–	–	–	–	0.01*** (0.00-0.04)	0.02*** (0.01-0.07)
Experience: 2-5 years (REF: 1 year or less)	–	–	–	–	3.35** (1.36-8.26)	3.50* (1.32-9.30)
Experience: 6-9 years (REF: 1 year or less)	–	–	–	–	1.22 (0.57-2.59)	2.60* (1.15-5.87)
Experience: 10+ years (REF: 1 year or less)	–	–	–	–	0.25*** (0.11-0.54)	0.68 (0.31-1.50)

Notes: *** $p < 0.001$, ** $p < 60.01$, * $p < 0.05$, + $p < 0.1$. Results are weighted to account for selection into each of the respective sequentially administered modes.

experience were less likely compared to those with less than a year of experience to code respondents during the interview, although this effect was not statistically significant (OR = 0.68, 95% CI: 0.31-1.50). Further to the above observations, the estimated ICC (rho = 0.67) showed that 67% of the variability in the look-up coding outcome, after accounting for interviewers' and participants' characteristics, was due to variability between interviewers.

6. Discussion

This study illustrated the feasibility of coding occupations during the interview using a standardized coding look-up system implemented in a large-scale sequential mixed-mode (web, telephone, face-to-face) survey of young adults in the UK. Occupation coding is particularly important in this population in the stages of transition into the labor market. To our knowledge, this is the first study to present findings on the feasibility of occupation coding during the interview in a mixed-mode survey, in which the occupation coding procedure is performed by respondents themselves in the first offered web mode, and by interviewers in the follow-up telephone and face-to-face modes. The design and implementation of occupation coding during the interview is challenging in online surveys, in the absence of an interviewer to guide or motivate respondents to perform the self-coding. However, these are even more challenging in mixed-mode studies, for which standardization of the measurement of occupation is desired to the maximum extent possible to ensure comparability of the collected data across modes.

The study yielded five main findings. First, the look-up coding method was considered highly effective as 82% of all respondents were assigned an occupation code during the interview. This is rather similar to the coding rates achieved in previous studies (Hacking et al. 2006; Brugiavini et al. 2017; Schierholz et al. 2018), acknowledging the differences in study populations, designs and occupation coding frames. This result suggests high potential for cost savings as only the remaining 18% of respondents required subsequent office coding.

Second, the success of the look-up coding method varied significantly by survey mode. It achieved a rate of about 90% in the web and telephone interviews, and about 70% in the face-to-face interviews. This finding contradicted our expectation that the look-up coding method would perform better in the interviewer-administered modes compared to the self-administered web mode. Nevertheless, it is a promising finding and suggests that respondents are not overly burdened with the task of looking up and assigning an occupation code to themselves. This is also a particularly timely finding, as web surveys are becoming more popular in survey research and established interviewer-administered surveys are increasingly transitioning to more online and mixed-mode data collection, as Next Steps has done since wave 5 onward.

Third, despite concerns that coding occupational descriptions may perform differently and sub-optimally in a web survey compared to interviewer-administered modes, we found them to be comparable across both mode types. There was no difference in the proportion of generic occupation codes (i.e., codes with last digit '0' or '9' at the 4-digit level of SOC) assigned during the interview or in post-interview office coding across the different survey modes. However, some occupations were more likely to be coded using the look-up method,

such as Professional occupations, than others, namely Process, plant and machine operatives. Almost all occupation descriptions captured with the open-text question (due to failure of the look-up method) were successfully office-coded across all modes.

Fourth, web respondents were about 11 and 13 seconds slower, on average, in using the look-up method to identify an appropriate occupation code compared to telephone and face-to-face respondents, respectively. Web respondents also took longer to describe their jobs if they could not assign themselves an occupation code using the look-up method: 16 seconds longer than telephone and seven seconds longer than face-to-face respondents. However, the longer online durations led to longer descriptions (about 17 more characters, on average) compared to those recorded in the telephone and face-to-face interviews. Again, these observations give positive insights into implementing occupation coding in self-administered surveys; namely, that self-coding of occupation does not appear to substantially extend the interview duration and respondents who are unable to self-code themselves tend to enter more details in the open-text form compared to interviewers. The difference in the timing to code occupation with the look-up method or write a detailed occupation description was expected as web respondents may require more time to read, comprehend and respond to the requests, than respondents who are assisted by a trained interviewer who is already familiar with the procedure. This, however, could also be a result of less time pressure on the web.

Lastly, we found that both study members' and face-to-face interviewers' characteristics influenced whether an occupation code was assigned during the interview, even after accounting for selection into each sequential mode. Study members' ethnic background, university participation and cohabitation status affected occupation coding during the online interviews; while there was no evidence that these characteristics, apart from cohabitation status, influenced occupation coding during the telephone and face-to-face interviews. We observed a notable interviewer effect in the face-to-face interviews: interviewers' sex, age and years of interviewing experience strongly impacted the likelihood of assigning an occupation code during the interview, as males, older and less experienced interviewers were less likely to succeed in assigning a code using the look-up method.

The effect of interviewers in the face-to-face survey is concerning and raises the question of whether the look-up coding method used during the interview may be more burdensome for interviewers to administer to respondents than for respondents to administer to themselves in a self-administered setting. However, use of the look-up method did not appear to be as problematic for telephone interviewers who performed their interviews from a centralized telephone unit under continuous monitoring by supervisors, suggesting that the higher coding rates in the telephone interviews may have been influenced by the tighter level of control and supervision of the interviewers. Although unstandardized interviewer behavior does not necessarily have a negative impact on data quality (Schierholz et al. 2018), for those interviewers for whom the look-up method was less successful, further training and supervision may be needed. The fact that all interviewer-collected descriptions were successfully coded post-interview raises the question of whether these interviewers invested the necessary effort in using the look-up method to assign an occupation code, as opposed to reverting to their prior experience and habits of collecting occupation information using standard open-ended questions only.

In our view, this suggests that monitoring of interviewer performance and more specialized training on the use of within-interview coding methods is needed to improve their successful application. The training should make clear to interviewers the benefits of using the new coding system (e.g., processing time reduction, respondent confirmation of assigned code, etc.), while monitoring should be used as part of a feedback loop to continuously improve the application of the coding system during the field period.

Our findings support the existing research that occupation coding during the interview reduces the need for (and associated costs of) post-interview office coding. This is because the large majority of occupations could be coded using the look-up method in each of the three survey modes. However, this finding should be weighed against the added costs of increasing the length of the interview, which could be valued differently depending on the mode of administration and other survey constraints. Furthermore, when the look-up method was not successful, almost all occupation descriptions collected via the open-text question were successfully coded post-interview, which is indicative of the quality of the verbatim information provided by respondents, including that which the web respondents provided without interviewer assistance. This is encouraging as previous research has noted that "the largest source of error lies in shortcomings of the verbatim raw material," as opposed to errors resulting from coding (Hoffmann et al. 1995, 13). It is also positive that web survey respondents provided lengthier descriptions, on average, than the descriptions recorded by interviewers in telephone and face-to-face modes, which is indicative of respondents' engagement.

It has been hypothesized by Conrad et al. (2016) that, as writing or typing requires more effort than speaking for most people, it could be the case that occupation descriptions might be shorter in self-administered (visual) modes, and flagged this as an area – especially with the growth of online surveys – that warrants further study. The fact that we find the opposite effect – that web respondents offer longer descriptions – is reassuring, particularly for the more complex occupations which respondents are unable to locate using the look-up method. As there is no interviewer to probe for more specific information, providing longer descriptions during the interview to facilitate post-interview coding may be more useful than shorter descriptions. However, acknowledging that it may not necessarily be the length of the description that leads to an optimal occupation code (Conrad et al. 2016), survey designers may benefit from offering more specific instructions (including examples) to respondents and interviewers about what constitutes a good occupational description. In addition to more specific instructions, survey designers may consider following up on suboptimal or generic coding (i.e., allocation of a code ending '0' and '9') during the interview with an open-text question for more details. This could potentially enable allocation of a more specific occupation code post-interview with the generic code used as a starting point.

This research complements the existing literature with evidence about the feasibility and effectiveness of occupation coding during the interview in a large-scale, probability-based online and mixed-mode survey. It also provides insights on the performance of coding during the interview and the characteristics of the provided occupation descriptions which are likely to impact occupation data quality. To our knowledge, such an assessment has not surfaced in the research literature. Our work also identifies respondent and interviewer factors that affect the performance of the coding method during the interview and suggests ways for improvement.

There are, however, limitations that could be addressed in future work. Next Steps uses a sequential mixed-mode design, which makes it difficult to remove selection mode effects from measurement mode effects. We addressed this limitation by performing an extensive data-driven "back-door" weighting procedure utilizing seven waves of Next Steps data and numerous covariates to adjust for selection into each phase of the sequential mode design. However, there is still the potential that some factors influencing selection into mode were unaccounted for in the weighting procedure. Another limitation is that this study was performed on a panel population in its eighth wave of data collection. This population is likely to be more cooperative and perhaps more patient in engaging with the occupation coding system, than a freshly recruited sample of the general population. Nonetheless, it is reassuring that the look-up rates observed here were comparable to those observed in other studies (Hacking et al. 2006; Brugiavini et al. 2017; Schierholz et al. 2018). Furthermore, the study lacked relevant pieces of information that would provide further insights on the application of the coding method, including characteristics of the telephone interviewers, the length and content of the terms entered into the look-up search box, as well as change-logs to view the iterative process that respondents and interviewers undertook to identify an appropriate occupation code. Lastly, this study did not directly assess the quality (e.g., validity, reliability) of the occupation codes assigned using the look-up method, as the allocation of an occupation code does not necessarily imply that the optimal code was assigned. We plan to address these issues in future rounds of Next Steps and encourage future studies and other survey institutions to consider them as well. Future work is also needed on developing a theoretical framework for occupation coding.

7. References

Belloni, M., A. Brugiavini, E. Meschi, and K. Tijdens. 2016. "Measuring and detecting errors in occupational coding: an analysis of SHARE data." *Journal of Official Statistics*, 32(4): 917–945. DOI: https://doi.org/10.1515/jos-2016-0049.

Bergmann, M.M., and D. Joye. 2005. "Comparing Social Stratification Schemata: CAMSIS, CSP-CH, Goldthorpe, ISCO-88, Treiman, and Wright." *Cambridge Studies in Social Research* 10: 1–35. Available at: https://www.sociology.cam.ac.uk/system/-files/documents/cs10.pdf (accessed October 2019).

Brugiavini, A., M. Belloni, R.E. Buia, and M. Martens. 2017. *The "Job Coder"*. In SHARE Wave 6: Panel innovations and collecting Dried Blood Spots. Edited by F. Malter and A. Börsch-Supan. Munich: MEA, Max Planck Institute for Social Law and Social Policy: 51–70. Available at: http://www.share-project.org/uploads/tx_sharepublications/ 201804_SHARE-WAVE-6_MFRB.pdf (accessed October 2019).

Burstyn, I., A. Slutsky, D.G. Lee, A.B. Singer, Y. An, and Y.L. Michael. 2014. "Beyond Crosswalks: Reliability of Exposure Assessment Following Automated Coding of FreeText Job Descriptions for Occupational Epidemiology." *The Annals of Occupational Hygiene* 58(4): 482–492. DOI: https://doi.org/10.1093/annhyg/meu006.

Campanelli, P., K. Thompson, N. Moon, and T. Staples. 1997. "The Quality of Occupational Coding in the United Kingdom." In *Survey Measurement and Process Quality*. Edited by L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin: 437–453. New York: Wiley.

Cantor, D., and J.L. Esposito. 1992. "Evaluating Interviewer Style for Collecting Industry and Occupation Information." In Proceedings of the Section on Survey Methods, American Statistical Association: 661–666. Available at https://www.bls.gov/osmr/research-papers/1992/pdf/cp920010.pdf (accessed March 2021).

Centre for Longitudinal Studies. 2017. *Next Steps Age 25 Survey*. Technical Report. University College London. Available at: http://doc.ukdataservice.ac.uk/doc/5545/mrdoc/pdf/5545age_25_technical_report.pdf (accessed October 2019).

Conrad, F., M. Couper, and J.W. Sakshaug. 2016. "Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes." *Journal of Official Statistics* 32(1): 75–92. DOI: https://doi.org/10.1515/jos-2016-0003.

Creecy, R.H., B.M. Masand, S.J. Smith, and D.L. Waltz. 1992. "Trading MIPS and memory for knowledge engineering". *Communications of the ACM* 35(8): 48–64. DOI: https://doi.org/10.1145/135226.135228.

Department for Education. 2011. *LSYPE User Guide to the Datasets: Wave 1 to Wave 7*. Available at: http://doc.ukdataservice.ac.uk/doc/5545/mrdoc/pdf/5545lsype_user_guide_wave_1_to_wave_7.pdf (accessed October 2019).

Elias, P., M. Birch, and R. Ellison. 2014. *CASCOT International version 5. User Guide*. Institute for Employment Research, University of Warwick, Coventry. Available at: https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/cascot_international_user_guide.pptx (accessed October 2019).

Gweon H., M. Schonlau, L. Kaczmirek, M. Blohm, and S. Steiner. 2017. "Three Methods for Occupation Coding Based on Statistical Learning." *Journal of Official Statistics* 33(1): 101–122. DOI: http://dx.doi.org/10.1515/JOS-2017-0006.

Hacking, W., J. Michiels, and S. Jansen, S. 2006. "Computer Assisted Coding by Interviewers." In Proceedings of the 10th International Blaise Users Conference, IBUC 2006, 9–12 May, Arnhem, The Netherlands. Available at: http://blaiseusers.org/2006/Papers/291.pdf (accessed October 2019).

Helppie-McFall, B. and A. Sonnega. 2018. *Feasibility and Reliability of Automated Coding of Occupation in the Health and Retirement Study*. Ann Arbor MI: University of Michigan Retirement Research Center (WP 2018-392). Available at: https://mrdrc.isr.umich.edu/publications/papers/pdf/wp392.pdf (accessed October 2020).

Hoffman, E. 1995. *What Kind of Work Do You Do? Data collection and processing strategies when measuring "occupation" for statistical surveys and administrative records.* ILO. (WP 1995: 95-1). Available at: https://www.ilo.org/wcmsp5/groups/public/--dgreports/--stat/documents/publication/wcms_087880.pdf (accessed October 2019).

Hox, J.J., E.D. De Leeuw, and E.A. Zijlmans. 2015. "Measurement Equivalence in Mixed Mode Surveys." *Frontiers in Psychology* 6(87): 1–11. DOI: https://doi.org/10.3389/fpsyg.2015.00087.

Klausch, T., B. Schouten, and J.J. Hox .2017. "Evaluating Bias of Sequential Mixed-Mode Designs against Benchmark Surveys," *Sociological Methods and Research* 46(3): 456–489. DOI: https://doi.org/10.1177/0049124115585362.

Lyberg, L., and P. Dean. 1992. *Automated Coding of Survey Responses: An International Review*. R&D Reports (1992–2). Statistics Sweden, Stockholm, Sweden. Available at:

https://www.scb.se/contentassets/7c4edb581f8745e3a081e1ba9b332eb4/rnd-report-1992-02-green.pdf (accessed October 2019).

Massing, N., M. Wasmer, C. Wolf, and C. Zuell. 2019. "How Standardized is Occupational Coding? A Comparison of Results from Different Coding Agencies in Germany." *Journal of Official Statistics* 35(1): 167–187. DOI: http://dx.doi.org/10.2478/JOS-2019-0008.

Office for National Statistics. 2010a. *Standard Occupational Classification 2010 Volume 1 Structure and descriptions of unit groups.* Available at: https://www.ons.gov.uk/-methodology/classificationsandstandards/standardoccupationalclassificationsoc/-soc2010/soc2010volume1structureanddescriptionsofunitgroups (accessed October 2019).

Office for National Statistics. 2010b. *Standard Occupational Classification 2010 Volume 2: the structure and coding index.* Available at: https://www.https://www.ons.gov.uk/-methodology/classificationsandstandards/standardoccupationalclassificationsoc/-soc2010/soc2010volume2thestructureandcodingindex (accessed October 2020).

Ossiander, E.M., and S. Milham. 2006. "A computer system for coding occupation." *American Journal of Industrial Medicine* 49: 854–857. DOI: https://doi.org/10.1002/a-jim.20355.

Schierholz, M., M. Gensicke, N. Tschersich, and F. Kreuter. 2018. "Occupation Coding During the Interview." *Journal of the Royal Statistical Society* A 181: 379–407. DOI: http://dx.doi.org/10.1111/rssa.12297.

Schierholz, M., and M. Schonlau. 2020. "Machine Learning for Occupation Coding – a Comparison Study." *Journal of Survey Statistics and Methodology, smaa023*. DOI: https://doi.org/10.1093/jssam/smaa023.

Tijdens, K. 2014. *Reviewing the measurement and comparison of occupations across Europe* (WP 149, AIAS). Available at: https://pure.uva.nl/ws/files/2172301/154005_WP149_Tijdens_1.pdf (accessed October 2019).

Tijdens, K. 2015a. The design of a tool for the measurement of occupations in web surveys using a global index of occupations. Leuven. (WP InGRID project M21.2). Available at: https://inclusivegrowth.be/downloads/output/m21-4-coding-tool-eind.pdf (accessed October 2020).

Tijdens, K. 2015b. "Self-identification of occupation in web surveys: requirements for search trees and look-up tables" *Survey Methods: Insights from the Field*. Available at: https://surveyinsights.org/wp-content/uploads/2015/06/Self-identification-of-occupation-in-web-surveys-requirements-for-search-trees-and-look-up-tables-Survey-Methods-Insights-from-the-Field-SMIF.pdf (accessed October 2020).

Tijdens, K. 2016. "*Measuring occupations: respondent's self- identification from a large database*." In Proceedings of European Conference on Quality of Official Statistics, Special session: Synergies for Europe's Research Infrastructures in the Social Sciences and Official Statistics (SERISS), 2 June 2016. Available at: https://seriss.eu/wp-content/uploads/2016/06/Measuring-Occupations-Respondent%e2%80%99s-self-identification-from-a-large-database.pdf (accessed October 2020).

Tijdens, K., and S. Visintin. 2017. *EU-harmonised and comparative measurement of occupations and skills*. Leuven. (InGRID project Deliverable 21.1). Available at: https://inclusivegrowth.be/downloads/output/d21-1-eind.pdf (accessed October 2019).

Vannieuwenhuyze, J.T.A., and G. Loosveldt. 2013. "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects," *Sociological Methods and Research* 42(1): 82–104. DOI: https://doi.org/10.1177/0049124112464868.

Vannieuwenhuyze, J.T.A., G. Loosveldt, and G. Molenberghs. 2014. "Evaluating Mode Effects in Mixed-Mode Survey Data Using Covariate Adjustment Models," *Journal of Official Statistics* 30 (1): 1–21. DOI: https://doi.org/10.2478/jos-2014-0001.

Received May 2020
Revised October 2020
Accepted March 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 1009–1045, http://dx.doi.org/10.2478/JOS-2021-0043

Nowcasting Register Labour Force Participation Rates in Municipal Districts Using Survey Data

Jan van den Brakel[1] *and John Michiels*[1]

In the Netherlands, very precise and detailed statistical information on labour force participation is derived from registers. A drawback of this data source is that it is not timely since definitive versions typically become available with a delay of two years. More timely information on labour force participation can be derived from the Labour Force Survey (LFS). Quarterly figures, for example, become available six weeks after the calendar quarter. A well-known drawback of this data source is the uncertainty due to sampling error. In this article, a nowcast method is proposed to produce preliminary but timely nowcasts for the register labour force participation on a quarterly frequency at the level of municipalities and neighbourhoods, using the data from the LFS. As a first step, small area estimates for quarterly municipal figures on labour force participation are obtained using the LFS data and the unit-level modelling approach of Battese, Harter and Fuller (1988). Subsequently, time series of these small area estimates at the municipal level are combined with time series on register labour force participation in a bivariate structural time series model in order to nowcast the register labour force participation at the level of municipalities and neighbourhoods.

Key words: Small area estimation; unit-level model; survey sampling; register-based statistics; data integration.

1. Introduction

Official statistics on the labour force are traditionally obtained by using probability sampling in combination with design-based or model-assisted inference procedures (see Cochran (1977) or Särndal et al. (1992) for an introduction). For national statistical institutes (NSIs) this is a widely accepted approach, since it allows to draw valid inferences about finite populations based on relatively small samples, where it is understood that the size of the sample is small with respect to the population size. In addition, the uncertainty of observing a small sample instead of the entire target population can be quantified through variance calculation. Because of the continuing need of NSIs to reduce costs and response burden, alternative data sources are being explored. An important example is the use of data sources available outside statistical agencies, such as register data from the tax service or local population registers (e.g., see Wallgren and Wallgren 2007, chap. 1; Hand 2018). More recently, also other data sources that are not

[1] Statistics Netherlands (CBS), PO Box 4481, Heerlen 6401 CZ, the Netherlands. Emails: JBRL@cbs.nl and jjm.michiels@cbs.nl

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors are grateful to the Associate Editor and six anonymous referees for careful reading and commenting on two former drafts of this manuscript. Their comments proved to be very helpful to improve this article.

directly related to a statistical or an administrative purpose are considered in the production of official statistics, so-called big data (see e.g., Daas and Puts 2014; Pfeffermann et al. 2015).

Administrative data used in the application described in this article provide an almost complete enumeration of the target population and are therefore useful in the production of detailed statistics. This can be done by using administrative data as a primary data source for producing statistics (Wallgren and Wallgren 2007, chap. 1) or as auxiliary information in small area estimation procedures (Rao and Molina 2015, chap. 5, 7). Other benefits sometimes granted to administrative data and big data are their timeliness. This, however, varies between applications. If administrative data or big data sets become available at a higher frequency than repeated surveys, than their timeliness can be exploited in nowcasting methods and small area estimation methods. The auxiliary information from these more timely data sets can be used to produce more accurate estimates for the sample survey in real time even in instances when the statistics from a big data source become available but the sample data are not yet collected and processed (Vosen and Schmidt 2011; Choi and Varian 2012; Giannone et al. 2008).

Typical examples of big data sources that are available at a high frequency are Google trends, statistics derived from social media platforms, sensor data, mobile phone data, data obtained from GPS trackers, and scanner data.

Administrative data are not necessarily timely. Income tax registers, for example, come in the Netherlands with a delay of two years. In this case, their additional value comes from the fact that they cover the target population almost entirely. In such cases, more timely statistics can be produced with survey samples, with the obvious drawback that sufficient precise estimates are only available at high regional levels. If a repeated survey is conducted to produce more timely statistics than administrative sources can, nowcasting methods can be used to combine both data sources and produce preliminary estimates at a detailed level. An example of this is considered in this article, where the Dutch Labour Force Survey (LFS) and tax registers in the Netherlands provide information about the employed labour force.

For the purpose of producing statistics on the labour force participation rate at a low regional level, the traditional approach of survey sampling and design-based estimators was, until recently, the preferred method in the Netherlands. Until 2015, municipal estimates were produced annually by means of direct general regression estimation (GREG) (see, for example Särndal et al. 1992, chap. 6), for municipalities with at least 30,000 inhabitants. To improve annual municipal estimates, a model-based small area estimation method (SAE) has been implemented. Small area estimation refers to a class of model-based estimation procedures that explicitly rely on a statistical model to improve the precision of sample estimates. Mainstream SAE methods are based on cross-sectional multilevel models, which improve the precision of domain estimates with sample information observed in other domains observed in the same reference period. Official statistics, however, are generally produced repeatedly in time. For such situations, multivariate time series models are appropriate since they improve sample estimates with sample information observed in previous reference periods as well as information observed in other domains. A comprehensive overview of the literature on SAE is provided by Rao and Molina (2015). For a more compact, but nevertheless complete overview of the SAE literature, see Pfeffermann (2002, 2013).

Statistics Netherlands uses a unit level model (Battese et al. 1988), for the estimation of labour status at the level of municipalities at an annual frequency. See Boonstra et al. (2011) and CBS Statline, (2017) for details. Still, this latter approach is not ideal for statistics at a very low regional level, such as districts within municipalities, or other very small subdomains. Small area estimators for very small domains reduce to purely synthetic estimators, which may be severely biased.

An alternative approach is to use register data to produce statistics on the employed labour force. If tax data are available at the individual level about persons with paid employment and persons who are self-employed, detailed statistics on the labour force participation rate can be produced for small domains. The register alternative can deliver labour force statistics at a regional level within municipalities, although it may fall short when there is a need for timely statistics on the labour force participation rate.

For very detailed and timely regional statistics (or other statistics for small domains) each of the latter two techniques are therefore in many cases insufficient. They may, however, be combined to produce more accurate results using nowcasting methods. Many official statistics are produced repeatedly. Using information from the past to make more precise estimates for the last period is possible in a time series approach. This idea has been pursued by many authors dating back to research by Blight and Scott (1973). Moreover, auxiliary time series that correlate with the target series may be used to improve the accuracy of the target series. Other key references to authors that apply time series methods to survey data are Rao and Yu (1994), Pfeffermann and Burck (1990), Harvey and Chung (2000) and Pfeffermann and Tiller (2006).

The Netherlands is divided into 12 provinces and contains about 400 municipalities. Municipalities are further divided into neighborhoods, about 12,000 in total. The purpose of this article is to use more timely quarterly survey data on labour force participation at the municipal level to nowcast less timely register statistics at the more detailed level of neighbourhoods. To this end, the bivariate state space model developed by Van den Brakel et al. (2017) is further extended in this article. Quarterly time series of labour force participation rates for approximately 12,000 neighbourhoods have been derived from register sources. Alongside these register series quarterly survey time series of labour force participation rates at the municipal level are derived for the time period 2003–2014 using small area predictions obtained using a cross-sectional unit level model. The approach applied in this article is to use the register time series up to 2013 and the LFS time series up to 2014 in order to nowcast the register statistics for the periods 2012–2014. These predictions are then compared with the final register statistics.

The novelty of this article can be found in the following extensions. A timely time series obtained with a repeated survey is used as an auxiliary series in a state-space model to nowcast a very detailed target series that comes from a register with a time lag of two years. A cross-sectional unit level model is applied to the survey data as a first step, which is used to produce auxiliary series for the state-space model. The state-space model accounts for the uncertainty in the auxiliary series, as well as the autocorrelation, which arise from the rotating panel design of the LFS. Finally, the model is applied to the COVID-19 pandemic, for the purpose of evaluating the nowcast procedure directly after a turning point induced by the COVID-19 crisis.

In Section 2, the survey design of the labour force survey and its regional estimation procedure is described. Also described is the construction of regional statistics on the

employed labour force using register data from the tax service. In Section 3, structural time series models are proposed for the survey and register time series on labour force participation rates over the time period 2003–2014. Both univariate and bivariate models are considered and the results obtained by these models are presented in Section 4. The article concludes with a discussion in Section 5.

2. Official Statistics About the Dutch Labour Force

2.1. Labour Force Survey

2.1.1. Survey Design

The target population of the Dutch LFS is defined as persons aged 15 to 75 years residing in the Netherlands and not living in an institutional household. Since 2000, the Dutch LFS is based on a rotating panel design, in which respondents are interviewed five times at quarterly intervals. Each month a sample of addresses is selected through a stratified two-stage cluster design. Strata are formed by geographic regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. All households residing at an address, up to a maximum of three, are included in the sample. About 9,000 new respondents are interviewed monthly in the first wave of the panel design. Until 2010, data collection in the first wave was based on computer-assisted personal interviewing (CAPI) only. From 2010 until 2012, data collection in the first wave changed to a mixed-mode data collection approach of computer-assisted telephone interviewing (CATI) if a non-secret landline number was available, and CAPI for the remaining households. Since 2012, data collection in the first wave is based on a sequential mixed-mode design that starts with computer-assisted web interviewing (CAWI). After three reminders, non-respondents are approached by means of CATI if a non-secret landline number is available, and by CAPI for the remaining households. In the four subsequent waves of the panel, data are collected by means of CATI, since the start of the panel in 2000. During these re-interviews, a reduced questionnaire is completed to establish changes in the labour market position of the household members aged 15 years and over. When a household member cannot be contacted, proxy interviewing is allowed by members of the same household in each wave. Labour force participation is established by asking respondents about (not) having a paid job.

The aforementioned redesigns in 2010 and 2012 resulted in systematic effects in the LFS estimates. The series observed before 2012 are adjusted to the level of the sequential mixed mode design observed since 2012. See Van den Brakel and Krieg (2015) for more details on how to quantify and correct for these so-called discontinuities in the Dutch LFS.

2.1.2. Estimation Procedures for Official Labour Force Figures

LFS data are used to publish official statistics about the labour force on a monthly, quarterly and annual frequency. Inference methods of the LFS are based on a mixture of design-based and model-based techniques. Monthly figures about the unemployed, employed and total labour force are based on a multivariate structural time series modelling approach. This method improves the precision compared to direct estimates by

using sample information from preceding periods, accounts for rotation group bias and autocorrelation induced by the rotating panel design of the LFS. Monthly figures are published at the national level and as a breakdown in six domains based on a cross-classification of age and gender (see Van den Brakel and Krieg (2015) for details).

Quarterly figures at the national level are based on the GREG estimator. The weighting procedure contains the monthly estimates from the time series model to force consistency between monthly and quarterly figures about the labour force. Several registrations provide additional auxiliary information that is used in this weighting procedure. These auxiliary variables include registered paid employment (source: Polis administration), which is a strong predictor for labour force participation.

Regional quarterly and annual information on the labour force is provided by the LFS in official statistics using small area techniques for the estimation of labour force indicators at the municipal and province level, since sample sizes are too small to use direct estimators at this level. The backbone of this modelling approach is an hierarchical Bayesian version of the Battese-Harter-Fuller unit-level model (Battese et al. 1988; Datta and Ghosh 1991). Various labour force indicators are presented for a number of subpopulations (according to sex, age groups, educational level and ethnicity). The most detailed regional indicators (at the municipal level) appear on a yearly basis, within two months after the reference year has ended.

Quarterly sample sizes of the LFS amount to approximately 85,000 persons, distributed across just over 400 municipalities. Municipal sample sizes vary between zero for the smallest municipalities and over 3,000 for the capital, Amsterdam. In Figure 1 a Pareto graph visualises the distribution of the sample over the municipalities for the first quarter of 2014. For the purposes of this publication, the small area estimates of all municipalities have been estimated on a quarterly basis using the aforementioned Battese-Harter-Fuller unit-level model. For a given period t, let $y_{i,d}^{LFS}$ denote a binary indicator for labour force participation and $x_{i,d}$ a vector of covariates for sample unit i and municipality d. The linear

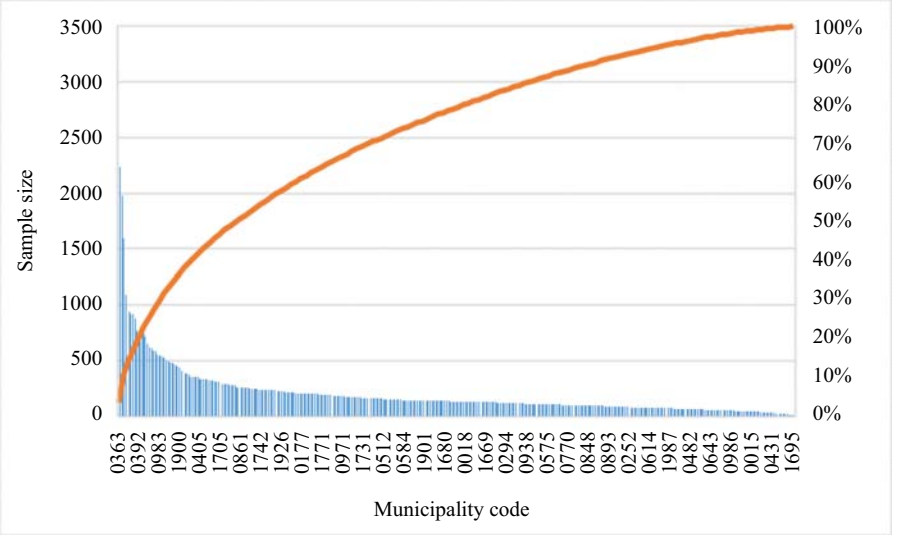

Fig. 1. Pareto graph of municipal sample sizes in the Dutch LFS for the first quarter of 2014.

unit level model is used as an approximation for the observed variables, that is, $y_{i,d}^{LFS} = \beta' x_{i,d} + v_d + \varepsilon_{i,d}$, with β a vector of regression coefficients, $v_d \cong N\left(0, \sigma_v^2\right)$ an i.i.d. random domain effect and $\varepsilon_{id} \cong N\left(0, \sigma_\varepsilon^2\right)$ i.i.d. random errors. Let $\bar{y}_d^{LFS}$ denote the sample mean of the labour force participation rate of municipality d. Based on this model, empirical best linear unbiased predictions for the domains are obtained by

$$y_d^{LFS} = \hat{\gamma}_d\left(\bar{y}_d^{LFS} + \hat{\beta}'(\bar{X}_d - \bar{x}_d)\right) + \left(1 - \hat{\gamma}_d\right)\hat{\beta}'\bar{X}_d, \tag{1}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \hat{\sigma}_\varepsilon^2/n_d}, \ \hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y.$$

Here, n_d is the number of observations in municipality d, $n = \sum_{d=1}^{D} n_d$ is the number of observations in a particular period t, D is the total number of municipalities, $\bar{x}_d$ is a vector with sample means of municipality d and $\bar{X}_d$ is a vector with the corresponding population means, X is the full matrix with covariates of the n units included in the sample, y is the n vector with observations of the n units included in the sample, and $\hat{\Sigma}_t = cov(y_t) = \hat{\sigma}_{\varepsilon,t}^2 I_{n_t} + \hat{\sigma}_{v,t}^2 \oplus_{d=1}^{D} J_{n_{t,d}}$, where I_n is the identity matrix of order n and $\oplus_{d=1}^{D} J_{n_{t,d}}$ is the block diagonal matrix of order $n \times n$ with J_{n_d} the $n_d \times n_d$ block diagonal elements with each element equal to 1. Note that the empirical best linear unbiased prediction estimator defined in Equation (1) can also be expressed as $y_d^{LFS} = \hat{\beta}'\bar{X}_d + \hat{\gamma}_d\left(\bar{y}_d^{LFS} - \hat{\beta}'\bar{x}_d\right)$. The advantage of Equation (1) is that it has the interpretation that the estimator is the weighted average of the survey regression estimator $\hat{\gamma}_d\left(\bar{y}_d^{LFS} - \hat{\beta}'\bar{x}_d\right)$ and the regression synthetic estimator $\hat{\beta}'\bar{X}_d$, see Rao and Molina (2015, subsec. 7.2). The model used in Equation (1) is:

$$\begin{aligned}&\text{EmployedReg} \times [\text{Gender} + \text{Age(3)} + \text{Ethinicity(3)} + \text{Bordermunicipality(2)}]\\ &+ \text{Gender} \times \text{Age(3)} + \text{Age(5)} + \text{Ethnicity(7)} + \text{EmploymentOffice(5)}\\ &+ \text{HHType(3)} + \text{Wage(6)} + \text{WaveNr},\end{aligned} \tag{2}$$

where the variables are explained in Appendix (Subsection 6.1). The demographic auxiliary variables in the model are available from the Municipal Basic Administration (MBA). This is a register of all people residing in the Netherlands and contains background variables such as age, sex, nationality, and marital status. Since Dutch citizens are required by law to report changes in their demographics to their municipalities, this register provides a very accurate list of the Dutch population (Bakker 2012). EmployedReg and EmploymentOffice are available from registers of the Employee Insurance Agency and can be linked with the MBA through a personal identifier. Each person residing in the Netherlands has a unique personal identifier that is available in most registrations.

Equation (1) is expressed as an hierarchical Bayesian model with a flat prior on β, σ_v^2 and σ_ε^2 and fitted using the R-package hbsae, (Boonstra 2012). Small area predictions and their standard errors are obtained from the posterior mean and posterior variance of y_d^{LFS}.

Applying a linear model directly to binary data or percentages might appear rigid at first sight but similar linear models are used to motivate the general regression estimator that is generally used in survey sampling to estimate sample means or totals of binary or categorical variables. Boonstra et al. (2007) conducted a large simulation study for the Dutch municipal labour force figures, in which it was found that logistic unit-level models do not improve upon normal linear models. Furthermore, the quantities of interest are area means. Interest is not

focused in predictions for individual units, but always aggregate such predictions to the area level. This makes it more reasonable to use normal linear models. Finally, prediction for non-linear models is computationally more cumbersome than for linear models and often not worthwhile the effort. See, for example, Bijlsma et al. (2020) for another application in which logistic models do not outperform the normal linear model. Selection and evaluation of this model is described in detail in Boonstra et al. (2007, 2008, 2011). Other examples where the area level model is applied to untransformed estimated percentages in the context of SAE are Datta et al. (1999), You et al. (2003), and Arima et al. (2017). Early references on the use of the logistic-normal model for SAE are MacGibbon and Tomberlin (1989) and Malec et al. (1997). See Hobza and Morales (2016), Hobza et al. (2018), and Marino et al. (2019) for more recent literature on alternative ways of obtaining small area predictions for binary data under a logistic mixed model. In this article, the linear Equation (1) is applied, since this is the model that is used by Statistics Netherlands for the production of official quarterly LFS figures.

Instead of cross-sectional small area estimation models, time series small area estimation models can be used as an alternative. Rao and Yu (1994) extended the cross-sectional area level model by Fay and Herriot (1979) with an AR(1) component to borrow strength over time and space. Harvey and Chung (2000) proposed a time series model for the LFS in the United Kingdom extended with a series of claimant counts. Pfeffermann and Burck (1990) and Pfeffermann and Tiller (2006) developed multivariate structural state space models to borrow strength over time and space. Boonstra and Van den Brakel (2019) emphasise that structural time series models can be considered as time series extensions of the area level model, which both can be expressed as state space models or time series multilevel models, resulting in similar estimation results.

2.2. Register Data for Low Regional Statistics on the Employed Labour Force

The register data contain information about the sources of income of persons on a monthly basis, including income from paid employment, self-employment and social benefits. These register data are based on data collected by the tax service (De Belastingdienst) and the Employee Insurance Agency. Self-employed taxpayers may opt for a delayed tax declaration and the final tax assessment may take several years. As a result, data collection is significantly slower than with the LFS. Approximately nine months after the reference year has ended a preliminary data set is produced and the final one is produced the year after. These data sets can be combined with other register data that contain individual information about demographic and regional variables among other variables.

In Figure 2, time series are presented on quarterly small domain estimates of labour force participation rates from the LFS and quarterly figures on labour force participation rates derived from the tax register. The municipal domain estimates are based on the Battese-Harter-Fuller unit-level model. The time series contain quarterly data over the period 2003–2014 for three municipalities: Amsterdam, the largest Dutch municipality (811,000 inhabitants, January 2014); Heerlen, a medium-size municipality in the south-east corner of the Netherlands (88,000 inhabitants, January 2014) and Haren, a relatively small municipality in the north (14,000 inhabitants, January 2014).

For the three municipalities, it follows that the results for labour force participation rates are very close, which is not trivial. In many situations, survey data and related register data

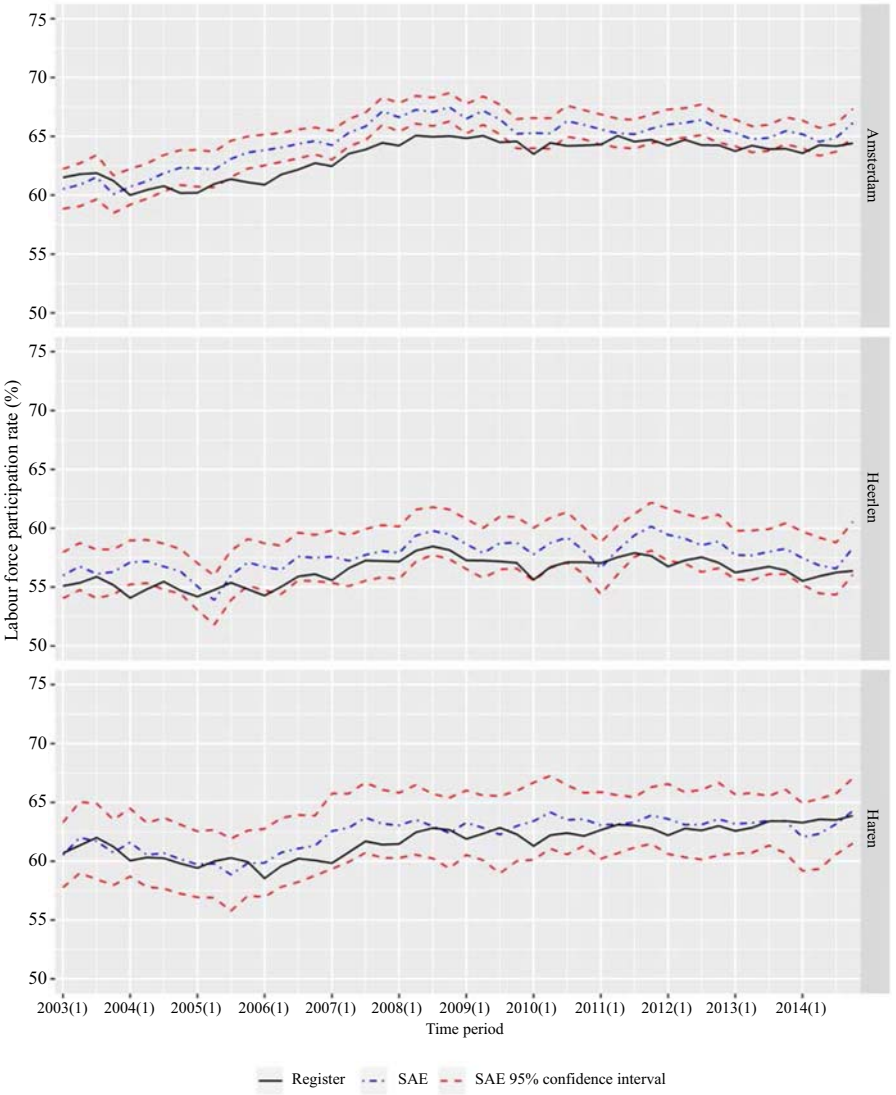

Fig. 2. LFS and register labour force participation rates for a large, medium-sized and small municipality respectively, quarterly data 2003–2014 with a 95% confidence interval for the LFS SAE statistics.

follow a similar evolution but generally at different levels. Although a small area estimation approach is used to estimate the labour force participation rate with the LFS data, the amount of uncertainty increases inversely proportional to the sample size of the domains.

2.3. *Accuracy of Survey and Register Statistics Concerning Regional Labour Force Statistics*

The accuracy of statistics is usually measured in terms of bias and variance. We here distinguish between selection bias, bias introduced by the estimation procedure and measurement bias. Variance constitutes an important part of the uncertainty of LFS labour

force indicators, especially for small domains (municipal regions) in which the sample sizes are limited. To overcome this problem, small area estimators are used for the calculation of labour force participation rates at the municipal level. These estimators are not unbiased, but in the presence of strong auxiliary information using models that meet their underlying assumptions, this bias might be limited and the variance reduction dominates. As a result, the mean squared error of these estimators is smaller than the variance of traditional direct (GREG) estimators.

Another contributing factor to the uncertainty of survey statistics is selection bias. In theory, this bias is zero under complete response. In practice, selection bias arises due to undercoverage of the sampling frame, the inability to reach the target population with the implemented field work strategies and data collection mode successfully, and selective nonresponse. In register data, there is no uncertainty due to a sampling design. However there still is uncertainty about labour force participation rate statistics due to incomplete data collection in the preliminary tax datasets. The uncertainty stemming from temporary nonresponse is partially removed by imputing the missing records for the self-employed in a deterministic procedure (by checking the existence of provisional assessments and companies' registry membership). Imputing for missing values may and does lead to biased estimates. Apart from temporary nonresponse, there are also groups that do not need to declare their income (e.g., some groups of household workers) and workers that do not declare their taxable income although they should (informal labour market).

In survey sample statistics, the measurement bias depends on the extent to which conceptual variables to be measured are operationalised correctly in the questionnaire. However, the mode of data collection and other contributing factors (such as the quality of the interviewers, and errors introduced elsewhere in the production process) are relevant.

Register data are not without measurement bias. In addition, taxpayers have to complete a questionnaire (their tax declaration) and usually do this without the help of an interviewer. The difficult nature of the tax form may lead to measurement bias. Furthermore, there may be underreporting of income sources. However, it is assumed in this application that these errors can be ignored.

3. Time Series Modelling of Register and Survey Indicators of Labour Force Participation

In this section, univariate and bivariate structural time series models are developed for the quarterly statistics on labour force participation rates obtained with the LFS and the tax register. With a structural time series model, a series is decomposed into a trend component, a seasonal component, other cyclic components, a regression component and an irregular component. For each component, a stochastic model is assumed. This allows the trend, seasonal, and cyclic component, but also the regression coefficients, to be time dependent. If necessary, ARMA components can be added to capture the autocorrelation in the series beyond these structural components. See Harvey (1989) or Durbin and Koopman (2012) for details about structural time series modelling.

The question addressed in this article is whether time series from an auxiliary source at an aggregate level (e.g., municipalities) with more timely data than the main time series, but hampered with sampling error, can be used to make more precise first predictions (nowcasts)

for the measurements in the main series at a more detailed regional level (neighbourhoods). The auxiliary series is considered useful if the combination of both series leads to smaller mean square errors for the nowcasted estimates than in the case of a univariate modelling approach of the main series. The main time series is based on register data on labour force participation rates at the level of municipalities and underlying neighbourhoods. The auxiliary series consists of timely estimates of labour force participation rates at the level of municipalities or more aggregated regional units obtained from the LFS. The question is tackled by developing a bivariate structural time series model for the register series (for a municipality or neighbourhoods within a municipality) and the LFS series (for the corresponding municipality or a higher level regional unit) and modelling the correlation between the disturbance terms of the different components of the structural time series for both models.

The model selection starts by building appropriate univariate time series models for the labour force participation rates obtained with the LFS and the tax register data. This forms the input for building an optimal bivariate model. The performance of this bivariate model will be compared with predictions based on the univariate model for labour force participation rates. Results are presented for three municipalities: a large municipality, Amsterdam, together with a mediumsized and a small municipality, respectively: Heerlen and Haren. In this fashion, the effect of larger standard errors for estimates in smaller municipalities can be included in the analysis. In a further stage of the analysis register, labour force participation rates for neighbourhoods will also be related to the corresponding survey indicator at the municipal level.

3.1. *Univariate Models for Register and Survey Statistics on Labour Participation*

Let $y^R_{t,d}$, denote the register measurements of labour force participation rate for period t of municipality d and $y^{LFS}_{t,d}$ the survey estimates for labour force participation rate for period t of municipality d based on the LFS and estimated with the HB model defined in Equation (1). For both time series, the following structural time series model is proposed:

$$y^x_{t,d} = L^x_{t,d} + S^x_{t,d} + E^x_{t,d} \text{ with } x \in \{R, LFS\}, \tag{3}$$

with L^x_t and S^x_t an appropriate stochastic model for the trend and the seasonal component and E^x_t the remaining unexplained variation.

For both time series, the smooth trend model appears to be the most appropriate model to capture both trend and cyclic components. The smooth trend model is defined as (Durbin and Koopman, 2012, Subsubsec. 3.2.1):

$$\begin{aligned} L^x_{t,d} &= L^x_{t-1,d} + R^x_{t-1,d}, \\ R^x_{t,d} = R^x_{t-1,d} &+ \eta^x_{t,d}, \eta^x_{t,d} \cong N(0, \sigma^{x^2}_{\eta,d}),\ Cov(\eta^x_{t,d}, \eta^x_{t',d}) = 0 \text{ if } t \neq t'. \end{aligned} \tag{4}$$

In Equation (4) $L^x_{t,d}$ is the level of the trend, which is defined as the level of the previous period plus a change, $R^x_{t-1,d}$ which can be interpreted as a slope parameter. The slope on its turn is modelled as a random walk. As a result, the trend can gradually change over time. Therefore, Equation (4) models the low frequency variation of the series and has the flexibility to model the trend and long-term economic cycles. The flexibility of the trend component is determined by the value of the variance of the slope disturbance terms, that is, $\sigma^{x^2}_{\eta,d}$. If $\sigma^{x^2}_{\eta,d} = 0$, then Equation (4) defines a straight line, that is, $L^x_{t,d} = L^x_{0,d} + tR^x_{0,d}$,

with $L^x_{0,d}$ the intercept which defines the level of the trend at the start of the series and $R^x_{0,d}$ a time-invariant slope.

Alternative trend models are the local linear trend model and the local level model. The local linear trend model is obtained if an additional random component is added to $L^x_{t,d}$ in Equation (4). In that case, the first line of Equation (4) equals $L^x_{t,d} = L^x_{t-1,d} + R^x_{t-1,d} + \zeta^x_{t,d}$, with $\zeta^x_{t,d}$ a level disturbance term that is normally and independently distributed, that is $\zeta^x_{t,d} \cong N(0, \sigma^{x^2}_{\zeta,d})$, $Cov(\zeta^x_{t,d}, \zeta^x_{t',d}) = 0$ if $t \neq t$'.

The local level model is obtained if $L^x_{t,d}$ is modelled as a random walk without a slope parameter $R^x_{t,d}$, that is, $L^x_{t,d} = L^x_{t-1,d} + \zeta^x_{t,d}$. The linear trend model and the local level model generally give more volatile trend estimates compared to the smooth trend model.

A likelihood-ratio test shows that the local linear trend model for the register series does not significantly improve model fit since the log-likelihoods for a smooth trend model and the local linear trend model are almost equal. A local level model for the register series tends to overfit the data. This means that the maximum likelihood estimates for the variance of the trend disturbance terms are large, while the variance of the measurement errors tend to zero. As a result, the filtered and smoothed signals of the time series model are almost identical to the observed time series and have extremely small confidence intervals. The local level model also leads to nearly identical log-likelihoods as compared to the other two model types. Note that the likelihood of a structural time series model is based on the one-step-ahead-prediction error decomposition (Harvey 1989, subsec. 3.4). Since a model that overfit the data has low prediction power, the log-likelihood is not improved compared to the more parsimonious smooth trend model. For the survey series, the local level model and local linear trend model both tend to overfit, hence once again the smooth trend model is selected for the LFS series.

The seasonal component is modelled with a trigonometric model. For a quarterly series, this model is defined as (Durbin and Koopman, 2012, subsubsec. 3.2.2):

$$S^x_{t,d} = \sum\nolimits_{j=1}^{2} \gamma^x_{jt,d}, \tag{5}$$

with

$$\gamma^x_{jt,d} = \gamma^x_{jt-1,d} \cos\left(\lambda_j\right) + \tilde{\gamma}^x_{jt-1,d} \sin\left(\lambda_j\right) + \omega^x_{jt,d},$$

$$\tilde{\gamma}^x_{jt,d} = -\gamma^x_{jt-1,d} \sin\left(\lambda_j\right) + \tilde{\gamma}^x_{jt-1,d} \cos\left(\lambda_j\right) + \tilde{\omega}^x_{jt,d},$$

and

$$\lambda_j = \frac{\pi \times j}{2}.$$

For the disturbance terms it is assumed that:

$$\omega^x_{jt,d} \cong N\left(0, \sigma^{x^2}_{\omega,d}\right), \quad \tilde{\omega}^x_{jt,d} \cong N\left(0, \sigma^{x^2}_{\omega,d}\right),$$

$$\mathrm{Cov}\left(\omega^x_{jt,d}, \omega^x_{j't',d}\right) = 0 \text{ and } \mathrm{Cov}\left(\tilde{\omega}^x_{jt,d}, \tilde{\omega}^x_{j't',d}\right) = 0 \text{ if } t \neq t' \text{ or } j \neq j',$$

$$\mathrm{Cov}\left(\omega^x_{jt,d}, \tilde{\omega}^x_{j't',d}\right) = 0.$$

This means that the variance structure for all the frequency contributions is assumed to be identical with no correlation between the disturbance terms of these contributions. The

interpretation of Equation (5) is that a quarterly seasonal pattern is modelled with a set of trigonometric terms at the seasonal frequencies $\lambda_1 = \pi/2$ and $\lambda_2 = \pi$. Equation (5) shows that $\tilde{\gamma}^x_{1t,d}$ contributes indirectly to the seasonal pattern via $\gamma^x_{1t,d}$. Since $\sin(\pi) = 0$, there are effectively three parameters, that is, $\gamma^x_{1t,d}$, $\tilde{\gamma}^x_{1t,d}$ and $\gamma^x_{2t,d}$ to model the quarterly seasonal pattern. For a more detailed discussion of the trigonometric seasonal model, see Harvey (1989, subsubsec. 2.3.4).

The unexplained variation for the register series can be modelled as a white noise process, that is, $E^R_{t,d} \equiv \xi^R_{t,d} \cong N(0, \sigma^{R^2}_{\xi,d})$. For the LFS series, the situation is more complicated since the unexplained variation contains the unexplained variation of the real but unknown population parameter, say $\xi^{LFS}_{t,d}$, plus the sampling error, say $e^{LFS}_{t,d}$. As a result, it follows that $E^{LFS}_{t,d}$ can be decomposed as $E^{LFS}_{t,d} = \xi^{LFS}_{t,d} + e^{LFS}_{t,d}$. The unexplained variation of the population parameter is modelled as a white-noise disturbance term, $\xi^{LFS}_{t,d} \cong N(0, \sigma^{LFS^2}_{\xi,d})$. Due to the rotating panel design of the LFS, there is a large sample overlap between adjacent quarters, which results in a significant positive autocorrelation in the sampling error $e^{LFS}_{t,d}$. Furthermore, the variance of the sampling error changes over time because the sample size of the LFS change over time. Let $MSE(y^{LFS}_{t,d})$ denote the estimated variances of the quarterly small domain predictions obtained with Battese-Harter-Fuller unit-level Equation (1). The time series model accommodates for the heteroscedasticity in the sampling error by scaling the sampling errors with the estimated mean squared errors of the quarterly small domain predictions, that is, $e^{LFS}_{t,d} = \sqrt{MSE(y^{LFS}_{t,d})}\tilde{e}^{LFS}_{t,d}$. In a next step, the Yule-Walker equations are applied to the autocovariances of the estimated sampling errors to derive an appropriate AR model for the autocorrelation in the sampling errors. It is established that an AR(1) model is sufficient to model the autocorrelation between the survey errors. As a result, the scaled sampling errors are modelled as $\tilde{e}^{LFS}_{t,d} = \rho_d \tilde{e}^{LFS}_{t-1,d} + \varsigma^{LFS}_{t,d}$, with ρ_d the autoregressive parameter obtained from the Yule-Walker equations and $\varsigma^{LFS}_{t,d} \cong N(0, \sigma^2_{\varsigma,d})$. This is in line with the findings of Van den Brakel and Krieg (2015) and Boonstra and Van den Brakel (2019). Using the variance of $e^{LFS}_{t,d}$ and the coefficient for the AR(1) model derived from the micro data as a priori information in the time series models allows the identification of the following structure for the unexplained variation of the LFS series:

$$E^{LFS}_{t,d} = \xi^{LFS}_{t,d} + e^{LFS}_{t,d} \equiv \xi^{LFS}_{t,d} + \sqrt{MSE\left(y^{LFS}_{t,d}\right)}\tilde{e}^{LFS}_{t,d}, \tag{6}$$

$$\tilde{e}^{LFS}_{t,d} = \rho_d \tilde{e}^{LFS}_{t-1,d} + \varsigma^{LFS}_{t,d},$$

Note that $Var(e^{LFS}_t) = MSE(y^{LFS}_{t,d})\sigma^2_{\varsigma,d}/(1-\rho^2_d)$. The variance component $\sigma^2_{\varsigma,d}$ is estimated via maximum likelihood (see Subsection 3.4). If $MSE(y^{LFS}_{t,d})$, is a good approximation of the variance of e^{LFS}_t, that is, if $Var(e^{LFS}_{t,d}) \approx MSE(y^{LFS}_{t,d})$, then $\sigma^2_{\varsigma,d}/\left(1-\rho^2_d\right)$ is expected to be close to one, and thus $\sigma^2_{\varsigma,d} \approx \left(1-\rho^2_d\right)$. In this way, $\sigma^2_{\varsigma,d}$ is in fact a scaling factor for $MSE(y^{LFS}_{t,d})$ to correct for possible bias in the estimates for $MSE(y^{LFS}_{t,d})$ that are used as a priori information in the time series model.

3.2. Bivariate Model for Register and Survey Statistics on Labour Participation

The univariate models for the register and LFS series can be combined in one bivariate model. This model can be used to nowcast the register series at the level of

neighbourhoods with the more timely LFS series at the municipal level. For the bivariate model, the following structure is proposed

$$\begin{pmatrix} y^R_{t,d} \\ y^{LFS}_{t,d} \end{pmatrix} = \begin{pmatrix} L^R_{t,d} \\ L^{LFS}_{t,d} \end{pmatrix} + \begin{pmatrix} S^R_{t,d} \\ S^{LFS}_{t,d} \end{pmatrix} + \begin{pmatrix} 0 \\ e^{LFS}_{t,d} \end{pmatrix} + \begin{pmatrix} \xi^R_{t,d} \\ \xi^{LFS}_{t,d} \end{pmatrix}. \tag{7}$$

The models for the trend, seasonal components, the sampling error, and the population white noise are defined in Subsection 3.1. To improve the precision of the nowcasts for the register series with the additional timely information from the LFS series, the correlation between the disturbance terms of the trend, seasonal component and population white noise, can be modelled in Equation (7). In the final selected model, only a non-zero correlation for the slope disturbance terms of the trend components is defined. The seasonal components turned out to be time invariant, while no correlation is detected between population white noise terms. This results in the following variance structure for the slope disturbances of the smooth trend models for the register and LFS series:

$$Cov\left(\eta^x_{t,d}, \eta^x_{t',d}\right) = \begin{cases} \left(\sigma^x_{\eta,d}\right)^2 & if\ t = t' \\ 0 & if\ t \neq t' \end{cases} \quad \text{for } x \in \{R, LFS\}, \tag{8}$$

$$Cov\left(\eta^R_{t,d}, \eta^{LFS}_{t',d}\right) = \begin{cases} \sigma^R_{\eta,d}\sigma^{LFS}_{\eta,d}\,\rho_{\eta,d} & if\ t = t' \\ 0 & if\ t \neq t' \end{cases}.$$

If the model detects (strong) correlation between the trends of both series, then this implies that the trends of both series develop more or less in the same direction. Note that Equation (8) defines a 2×2 covariance matrix for the slope disturbance terms of the trends. In the case of strong correlation, $\rho_{\eta,d} \rightarrow 1$ (or $\rho_{\eta,d} \rightarrow -1$), the rank of this covariance matrix will reduce to one. In that case, the trend components of both series are said to be cointegrated, since they are driven by one underlying common trend. Expressing the model in terms of a common factor model results in a more parsimonious model and improves the efficiency of the estimation procedure. See Harvey (1989, subsec. 8.5), or Koopman et al. (2007, subsec. 9.1), for more details concerning cointegration and common factor state space models.

In the case of strong correlation or even cointegration, the LFS series contains valuable information for the prediction of register labour force participation rates at times when those register data are not yet available while the LFS estimates have already been calculated. In the univariate case, the predictions are based on past observations of the register series alone. In the bivariate case, the correlation between both series may improve the accuracy of these predictions. To see whether more accurate small area estimates of the auxiliary series lead to greater improvements in the accuracy of the nowcasted main series results, the model calculations are repeated for a number of differently sized municipalities.

3.3. Multiplicative Models

Equations (3) and (7) are additive models. Since the target variables are percentages, multiplicative models can be considered as an alternative. This is achieved by taking the

log of the observed series. Thus $y^R_{t,d}$ and $y^{LFS}_{t,d}$ are replaced by $log(y^R_{t,d})$ and $log(y^{LFS}_{t,d})$ in Equations (3) and (7) respectively. For the variance structure of the measurement equation of the LFS, an approximation of the MSE of the $log(y^{LFS}_{t,d})$ in Equation (6) is required. Based on a first-order Taylor approximation of $log(y^{LFS}_{t,d})$ it follows that $MSE(log(y^{LFS}_{t,d})) \approx MSE(y^{LFS}_{t,d})/(y^{LFS}_{t,d})^2$. In a similar way it follows for the auto-covariance that $cov(log(y^{LFS}_{t,d}), log(y^{LFS}_{t-1,d})) \approx cov(y^{LFS}_{t,d}, y^{LFS}_{t-1,d})/(y^{LFS}_{t,d} \times y^{LFS}_{t-1,d})$. As a result, the auto-correlations are unaffected and the AR(1) structure derived for the additive model can be used for the multiplicative model.

3.4. Estimation of Structural Time Series Models

A widely applied approach to fit structural time series models is to put them in the so-called state space form. Under the assumption of normally distributed disturbance terms, optimal estimates of the state variables and linear combinations of these variables are obtained with the Kalman filter. State variables are the variables that define the unobserved components, that is, the $L^x_{t,d}$ and $R^x_{t,d}$ of the trend, the $\gamma_{jt,d}$ and $\tilde{\gamma}_{jt,d}$ of the seasonal component and the sampling errors $\tilde{e}^{LFS}_{t,d}$. The Kalman filter is a recursive procedure that runs from period $t = 1$ to T and gives, for each time period, an optimal estimate for the state variables based on the information available up to and including period t. These estimates are referred to as the filtered estimates. The filtered estimates of past state vectors can be updated, if new data after period t become available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the complete time series. In this article, the fixed-interval smoother is used. Variances for the state variables and signals are obtained with the standard Kalman filter recursions.

The Kalman filter assumes that the variances and covariances of the disturbance terms in the models for the trend, seasonal components, sampling error and population white noise, are known in advance. These components are often referred to as the hyperparameters of the state space model. In practice, these parameters are not known and therefore have to be estimated. In this article, maximum likelihood estimates for the hyperparameters are obtained by numerically optimising the likelihood function with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm and repeatedly running the Kalman filter.

The maximum likelihood estimates for the hyperparameters are inserted in the Kalman filter. The uncertainty due to replacing the unknown hyperparameter values for their maximum likelihood estimates is further ignored in the standard errors of the Kalman filter estimates for the state variables and signals. Pfeffermann and Tiller (2005) developed a bootstrap to account for the additional uncertainty of the maximum likelihood hyperparameter estimates in the standard errors of the Kalman filter estimates. An alternative approach is to express the bivariate model as an hierarchical Bayesian time series multilevel model and fit the model using MCMC simulations. This approach also accounts for the additional uncertainty of the hyperparameter estimates (see, for example, Boonstra and Van den Brakel 2019). Bollineni-Balabay et al. (2017) analysed the additional uncertainty in the standard errors of the Kalman filter estimates in the Dutch LFS because maximum likelihood estimates for the hyperparameters are plugged into the Kalman filter recursions. They conclude that this additional increase is ignorable in their

application. In this article, we follow the standard state-space approach and ignore this uncertainty in this application.

The Kalman filter recursions provide predictions for missing observations in the time series. These predictions can be interpreted as imputations obtained by the EM algorithm (Durbin and Koopman 2012, subsubsec. 7.3.4). This property is used here for nowcasting the register series. The last four quarters of the register series are missing. The Kalman filter provides predictions or nowcasts for these missing values, including their variances, which can be interpreted as imputations obtained by the EM algorithm.

All state variables, except for the sampling error $\tilde{e}_{t,d}^{LFS}$ of the LFS series, are non-stationary. For the non-stationary state variables a diffuse initialization of the Kalman filter is used. This means that their starting values at $t = 0$ are chosen equal to zero with a large variance (10E7). The sampling errors are stationary and for their state variables an exact initialization can be applied. This means that their starting values are chosen equal to zero with a variance derived from the assumed AR(1) process. More technical details about state space models and their analysis can be found in Harvey (1989) or Durbin and Koopman (2012).

Performance of the models for nowcasting register labour force participation out of sample forecasts is evaluated for the periods 2012 through 2014. Let T denote the quarter of the last year that is observed with the register. For T equal to the fourth quarter of respectively 2011, 2012 and 2013, out of sample forecasts for the state variables and signals for the register employed labour force are made for the next four quarters, that is, $T+1$, $T+2$, $T+3$ and $T+4$. Results for nowcasts are estimated in real time, that is, the hyperparameters are re-estimated as a new observation becomes available.

Model selection is based on likelihood diagnostics like AIC (Durbin and Koopman 2012, subsec. 7.4) and likelihood-ratio tests (Harvey 1989, subsec. 5.1) and an evaluation of the model assumptions. The likelihood ratio test is used to test restrictions on hyperparameters, for example, to test whether the correlation between the slope disturbance terms in the bivariate model are equal to zero and to whether the variance of the level disturbance terms of the local linear trend model equals zero and thus can be restricted to the smooth trend model. The AIC criteria are used to compare non-nested models, for example, to compare the smooth trend model with the local level model. The state space models considered here assume that all disturbance terms are normally and independently distributed. As a result, the one-step-ahead predictions or innovations should also be normally and independently distributed. The model assumptions can be checked by performing a number of model diagnostics on the standardised innovations: calculation of mean, variance, skewness, kurtosis, Bowman-Shenton-normality test, Ljung-Box test on serial correlation for the first twelve lags, Durbin-Watson test on serial correlation (Durbin and Koopman 2012, subsec. 2.12) and a visual check on outliers, by plotting the standardised innovations. Also, an F-test has been used on the first and last twelve innovations in each of the series to test for heteroscedasticity (Durbin and Koopman 2012, 39).

The time series models are implemented in the OxMetrics software package, in combination with subroutines of SsfPack 3.0. For more information, see Doornik (2009) and Koopman et al. (2008).

4. Results

4.1. Model Parameters for Univariate and Bivariate Models

The univariate and bivariate analyses are based on the models specified in Subsections 3.1 and 3.2. They are applied to two series: the series of LFS municipal small area predictions using the Battese-Harter-Fuller unit level model, as described in Subsection 2.1, and a register series based on tax service data sources that include persons with paid employment and self-employment. The LFS and register series run from the first quarter in 2003 up to and including the final quarter of 2014. Due to the diffuse initialisation of five state variables in the Kalman filter, the first six quarters of the time series are removed from the presentation of the results to allow the Kalman filter to converge to a stable distribution for the state variables. Time series results, therefore, start at the third quarter of 2004 and run up to and including the last quarter of 2014. This period contains 42 quarters.

The performance of the proposed method to estimate the register series is evaluated over the last three years of the observed series in real time. Producing nowcasts in real time is achieved as follows. To obtain nowcasts for the register series for 2012, models are fitted to register series observed until the last quarter of 2011. With the univariate model, applied to the register series only, predictions are made for the four quarters of 2012 based on the register series observed until the last quarter of 2011. With the bivariate model, nowcasts for the four quarters of 2012 are based on the register series observed until the last quarter of 2011 and the LFS observed until the particular quarter of 2012 for which a nowcast of the register is calculated. This process is repeated for 2013 and 2014 using the register series observed until the last quarter of 2012 and 2013 respectively.

Variances and covariances of the innovations, and results for the model diagnostics, in order to establish that the innovations are normally and independently distributed, are included in Appendix (Subsection 6.2) for both the univariate and bivariate models. We also mention that plots of the standardised innovations do not indicate that there are outliers. These diagnostics show that no severe deviations of the underlying model assumptions exist. The maximum likelihood estimates for the additive univariate and bivariate model hyperparameters are presented in Tables 1 and 2. The hyperparameters are variances and are estimated on the log-scale to avoid negative maximum likelihood estimates. As a result, the confidence intervals are asymmetrical. Both the register and the LFS series are at the municipal level.

The small values for the variance of the seasonal disturbance terms for the register series in both the univariate and bivariate model illustrate that the seasonal component hardly changes over time. This also holds for the variance of the seasonal disturbance terms for

Table 1. Maximum likelihood estimates hyperparameters register univariate model (full series) with 95% confidence interval between brackets.

Hyperparameter	Amsterdam	Heerlen	Haren
Trend ($\sigma^R_{\eta,d}$)	0.135 (0.091-0.201)	0.157 (0.108-0.231)	0.150 (0.085-0.267)
Seasonal ($\sigma^R_{\omega,d}$)	0.027 (0.014-0.055)	0.014 (0.005-0.049)	0.027 (0.011-0.067)
Measurement equation ($\sigma^R_{\xi,d}$)	0.128 (0.084-0.197)	0.140 (0.097-0.202)	0.209 (0.140-0.280

Table 2. Maximum likelihood estimates hyperparameters bivariate model (full register series) with 95% confidence interval between brackets.

Hyperparameter	**Amsterdam**	**Heerlen**	**Haren**
Trend	0.134	0.155	0.139
register ($\sigma^R_{\eta,d}$)	(0.091-0.197)	(0.106-0.227)	(0.077-0.250)
Seasonal	0.027	0.014	0.026
register ($\sigma^R_{\omega,d}$)	(0.013-0.054)	(0.004-0.050)	(0.011-0.067)
Measurement equation	0.131	0.142	0.220
register ($\sigma^R_{\xi,d}$)	(0.086-0.198)	(0.099-0.206)	(0.147-0.329)
Trend	0.112	0.128	0.103
LFS ($\sigma^{LFS}_{\eta,d}$)	(0.084-0.157)	(0.089-0.185)	(0.047-0.244)
Seasonal	< 0.001	0.014	0.026
LFS ($\sigma^{LFS}_{\omega,d}$)	(0.000-9.870)	(0.001-2.309)	(0.002-0.384)
Measurement	0.231	<0.001	<0.001
eq. LFS ($\sigma^{LFS}_{\xi,d}$)	(0.057-0.940)	(0.000-1.984)	(0.000-9.430)
Corr. slope	0.899	0.999	0.980
residuals ($\rho_{\eta,d}$)			
Survey error	0.758	0.766	0.578
AR(1)-noise ($\sigma_{\varsigma,d}$)	(0.408-1.408)	(0.612-0.959)	(0.462-0.702)

the LFS series. The variance of the slope disturbance terms for the register series and the LFS series are large enough to give the trend the flexibility to accommodate for cyclic movements, as can be seen for the LFS series in Figure 3. For the register series, the variance of the measurement errors are about the same size as the variances of the slope disturbance terms. For the LFS series, this is only true for Amsterdam. For the other two cities the variance of the measurement errors are almost zero. A possible reason for this is that the sampling error in the two smaller cities (Heerlen and Haren) is much larger than the capital city Amsterdam. It appears that the sampling error dominates the uncertainty in the series of Heerlen and Haren and that the population white noise is absorbed in the sampling error component. The variance estimates for the three components (trend, seasonal and measurement error) of the register series under the univariate model in Table 1 are almost similar to the values under the bivariate model in Table 2.

The AR(1) autoregressive parameter for the sampling errors in the LFS series is estimated from the micro data as described in Subsection 3.1 (its value is 0.59). Recall from Subsection 3.1 that we expect values for the survey error AR(1) noise that are close to $\sigma^2_{\varsigma,d} \approx \left(1 - \rho^2_d\right)$, which implies that $\sigma_{\varsigma,d} \approx 0.8$. This holds reasonably for Amsterdam and Heerlen. The value for Haren is clearly smaller, which implies that the model reduces the MSE estimates for the BHF domain predictions of the LFS input series. The interpretation of the time series model is that the volatility of the LFS input series is smaller than the MSE estimates of the LFS domain predictions simply. One possible reason is that the BHF unit level model over shrinks the domain estimates, in particular for a small municipality like Haren.

In this application, the results obtained with the additive model are very similar to the results obtained under the multiplicative model. In Figure 3, the results under the bivariate additive and bivariate multiplicative models are compared for Heerlen. The confidence interval for the multiplicative model is asymmetric due to the anti-log transformation, but even this is hardly visible. Also, the model diagnostics are very similar under both models.

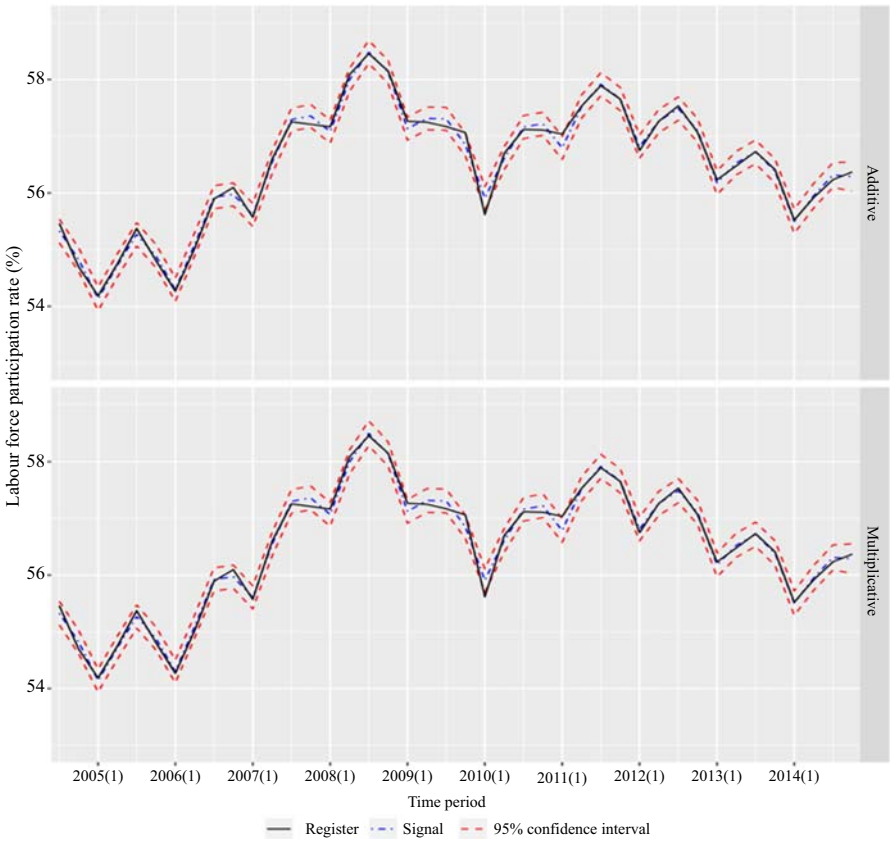

Fig. 3. Comparison Additive and Multiplicative Bivariate Model for Heerlen.

In Appendix (Subsection 6.2) the model diagnostics for the univariate and bivariate multiplicative model are included. In what follows, additive models are used.

The bivariate model detects a strong correlation between the LFS and register slope components in the bivariate analyses. A likelihood ratio test has been applied to test the significance of the correlation between the slope disturbances. If the correlation parameter is set to zero, the likelihood is reduced significantly. For the municipalities Amsterdam, Heerlen and Haren, the p-values of the corresponding likelihood ratio tests are 0.0021, 0.0025, and 0.0500 respectively.

4.2. Results for Univariate and Bivariate Models at Municipal Level

In Figures 4 and 5, two examples are presented of the nowcasting analyses of the univariate and bivariate models for the register series. The figures compare the nowcasts estimated in real time under the univariate and bivariate model with the final observations from the register. For both time series models, three series are compared; one based on the register information available up until the end of 2012, one based on the register information available up until the end of 2013, and one based on the register information available up until the end of 2014. The three time series for the univariate model are plotted in red, while the three time series of the bivariate model are plotted in blue. The end of the time series

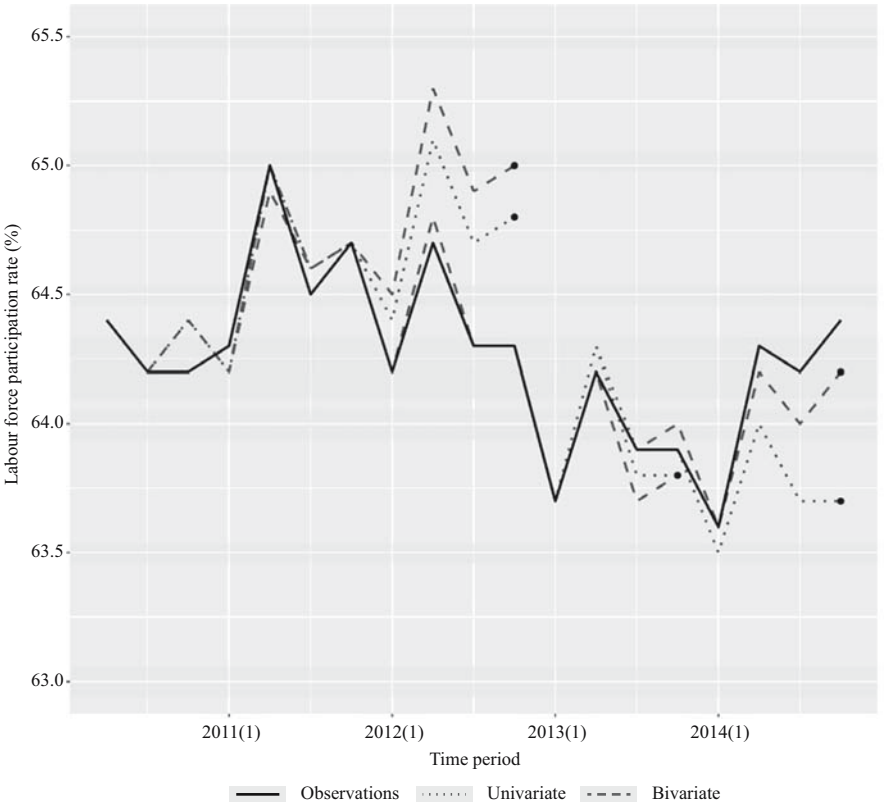

Fig. 4. Amsterdam: Smoothed signal of the labour force participation rate for the register series in the univariate and bivariate models with nowcasts for the last four measurements, compared to the final observations from the register.

estimates under both models are indicated with a red or blue dot. This in real time analysis results give an indication of the size of the revisions if new data become available afterwards.

Both the univariate and bivariate model results show nearly identical smoothed time series up to the last four quarters of 2014. However, for the last four quarters, where one-step, two-step, three-step and four-step ahead predictions were made, the results from the univariate and bivariate analysis differ. For the case of Amsterdam, the bivariate analysis nowcasts for 2014 are closer to the smoothed results of the full series than the nowcasts of the univariate analysis. In the case of Haren (Figure 5), a small municipality in the north of the Netherlands, the results lead to the opposite conclusion.

In Figure 6, standard errors based on the univariate and bivariate models are compared for the smoothed and nowcasted signal of the labour force participation rates in Amsterdam, Heerlen and Haren. For reference, the standard errors of the small area estimates of the LFS labour force participation rates are also presented.

The gain in accuracy obtained with the bivariate model with respect to the univariate model is approximately 20% in Amsterdam, Heerlen and Haren for 2014. For Haren, in 2012 and 2013 there is hardly any difference. The smaller gain in accuracy for Haren may be related to the smaller correlation between the register and LFS slope disturbance terms. However, the variance of the LFS small area estimates may also be a contributing factor. For Amsterdam,

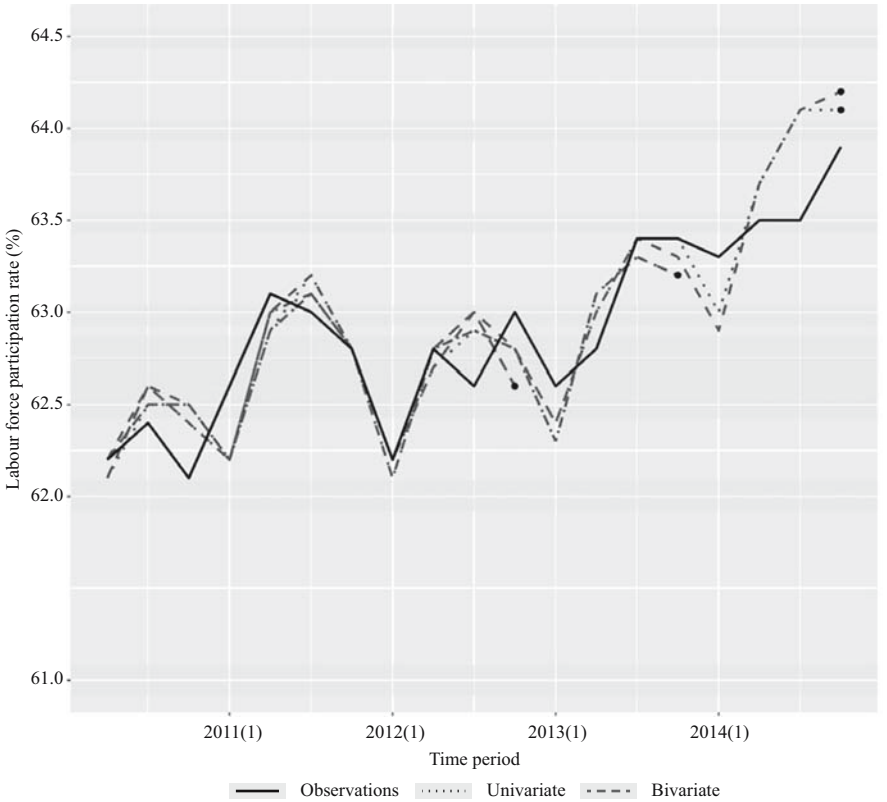

Fig. 5. Haren: Smoothed signal of the labour force participation rate for the register series in the univariate and bivariate models with nowcasts for the last four measurements, compared to the final observations from the register.

the standard errors of the time series nowcasts are comparable with the standard errors of the LFS small area predictions. For smaller municipalities, the nowcasted results of the register series are more accurate than the LFS estimates of the labour force participation rate.

As a second analysis, the time series of the GREG estimates of the labour force participation rate at the national level is used as an auxiliary series to nowcast the Haren register series. Although the correlation between these two series is smaller (0.783), the reduction in standard error in going from the univariate to the bivariate analysis is clearly larger: 39%. The relatively small standard errors of the national LFS estimates of the labour force participation rate seem to be an important contributing factor; they are much smaller than those of the Haren LFS auxiliary series (0.15 % point instead of 1.4% point).

This suggests that at least for small municipalities, where the variance of the LFS series is large and the correlation between the register series and the national LFS estimates is not too small, it may be better to use the national LFS series instead of the municipal series. Using auxiliary series at high aggregation levels, for example, national or province level, however, might introduce larger amounts of bias in the model predictions.

In a next step, the univariate and bivariate models are applied to all municipalities in the Netherlands. The estimated standard errors of the predicted signals of both the univariate and bivariate time series models are further evaluated with the coverage rates of the

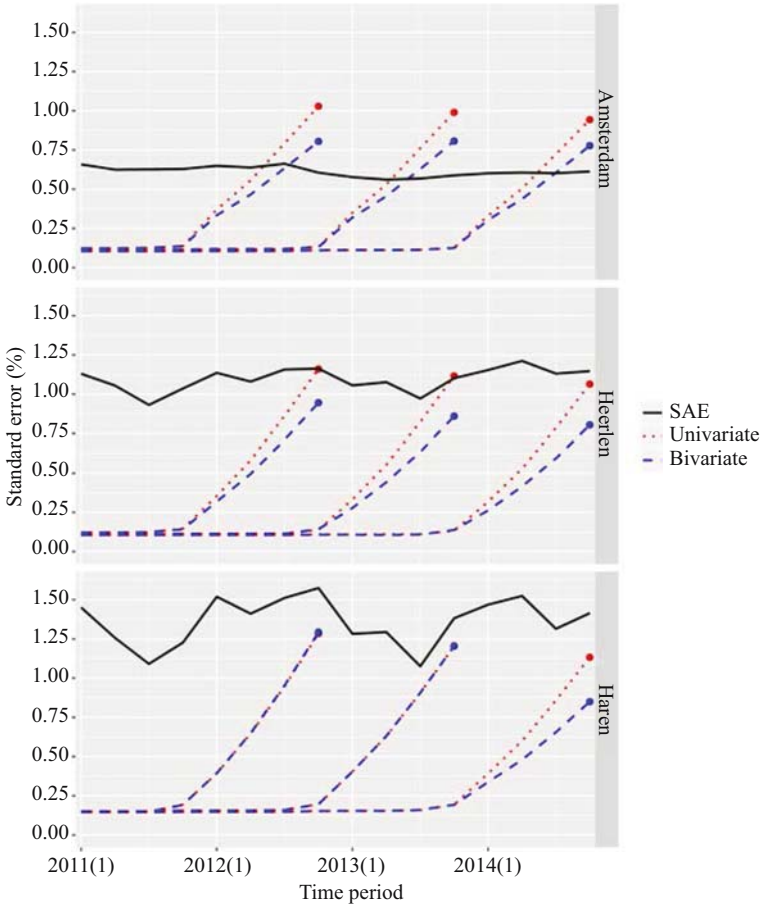

Fig. 6. Standard errors of predicted signal of the municipal labour force participation rate for the register series in the univariate and bivariate models compared with standard errors of small area estimates (nowcasting in each of the last three years of the time series).

corresponding 95% confidence intervals. For both time series models, the proportion of municipalities for which the measured value of the labour force participation rate is to be found within the estimated 95%-confidence interval is calculated. The results for the 2014 nowcasts are presented in Table 3.

Table 3. Coverages of the one-step to four-step ahead predictions of the register labour force participation rate (municipalities) for the univariate and bivariate time series models, 2014 (nowcasts for all municipalities in the Netherlands).

Type of estimate	**Univariate %**	**Bivariate [1]**	**Bivariate [2]**
1-step ahead	95.8	92.6	90.0
2-step ahead	99.0	95.1	94.9
3-step ahead	99.3	95.3	96.3
4-step ahead	99.3	94.9	97.5

[1]using municipal small area predictions as auxiliary series.
[2]using national GREG estimates as auxiliary series.

Table 4. Root relative mean square error of the one-step to four-step ahead predictions of the register labour force participation rate (municipalities) for the univariate and bivariate time series models, 2014 nowcasts.

RRMSE	Univariate %	Bivariate [1]	Bivariate [2]
1-step ahead	0.54	0.55	0.56
2-step ahead	0.72	0.68	0.66
3-step ahead	0.95	0.91	0.81
4-step ahead	1.31	1.19	1.04

[1]using municipal small area predictions as auxiliary series.
[2]using national GREG estimates as auxiliary series.

It turns out that the confidence regions of the univariate nowcasts are approximately equal to 95% for the one-step-ahead prediction but are too wide for the two-, three- and four-step ahead predictions. For the bivariate time series model (1) the confidence intervals are too small for the one-step-ahead predictions and approximately correct for the two-, three- and four-step ahead predictions. For the bivariate model (2) only the two-step ahead predictions are approximately correct.

To compare the accuracy of estimates obtained with the univariate and bivariate model, the root relative mean square errors (RRMSE's) are calculated:

$$RRMSE_{t,j} = 100\% \times \frac{1}{n_{mun}} \sqrt{\sum_{n_{mun}}^{d=1} \left(\frac{y^R_{d,t+j|t} - y^R_{d,t+j|t+4}}{y^R_{d,t+j|t+4}} \right)^2}$$

with $y^R_{d,t+j|t}$ for $j = 1, \ldots, 4$, the prediction or nowcast for the register labour force participation rate in municipality d for period $t + j$ based on the information observed until period t, $y^R_{d,t+j|t+4}$ the register value for the register labour force participation rate in municipality d as it eventually becomes available, and n_{mun} the number of municipalities. In Table 4, the RRMSEs of the 2014 nowcasts for the univariate and bivariate models are presented. Using this measure, the register estimates for the labour force participation rate in the bivariate model are more accurate than the univariate estimates (except for the first step ahead prediction), although the gain in accuracy seems less than as suggested by the estimated standard errors. Furthermore, these results show that using the national LFS labour force participation rates as an auxiliary series gives better estimates for the nowcasts.

4.3. Results for Univariate and Bivariate Models at Neighbourhood Level

In this subsection, the univariate and bivariate modelling approach for nowcasting the register labour force participation rates is extended to the level of neighbourhoods. As it is not practical to perform this kind of analysis on all neighbourhoods, municipality Heerlen is selected as an example and several of its neighbourhoods are considered in the analysis. In Table 5, the selected neighbourhoods from Heerlen together with a number of their characteristics are presented.

Figure 7 compares the measured labour force participation rates for several Heerlen neighbourhoods (indicated by 'Observations') with the predicted signals of the bivariate time series analysis (indicated by 'Bivariate'. Similar to Figures 4 and 5, this figure

Table 5. Selection of ten neighbourhoods in municipality Heerlen (January 2014).

Ten Heerlen neighbourhoods	Total	Sex		Age					Migration background	
	count	Men	Women	0–15	15–25	25–45	45–65	65 or older	Western	Non-Western
				%						
Heerlen (municipality)	88260	43780	44480	13	11	24	31	21	19	9
Maria Gewanden	3690	1905	1785	13	11	25	35	16	18	8
Weggebekker	400	195	200	20	9	29	27	14	22	17
Heksenberg	2410	1250	1155	14	11	25	33	16	20	7
Zeswegen	2350	1140	1210	19	13	31	30	7	25	24
Meezenbroek	2900	1410	1485	17	9	26	29	19	22	17
Schaesbergerveld	2515	1305	1210	14	12	30	30	14	18	13
't Loon	725	325	395	2	9	17	22	49	21	4
Eikenderveld	2555	1285	1265	11	10	32	29	18	18	14
Bekkerveld	1490	715	770	16	11	23	30	20	13	3
Douve Weien	3670	1700	1965	9	8	29	30	24	16	3

See Appendix (Subsection 6.1) for definitions of Western and non-Western immigrants.

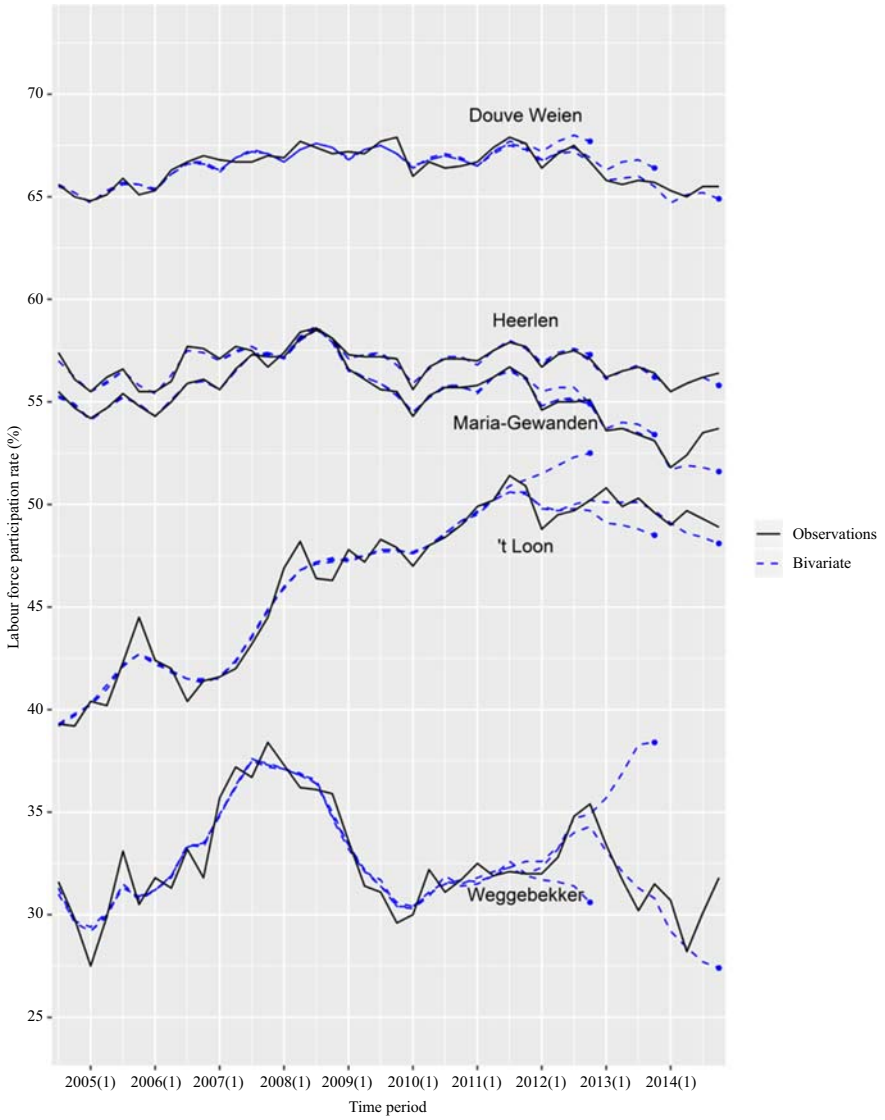

Fig. 7. Time series labour force participation rates for Heerlen and several of its neighbourhoods, comparing register measurements with predicted signals from the bivariate model (series run from third quarter of 2004 to the last quarter of 2014, with observations for each of the last three years).

compares the nowcasts estimated in real time using the bivariate model with the final observations from the register. To this end, three series of time series estimates are compared; one based on the information available up until the end of 2012, one based on the information available up until the end of 2013, and one based on the information available up until the end of 2014. The end of the time series estimates under the bivariate model are indicated with a blue dot. For some neighbourhoods the nowcasts are close to the final values of the register, for example, in the case of Douve Weien and Maria Gewanden. These neighbourhoods typically follow the municipal pattern. For other neighbourhoods, for example, Weggebekker, the model fails to predict unexpected turning

points. This might be explained by the fact that Weggebekker is a small and rather atypical Heerlen neighbourhood. It contained 365 inhabitants in 2018, which partly explains the volatility of the labor force participation rates over the years; also other characteristics deviate from the municipal average (e.g., in 2018: home ownership is only 5% compared to 46% in all of Heerlen; and the percentage of inhabitants with social benefits due to disability or unemployment is 30%, compared to 16% in Heerlen municipality).

In order to evaluate the accuracy of the univariate and bivariate nowcasts for these neighbourhoods, a number of comparisons have been produced. First, like in Subsection

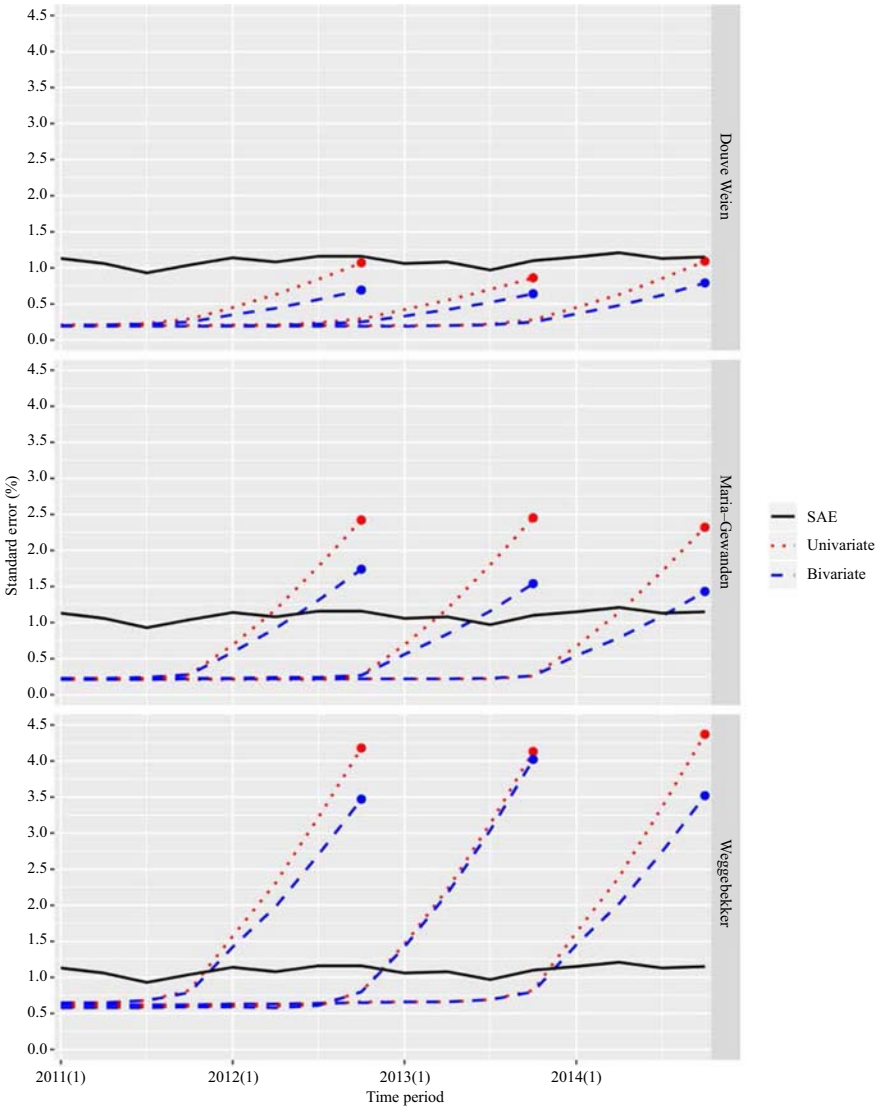

Fig. 8. Standard errors of the predicted signals of neighbourhood labour force participation rates for the register series at the neighbourhood level in the univariate and bivariate models compared with standard errors of small area estimates at the corresponding municipal level (nowcasting in each of the last three years of the time series). Results are shown for three neighbourhoods within municipality Heerlen.

4.2, the standard errors of the nowcasts obtained with both the univariate and bivariate time series models are compared with one another, in this case for several Heerlen neighbourhoods (see Figure 8). For reference, the standard errors of the LFS small area estimates for municipality Heerlen, obtained with the unit-level model, are included in the presentation. As with the municipal series, the standard errors obtained with the bivariate models for the Heerlen neighbourhoods are smaller than the standard errors of the corresponding univariate models. The standard errors for the nowcasts of the two larger neighbourhoods (Douve Weien and Maria-Gewanden) are of the same order of magnitude as the standard errors of the LFS municipal small area estimates for Heerlen. For neighbourhood Weggebekker, a small neighbourhood, the standard errors of both the univariate and bivariate analyses are relatively large (factor three to four times the standard errors of the LFS municipal small area estimates).

In Table 6, the coverage rates of the 95% confidence intervals for the neighbourhood nowcasts in 2014 are compared for several models: the univariate model, and the bivariate model where either the LFS series for Heerlen is used or the LFS series at the national level. Coverage rates for all three approaches are smaller than expected. This might be a result of the erratic behaviour of the register series at the very detailed level of neighbourhoods.

In Table 7, the RRMSE for the nowcasts of the Heerlen neighbourhoods in 2014 are compared for the univariate and the bivariate models. The smallest RRMSEs are obtained with the bivariate model where the register series is combined with the LFS series at the national level. The results for the bivariate model with municipal small area estimation predictions as auxiliary variables are comparable with the results obtained with the univariate model.

Table 6. Coverages of the one-step to four-step ahead predictions of the register labour force participation rate in Heerlen neighbourhoods for the univariate and bivariate time series models, 2014 nowcasts.

Type of estimate	Univariate %	Bivariate [1]	Bivariate [2]
1-step ahead	85.7	83.9	82.1
2-step ahead	80.4	80.4	75.0
3-step ahead	85.7	87.5	71.4
4-step ahead	92.9	85.7	82.1

[1]using municipal small area predictions as auxiliary series.
[2]using national GREG estimates as auxiliary series.

Table 7. Root relative mean square error of the one-step to four-step ahead predictions of the register labour force participation rates in Heerlen neighbourhoods for the univariate and bivariate time series models, 2014 nowcasts.

RRMSE	Univariate %	Bivariate [1]	Bivariate [2]
1-step ahead	4.78	5.38	4.67
2-step ahead	5.67	5.67	4.18
3-step ahead	6.28	6.26	4.55
4-step ahead	5.76	6.24	5.41

[1]using municipal small area predictions as auxiliary series.
[2]using national GREG estimates as auxiliary series.

4.4. Nowcasting During the COVID-19 Pandemic

An important issue with nowcasting is how accurate turning points are detected. To evaluate how well the nowcasts with the bivariate model pick up turning points in the target series, the nowcast exercise is extended to the COVID-19 pandemic for the two municipalities Amsterdam and Heerlen. The municipality Haren merged with another large municipality (Groningen) in 2019 and is therefore left out of this analysis. The Netherlands went into a lockdown at 16 March 2020. Any effect of the lockdown on the employed labour force will therefore be visible from the second quarter of 2020 onwards. The auxiliary series for the LFS are available up to the second quarter of 2020. The register series are available up to the fourth quarter of 2018. This implies that nowcasts are produced for the last six quarters. Nowcasts are calculated using the bivariate state space model with the LFS municipal small area predictions and the LFS GREG estimates at the national level being used as auxiliary series.

In Figure 9, the nowcasts of the register series for Amsterdam and Heerlen obtained with the bivariate model and the LFS municipal small area predictions as an auxiliary series, are compared with the LFS municipal small area prediction and the register series observed up to the fourth quarter of 2018. In Figure 10, the nowcasts of the register series for Amsterdam and Heerlen obtained with the bivariate model and GREG estimates of the LFS at the national level as the auxiliary series are compared with the LFS municipal small

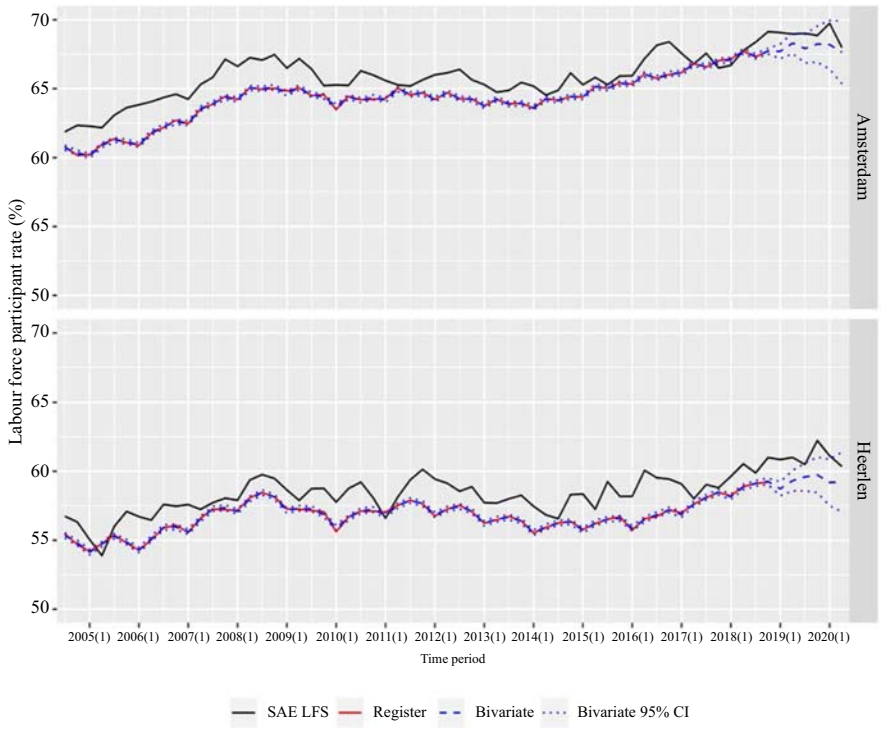

Fig. 9. Nowcasting the register series in the COVID-19 pandemic for the municipalities Amsterdam and Heerlen (municipal small area estimates as auxiliary series).

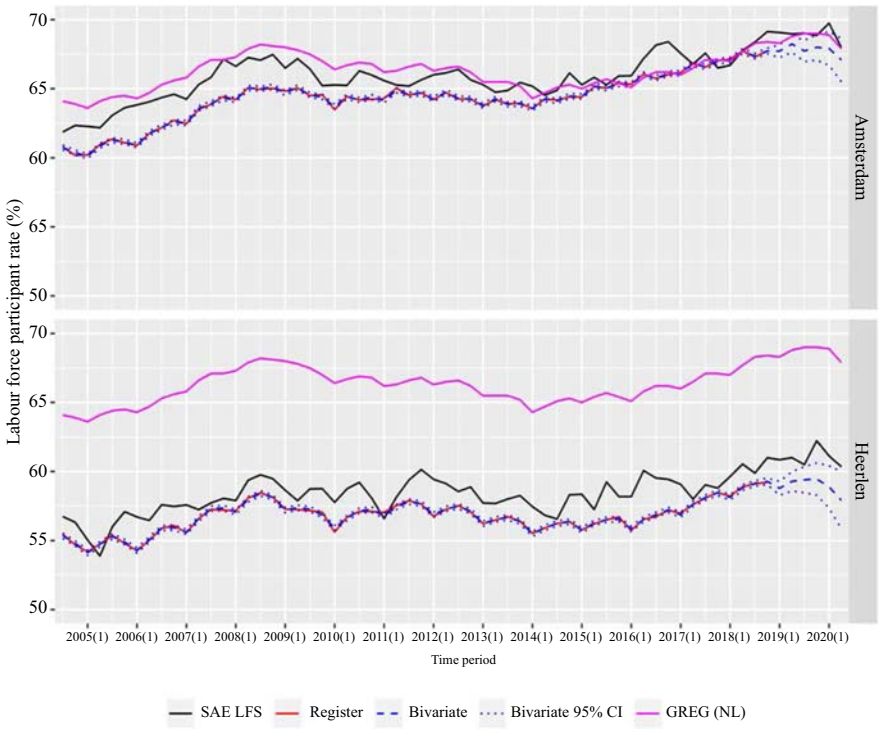

Fig. 10. Nowcasting the register series in the COVID-19 pandemic for the municipalities Amsterdam and Heerlen (using the GREG LFS estimates at national level as auxiliary series).

area prediction and the register series observed up to the fourth quarter of 2018. For reference, the LFS national GREG estimates are also included.

If the LFS GREG estimates at the national level are used as an auxiliary series, a correlation between the slope disturbance terms of 0.9 is found for Amsterdam and Heerlen. If the municipal small area predictions are used as an auxiliary series, the correlation between the slope disturbance terms is 0.9 again for Amsterdam and 1.0 for Heerlen. The LFS GREG estimates at the national level as well as the municipal small area predictions for Heerlen and Amsterdam show a clear drop for the employment rate in the second quarter of 2020. This drop is clearly picked up by the nowcasts for Amsterdam and Heerlen if the LFS GREG estimates at the national level are used as an auxiliary series. If the LFS municipal small area prediction are used as an auxiliary series, then the nowcast for Amsterdam still picks up a clear, but less pronounced, drop for the second quarter of 2020. The nowcast for Heerlen does not show a drop for the last quarter, despite the correlation between the slope disturbance terms being larger if the municipal small area prediction are used as an auxiliary series. The reason behind this observation is that the standard errors of the LFS GREG estimates at the national level are clearly smaller compared to the MSEs of the municipal small area predictions. Using an auxiliary series at a high aggregation level, however, might result in larger bias in the nowcasts. At the moment of doing this nowcast exercise, it is not possible to evaluate the nowcasts, since the final values of the register series are not available.

5. Discussion

An advantage of using registers for the production of official statistics rather than sample surveys is that they allow for the compilation of very accurate statistics at very detailed breakdowns, since it approximates a complete enumeration of the finite target population. In addition, data collection costs are small compared to sample surveys and do not contribute to response burden. Examples in the Netherlands in which reliable information for statistical purposes is obtained from administrative data are income statistics, statistics on poverty rates, short term business statistics and statistics on the labour force participation, since they are all derived from tax administrations and administrations on social benefits. On the other hand, these data sources are not always timely. This applies, in particular, to tax administrations on income data.

In this article, a nowcast method is proposed to improve the timeliness of very detailed regional statistics on the labour force participation rate derived from a tax register, by using more timely survey estimates obtained from the labour force survey (LFS). The LFS data for a particular reference year become available eight months in advance of the preliminary register statistics, and one year and eight months in advance of the final register statistics. As a first step, estimates for quarterly municipal figures on labour force participation rates are constructed using the LFS data with a small area estimation technique that is based on a cross-sectional small area estimation method. In this application, the unit level model of Battese et al. (1988) is used. This gives rise to a time series of timely quarterly small area estimates at the municipal level. These series are combined with the more precise, but less timely, quarterly series on labour force participation rates derived from the tax register in a bivariate structural time series model. The model uses the correlation between the disturbance terms of the stochastic trend component of the labour force participation rates in the survey and register sources. This idea was introduced by Harvey and Chung (2000) the other way around, that is, to improve UK LFS estimates with claimant counts as an auxiliary series. Van den Brakel et al. (2017) used a similar state space model to improve the timeliness of the Dutch consumer confidence survey with a related more timely series derived from social media platforms. The model proposed in this article accounts for the uncertainty in the auxiliary series by modelling the sampling error in the small area estimates of the LFS as well as the autocorrelation in the sampling error due to the panel overlap.

The proposed methods are applied to estimate labour force participation rates at both the municipal and neighbourhood level using the register data with a one year lag as the main series and the LFS small area estimates as the auxiliary series. The introduction of an auxiliary survey series at the municipal level decreases the standard errors of the predicted measurements of the one-year lag (nowcasts) on average by 20%, both at the municipal and at the neighbourhood level. The gains in terms of RRMSEs are smaller. The reduction in the variance of the predicted estimates hinges on two contributing factors: (1) the strength of the correlation between the register and survey series and (2) the amount of uncertainty in the small area estimates of the auxiliary series.

By choosing a series that correlates well with the main series and has a small variance component, for example, LFS estimates for labour force participation rates at the national level, the reduction in the standard errors in the bivariate analysis may be as large as 40%.

On the other hand, using a noisy auxiliary series in the bivariate analysis does not lead to large improvements in the accuracy of the predicted estimates. For smaller municipalities, replacing the municipal auxiliary LFS series of the labour force participation rate by a series at a higher regional level (province or even at the national level) may therefore be a better choice.

Predictions of labour force participation rates from the time series modelling approach at the regional level (municipalities and neighbourhoods) were compared with the actual, final measurements. Predictions root relative mean squares measures were calculated for both the univariate and bivariate modelling results. These measures show that in the case of a noisy (municipal) auxiliary series on labour force participation rates, the bivariate time series model does not produce more accurate nowcasts than the univariate model. When using a less volatile auxiliary series on labour force participation at a higher regional level (the national level), the bivariate model produces more accurate results than the univariate model.

The proposed method is extended to the start of the COVID-19 pandemic. It is found that the extent to which the nowcasts pick up the turning point induced by the lockdown of the Corona crisis depends on the accuracy of the auxiliary series. If the series of LFS GREG estimates at the national level with small standard errors are used, then the nowcasts pick up the turning point induced by the lockdown. If municipal small area predictions are used, then the turning point in the nowcasts is less pronounced because these auxiliary series have larger MSEs.

In the approach discussed in this article, the explanatory variable is observed with sampling error. From this point of view, the approach followed in this article has some similarities with the area level models used in small area estimation, where the auxiliary variables are observed with measurement error (Ybarra and Lohr 2008). In particular, the bivariate structural time series model in this article can be seen as a temporal version of the structural measurement error model, as discussed in Bell et al. (2019).

The method introduced in this article may be used for other applications where lagged register statistics can be nowcasted by using correlated auxiliary series observed from a more timely sample survey. One research line is to develop multivariate structural time series models in which the register series for all neighbourhoods within a municipality are combined with the LFS series at a higher regional level. Such models can also account for and profit from the correlation between the register series of the neighbourhoods. High-dimensionality problems will likely occur but can be handled with a dynamic factor model in state space form, where common factors are derived in a first step using principal components (see e.g., Giannone et al. 2008). Another research direction would be to extend the small area estimation approach for LFS figures to smaller regional domains using temporal and cross-sectional correlations. Finally, the (bivariate) time series model can also be applied to improve the accuracy of the LFS series as well, even if this series contains more timely data. Van den Brakel and Krieg (2016) showed that monthly estimates for the unemployed labour force derived from LFS data can be improved with claimant count series, even if the last two months of claimant counts are not available. Analyses on municipal labour force participation rates suggest that a univariate time series modelling approach can reduce the standard errors of the survey series small area estimates. Bivariate analyses with lagged register data seem to lead to further small improvements.

6. Appendix

6.1. Explanation Variables

Employed Reg: employed according to register (Polis-register);
- 0: not employed
- 1: employed

Sex:
- 1: men
- 2: women

Age (5): age in five classes:
- 1: 15–24
- 2: 25–34
- 3: 35–44
- 4: 45–54
- 5: 55–75

Age (3): age in three classes:
- 1: 15–24
- 2: 25–44
- 3: 45–75

Ethnicity (3): migration background in three classes;
- 1: Dutch background
- 2: Western background[1]
- 3: Non-western background[2]

Ethnicity (7): migration background in seven classes;
- 1: Dutch background
- 2: Moroccan
- 3: Turkish
- 4: Surinam
- 5: Antillen/Aruba
- 6: Other non-Western[2]
- 7: Western background[1]

Employment Office:
- 1: Not registered at the Employment Office
- 2: Registered at the employment office with a job and searching for work
- 3: Registered as unemployed less than one year
- 4: Registered as unemployed for one to four years
- 5: Registered as unemployed for more than four years

Border municipality (2) municipality is close to the border:
- 1: Not a border municipality
- 2: Is a border municipality

HH Type: Household type in three classes:
- 1: One person household
- 2: Household with children
- 3: Other household types

WaveNr.: wave number in the panel (1, . . ., 5)

[1]Western immigrant are defined as: all European countries excluding Turkey, North America, Canada, Australia, New-Zealand, Japan and Indonesia.

[2]Non-Western immigrant are defined as: Turkey, African countries, Latin America, all Asian countries with the exception of Japan and Indonesia.

6.2. Model Diagnostics (2003–2014 Series)

Meaning of table columns:
Value = value of test statistic
L = lower bound of 95%-confidence interval (n.a. = not applicable)
U = upper bound of 95%-confidence interval

Table 8. Univariate model register.

Diagnostic	**Amsterdam**			**Heerlen**			**Haren**		
	Value	L	U	Value	L	U	Value	L	U
Mean	0.089			0.051			0.039		
Variance	0.991			0.997			0.998		
Skewness	0.689			0.353			0.558		
Kurtosis	4.172			2.894			3.590		
Bowman-Shenton[1]	5.870		5.991	0.916		5.991	2.861		5.991
Ljung-Box[2]	4.278		21.02	8.953		21.02	10.16		21.02
Durbin-Watson[3]	1.979	1.402	2.597	1.932	1.402	2.597	1.807	1.402	2.597
F-test heteroscedasticity[4]	11.40*	0.305	3.277	2.201	0.305	3.277	1.768	0.305	3.277

*test statistic outside 95% confidence interval under null hypothesis.

Table 9. Bivariate model register innovations.

Diagnostic	**Amsterdam**			**Heerlen**			**Haren**		
	Value	L	U	Value	L	U	Value	L	U
Mean	0.107			0.043			0.094		
Variance	0.976			0.991			0.915		
Skewness	0.683			0.417			0.608		
Kurtosis	3.927			3.044			3.832		
Bowman-Shenton[1]	4.766		5.991	1.224		5.991	3.804		5.991
Ljung-Box[2]	4.526		21.02	10.39		21.02	9.010		21.02
Durbin-Watson[3]	1.988	1.395	2.604	1.901	1.395	2.604	1.858	1.395	2.604
F-test heteroscedasticity[4]	7.151*	0.305	3.277	1.336	0.305	3.277	1.748	0.305	3.277

*test statistic outside 95% confidence interval under null hypothesis.

Table 10. Bivariate model LFS innovations.

Diagnostic	**Amsterdam**			**Heerlen**			**Haren**		
	Value	L	U	Value	L	U	Value	L	U
Mean	-0.117			-0.100			-0.083		
Variance	1.028			0.951			0.933		
Skewness	0.406			0.070			0.068		
Kurtosis	2.644			2.953			2.108		
Bowman-Shenton[1]	1.378		5.991	0.038		5.991	1.424		5.991
Ljung-Box[2]	13.05		21.02	12.63		21.02	13.76		21.02
Durbin-Watson[3]	1.880	1.395	2.604	1.688	1.395	2.604	1.569	1.395	2.604
F-test heterosce-dasticity[4]	1.148	0.305	3.277	2.363	0.305	3.277	1.092	0.305	3.277

Table 11. Autocorrelations and cross correlation of standardised innovations (bivariate model).

Type of correlation	**Amsterdam**	**Heerlen**	**Haren**
Autocorrelation (register innovations)			
Lag 1	0.000	0.016	0.043
Lag 2	-0.038	0.050	-0.223
Lag 3	0.130	0.110	0.013
Cross correlation (register-LFS innovations)	0.166	0.075	0.132

Table 12. Bivariate model register innovations (multiplicative model).

Diagnostic	**Heerlen**		
	Value	L	U
Mean	0.043		
Variance	0.989		
Skewness	0.398		
Kurtosis	3.041		
Bowman-Shenton[1]	1.112		5.991
Ljung-Box[2]	10.11		21.02
Durbin-Watson[3]	1.890	1.395	2.604
F-test heteroscedasticity[4]	1.438	0.305	3.277

Table 13. Bivariate model LFS innovations (multiplicative model).

Diagnostic	**Heerlen**		
	Value	L	U
Mean	-0.100		
Variance	0.951		
Skewness	0.058		
Kurtosis	3.080		
Bowman-Shenton[1]	0.035		5.991
Ljung-Box[2]	11.94		21.02
Durbin-Watson[3]	1.707	1.395	2.604
F-test heteroscedasticity[4]	2.438	0.305	3.277

[1]Bowman-Shenton test on normality of the standardised innovations: distribution under H_0: χ^2_2
[2]Liung-Box test on autocorrelation of the first eight lags in the standardised innovations: distribution under H_0: χ^2_7
[3]Durbin-Watson test on autocorrelation in the standardised residuals: distribution under H_0 approximated with $N\left(2, \frac{4}{T}\right)$
[4]F-test for heteroscedasticity in the standardised innovations: distribution under H_0: χ^{11}_{11}

7. References

Arima, S., W.R. Bell, G.S. Datta, C. Franco, and B. Liseo. 2017. "Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error." *Journal of the Royal Statistical Society* Series A, 180: 1191–1209. DOI: https://cran.r-project.org/web/packages/hbsae/hbsae.pdf.

Bakker, B.F.M. 2012. "Estimating the Validity of Administrative Variables." *Statistica Neerlandica*, 66: 8–17. DOI: https://doi.org/10.1111/j.1467-9574.2011.00504.x.

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An error components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association*. 83: 28–36. DOI: https://doi.org/10.2307/2288915.

Bell, W.R., H.C. Chung, G.S. Datta, and C. Franco. 2019. "Measurement error in small area estimation: Functional versus structural versus naïve models." *Survey Methodology* 45: 61–80. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2019001/article/00005-eng.htm (accessed August 2020).

Bijlsma, I., J.A. van den Brakel, R. van der Velden, and J. Allen. 2020. "Estimating literacy levels at a detailed regional level; An application using Dutch data." *Journal of Official Statistics,* 36: 251–274. DOI: http://dx.doi.org/10.2478/JOS-2020-0014.

Blight, B.J.N., and A.J. Scott. 1973. "A stochastic model for repeated surveys." *Journal of the Royal Statistical Society* Series B, 35: 61–66. DOI: https://doi.org/10.1111/j.2517-6161.1973.tb00936.x.

Bollineni-Balabay, O., J.A. van den Brakel, and F. Palm. 2017. "State space time series modelling of the Dutch Labour Force Survey: Model selection and mean squared error estimation." *Survey Methodology,* 43: 41–67. Available at: http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14819-eng.htm (accessed November 2019).

Boonstra, H.J. 2012. hbsae: Hierarchical Bayesian Small Area Estimation.R package version 1.0. https://cran.r-project.org/web/packages/hbsae/hbsae.pdf (accessed February 2020).

Boonstra, H.J. and J.A. van den Brakel. 2019. "Estimation of level and change for unemployment using structural time series models." *Survey Methodology*, 45: 395–425. Available at: http://www.statcan.gc.ca/pub/12-001-x/2019003/article/00005-eng.htm (accessed February 2020).

Boonstra, H.J., J.A. van den Brakel, B. Buelens, S. Krieg, and M. Smeets. 2008. "Towards small area estimation at Statistics Netherlands." *Metron*, LXVI (1): 21-49.

Boonstra, H.J., B. Buelens, and M. Smeets. 2007. *Estimation of municipal unemployment fractions – a simulation study comparing different small area estimators*. Technical report, BPA-no. DMH-2007-04-20-HBTA, Statistics Netherlands, Heerlen. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjgldbMivrzAhVKDuwKHW-jCQEQFnoECAoQAQ&url=https%3A%2F%2Fwww.cbs.nl%2F-%2Fmedia%2Fimported%2Fdocuments%2F2011%2F02%2F2011-x10-02.pdf%3Fla%3Dnl-nl&usg=AOvVaw3n58-noqm4X0g4z5v28QSL (accessed June 2020).

Boonstra, H.J., B. Buelens, K. Leufkens, and M. Smeets. 2011. *Small area estimates of labour status in Dutch municipalities*. Technical Report 201102, Statistics Netherlands.

CBS Statline (2017). Statline statistical database: Avaliable at: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83524NEDtable?dl = 40C7/ (accessed November 2021).

Choi, H., and H. Varian. 2012. "Predicting the Present with Google Trends." *Economic Record* 88 Supplement s1: 2–9. DOI: https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Cochran, W. 1977. *Sampling Theory*. New York: John Wiley & Sons, Inc.

Daas, P., and M. Puts. 2014. "Big data as a source of statistical information." *The Survey Statistician,* 69: 22–31. Available at: http://isi-iass.org/home/wp-content/uploads/n69-2014-01-issn.pdf (accessed November 2019).

Datta, G.S., and M. Ghosh. 1991. "Bayesian Prediction in Linear Models: Applications to Small Area Estimation." *The Annals of Statistics*, 19(4): 1748–1770. DOI: https://doi.org/10.1214/aos/1176348369.

Datta, G., P. Lahiri, T. Maiti, and K. Lu. 1999. "Hierarchical Bayes estimation of unemployment rates for the states of the U.S." *Journal of the American Statistical Association* 94: 1074–1082. DOI: https://doi.org/10.2307/2669921.

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

Durbin, J., and S.J. Koopman. 2012. *Time Series Analysis by State Space Methods* (2nd edition). Oxford: Oxford University Press.

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: 269–277. DOI: https://doi.org/10.1080/01621459.1979.10482505.

Giannone, D., L. Reichlin, and D. Small. 2008. "Nowcasting: The real-time informational content of macroeconomic data." *Journal of Monetary Economics* 55: 665–676. DOI: https://doi.org/10.1016/j.jmoneco.2008.05.010.

Hand, D. 2018. "Statistical challenges of administrative and transaction data." *Journal of the Royal Statistical Society* series A, 181: 1–51. DOI: https://doi.org/10.1111/rssa.12315.

Harvey, A.C. 1989. *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Harvey, A.C., and C.H. Chung. 2000. "Estimating the underlying change in unemployment in the UK." *Journal of the Royal Statistical Society* Series A, 163: 303–339. DOI: https://doi.org/10.1111/1467-985X.00171.

Hobza, T., and D. Morales. 2016. "Empirical best prediction under unit-level logit mixed models." *Journal of Official Statistics*, 32: 661–692. DOI: https://doi.org/10.1515/jos-2016-0034.

Hobza, T., D. Morales, and L. Santamaria. 2018. "Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models." *TEST,* 27: 270–294. DOI: https://doi.org/10.1007/s11749-017-0545-3.

Koopman, S.J., A.C. Harvey, J.A. Doornik, and N. Shephard. 2007. *STAMP8: Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake.

Koopman, S.J., N. Shephard, and J.A. Doornik. 2008. *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*, London: Timberlake Consultants Press.

MacGibbon, B., and T.J. Tomberlin. 1989. "Small Area Estimates of Proportions via Empirical Bayes Techniques." *Survey Methodology*, 15: 237–252. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1989002/article/14565-eng.pdf (accessed August 2021).

Malec, D., J. Sedransk, C.L. Moriarity, and F.B. Leclere. 1997. "Small Area Inference for Binary Variables in the National Health Interview Survey." *Journal of the American Statistical Association*, 92: 815–826. DOI: https://doi.org/10.1080/01621459.1997.10474037.

Marino, M.F., M.G. Ranalli, N. Salvati, and M. Alfo. 2019. "Semiparametric empirical best prediction for small area estimation of unemployment indicators." *The Annals of Applied Statistics*, 13: 1166–1197. DOI: https://doi.org/10.1214/18-AOAS1226.

Pfeffermann, D. 2002. "Small area estimation – new developments and directions." *International Statistical Review* 70: 125–143. DOI: https://doi.org/10.1111/j.1751-5823.2002.tb00352.x.

Pfeffermann, D. 2013. "New important developments in small area estimation." *Statistical Science* 28: 40–68. DOI: https://doi.org/10.1214/12-STS395.

Pfeffermann, D., and L. Burck. 1990. "Robust Small Area Estimation combining Time Series and Cross-sectional Data." *Survey Methodology*, 16: 217–237. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1990002/article/14534-eng.pdf?st = pRQklC03 (accessed October 2019).

Pfeffermann, D., J.L. Eltinge, and L.D. Brown. 2015. "Methodological issues and challenges in the production of official statistics." *Journal of Survey Statistics and Methodology* 3: 425–483. DOI: https://doi.org/10.1093/jssam/smv035.

Pfeffermann, D., and R. Tiller. 2005. "Bootstrap approximation to prediction MSE for state-space models with estimated parameters." *Journal of Time Series Analysis,* 26: 893-916. DOI: https://doi.org/10.1111/j.1467-9892.2005.00448.x.

Pfeffermann, D., and R. Tiller. 2006. "Small area estimation with state space models subject to benchmark constraints." *Journal of the American Statistical Association* 101: 1387–1397. DOI: https://doi.org/10.1198/016214506000000591.

Rao, J.N.K., and I. Molina. 2015. *Small Area Estimation*, second edition. New York: John Wiley & Sons.

Rao, J.N.K., and M. Yu. 1994. "Small area estimation by combining time-series and cross-sectional data." *The Canadian Journal of Statistics* 22: 511–528. DOI: https://doi.org/10.2307/3315407.

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

Van den Brakel, J.A., E. Söhler, P. Daas, and B. Buelens. 2017. "Social media as a data source for official statistics; the Dutch Consumer Confidence Index." *Survey Methodology* 43: 183–210. Available at: http://www.statcan.gc.ca/pub/12-001-x/2017002/article/54871-eng.htm (accessed October 2019).

Van den Brakel, J.A., and S. Krieg. 2015. "Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design." *Survey Methodology* 41: 267–296. Available at http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14231- eng.pdf (accessed November 2019).

Van den Brakel, J.A., and S. Krieg. 2016. "Small area estimation with state-space common factor models for rotating panels." *Journal of the Royal Statistical Society* Series A. 179: 763–791. DOI: https://doi.org/10.1111/rssa.12158.

Vosen, M., and T. Schmidt. 2011. "Forecasting private consumption: survey-based indicators vs. Google trends." *Journal of Forecasting* 30: 565–578. DOI: https://doi.org/10.1002/for.1213.

Wallgren, A., and B. Wallgren. 2007. *Register-based statistics; Administrative data for statistical purposes*. John Wiley & Sons, West Sussex.

Ybarra, L.M.R., and S.L. Lohr. 2008. "Small area estimation when auxiliary information is measured with error." *Biometrika* 95: 919–931. DOI: https://doi.org/10.1093/biomet/asn048.

You, Y., J. Rao, and J. Gambino. 2003. "Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach." *Survey Methodology* 29: 25–32. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2003001/article/6602-eng.pdf (accessed November 2019).

Received March 2020
Revised September 2020
Accepted April 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 1047–1058, http://dx.doi.org/10.2478/JOS-2021-0044

The Robin Hood Index Adjusted for Negatives and Equivalised Incomes

Marion van den Brakel[1] *and Reinder Lok*[1]

Indisputable figures on income and wealth inequality are indispensable for politics, society and science. Although the Gini coefficient is the most common measure of inequality, the straightforward concept of the Robin Hood index (namely, the income share that has to be transferred from the rich to the poor to make everyone equally well off) makes it a more attractive measure for the general public. In a distribution with many negative values – particularly wealth distributions – the Robin Hood index can take on values larger than 1, indicating an intuitively impossible income transfer of more than 100%. This article proposes a method to normalise the Robin Hood index. In contrast to the original index, the normalised Robin Hood index always takes on values between 0 and 1 and ends up as the original index in a distribution without negatives. As inequality measures are commonly applied to equivalised income, we also introduce a method for adequately transferring equivalised incomes from the rich to the poor within the framework of the (normalised) Robin Hood index. An empirical application shows the effect of normalisation for the Robin Hood index, and compares it to the normalisation of the Gini coefficient from previous research.

Key words: Negative wealth; Pietra or Schutz index; normalisation; income inequality; Gini coefficient.

1. Introduction

For decades, inequality in income and wealth has been a continuous source of debate in politics, society and science. Studies on income and wealth inequality show a strong correlation with social and socio-economic changes. The OECD (2014, 2015) links globalisation and an increasingly flexible labour market to growing inequalities and demonstrates that in some Western countries, increasing inequality has had an inhibitory effect on economic growth. Piketty (2013) stated that growing economic inequality is accompanied by rising mistrust of citizens in fellow citizens and in politics, thus undermining the institutional structures of the society. The potential impact of economic inequality on societal changes, and vice versa, calls for adequate ways to describe the phenomenon.

Various criteria for measuring income (wealth) inequality have been developed over time. However, the complex structure of most of these criteria is a barrier to public

[1] Statistics Netherlands, Department of Statistics on Labour, Income and Living conditions CBS-weg 11 6412EX Heerlen, the Netherlands. Emails: mhfs@cbs.nl, and rlok@cbs.nl.

Acknowledgments: The authors are grateful to the unknown referees, the associate editors, Ferdy Otten (Statistics Netherlands) and Jan van den Brakel (Statistics Netherlands and Maastricht University) for reading and commenting on earlier drafts of the article. The views in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

understanding of financial prosperity inequality. Especially national statistical institutes, primarily responsible for providing inequality figures, are challenged with presenting indisputable figures in an appealing manner. The most commonly used measure of inequality is the Gini coefficient, developed in 1912 by the Italian statistician Corrado Gini. The coefficient owes its popularity in particular to its insightful graphic interpretation by the Lorenz curve. Less common, but conceptually much more insightful for the general public is the Robin Hood index, introduced by Gaetano Pietra in 1915 (see Pietra 2014 for an English translation) and also known as the Schutz index (Schutz 1951). The index expresses the share of the total income (or wealth) that has to be transferred from the rich to the poor half, in order to achieve an equal income for each household.

The Gini coefficient and the Robin Hood index normally take on values between 0 and 1, where 0 stands for perfect equality (everyone has the same income/wealth) and 1 for complete inequality (one household possesses everything). However, in a distribution with negatives, the Gini coefficient may take on values larger than 1, as pointed out by Chen et al. (1982). It is easy to see that the Robin Hood index is also sometimes faced with a distorded and intuitively impossible transfer of more than 100% of the total income. For instance: to achieve equality, Robin Hood has to transfer EUR 6,000 from a household with an income of EUR 8,000 to its neighbors who suffered losses of EUR 4,000; in other words, 1.5 times the total income (EUR 4,000) has to be shifted.

Elaborating on the work of Chen et al. (1982), Raffinetti et al. (2015) published a method to accurately incorporate negative incomes into the calculation of the Gini coefficient. This article discusses normalisation of the Robin Hood index, using the technique of Raffinetti et al. (2015) to proof the accurateness of the new index. In contrast to the original index, the normalised Robin Hood index always lies between 0 and 1 and ends up as the original index in a distribution without negative values. Distortion is no longer an issue and an outcome range regardless of the income (or wealth) distribution enables a proper inequality comparison between two or more populations, even if negative values occur. Furthermore, an upper and lower limit give meaning to the level of inequality of a distribution. As inequality measures are commonly applied to income and for comparability reasons equivalisation of income is customary international practice (United Nations 2011), we also introduce a method for adequately transferring equivalised incomes from the rich to the poor within the framework of the (normalised) Robin Hood index. An empirical application based on the Income and Wealth Statistics of Statistics Netherlands shows the effect of normalisation for the Robin Hood index as well as the Gini coefficient.

2. Normalising the Robin Hood Index

The simple concept of the share of income (or wealth) that has to be transferred from the rich half to the poor in order to achieve equality is captured in the formula of the classical Pietra or Robin Hood index R (see e.g., Pietra 2014; Ricci 1916):

$$R = \frac{\sum_{i \in D} (x_i - \mu)}{\sum_{i=i}^{N} x_i} = \frac{\sum_{i=1}^{N} |x_i - \mu|}{2 \sum_{i=1}^{N} x_i} \tag{1}$$

with N the number of units (households or persons) in the population, x_i the income of unit i, μ the mean income and D the subpopulation of units having an income larger than the mean. In words, R is the ratio of the total of all absolute unit distances to the mean income and twice the total income T. In distributions with only nonnegative values the outcome of R is always between 0 and 1, but as shown from the example in the introductory section this is not necessarily the case if negative values occur.

2.1. *Normalising (Positive Total)*

In a distribution $X = (x_1, \ldots, x_N)$ with both positive and negative values and T supposed to be positive, an upper bound of the Robin Hood index is T^+/T, where T^+ is the total amount of all positive incomes (*proposition A*).

Proof of proposition A: using that for any real values a and b the inequality $|a - b| \leq |a| + |b|$ applies and since $\mu > 0$, starting from the last term of Equation (1) it is easy to see that

$$R \leq \frac{\sum_{i=1}^{N}(|x_i| + |\mu|)}{2T} = \frac{\sum_{x_i<0}|x_i| + \sum_{x_i\geq 0}|x_i| + N\mu}{2T} = \frac{T^- + T^+ + T}{2T} = \frac{T^+}{T} \quad (2)$$

where T^- is equal to the absolute value of the total amount of all negative incomes.

It now is obvious to define the normalised Robin Hood index $R*$ by dividing R by T^+/T:

$$R^* = \frac{\sum_{i\in D}(x_i - \mu)}{T^+} \quad (3)$$

Both the denominator and nominator of $R*$ are not negative by construction, so the lower bound of $R*$ is zero. The upper bound is equal to 1, which directly follows from Equation (2). The thus normalised Robin Hood index $R*$ can be interpreted as the share of positive income that has to be transferred to achieve perfect equality, which intuitively is a logical way for Robin Hood to act. Furthermore, $R* = R$ in case of only nonnegative values.

2.2. *Visual Interpretation*

The normalisation of the Robin Hood index can be interpreted graphically by the Lorenz curve. To illustrate this, first of all note that the classical Robin Hood index (1) of X is identical to the longest vertical distance between the Lorenz curve, which is the cumulative portion of the total income held below a certain income percentile, and the 45-degree line representing perfect equality (see also Pietra (2014), who proved this for distributions with only nonnegative values). Defining this longest distance by R_X, this proposition (*B*) can be formulated as: $R = R_X$, in a distibution X with positive and negative values and $T > 0$.

Proof of propostion B: Let f be the discrete Lorenz curve of X (in which the units are ordered by income), thus $f(k) = \frac{\sum_{i=1}^{k} x_i}{T}$. Note that as long as $x_k < 0$ the Lorenz curve just decreases and gets further away from the equality line (slope < 0). As soon as $x_k \geq 0$ the decreasing stops and once $x_k > 0$ the curve increases (slope ≥ 0). Let k be first unit of X

for which the slope of the Lorenz curve f is larger than 1. For this slope applies:

$$\frac{\Delta f}{\Delta i} = \frac{f(k) - f(k-1)}{1/N} = N\left(\frac{\sum_{i=1}^{k} x_i}{T} - \frac{\sum_{i=1}^{k-1} x_i}{T}\right) = \frac{Nx_k}{T} > 1, \tag{4}$$

which implies that k is the first unit for which the income x_k is larger than the mean μ. From unit k on, the vertical distance between the Lorenz curve and the equality line (with constant slope 1) will only become smaller (as long as the slope of the Lorenz curve is smaller than 1, the distance between the curve and the equality line grows). The maximum vertical distance is therefore found at $m = k\text{-}1$ and is equal to

$$\frac{m}{N} - \frac{\sum_{i=1}^{\mathrm{m}} x_i}{T} = \frac{mT}{NT} - \frac{\sum_{i=1}^{\mathrm{m}} x_i}{T} = \frac{m\mu}{T} - \frac{\sum_{i=1}^{m} x_i}{T} = \frac{\sum_{i \in D^*} (\mu - x_i)}{T} \tag{5}$$

with $D*$ the subpopulation of units having an income smaller than or equal to the mean. Note that Equation (5) is equal to the classical Robin Hood index (1) and for the proof it only matters that T is positive.

Distribution X has positive as well as negative values and the highest value of the Lorenz curve is equal to 1 and the lowest to $\frac{-T^-}{T}$. The difference between the highest and lowest value (d) is obviously an upper bound for R_X. As d equals $1 + \frac{T^-}{T} = \frac{T^+}{T}$, which is just the ratio derived in Equation (2), the nomalised Robin Hood index can be interpreted as the ratio between distance R_X and distance d.

Example To further clarify the visual interpretation consider the distribution with values (-8;-3;3;8;10) and its Lorenz curve in Figure 1. For this distribution the distance R_X is

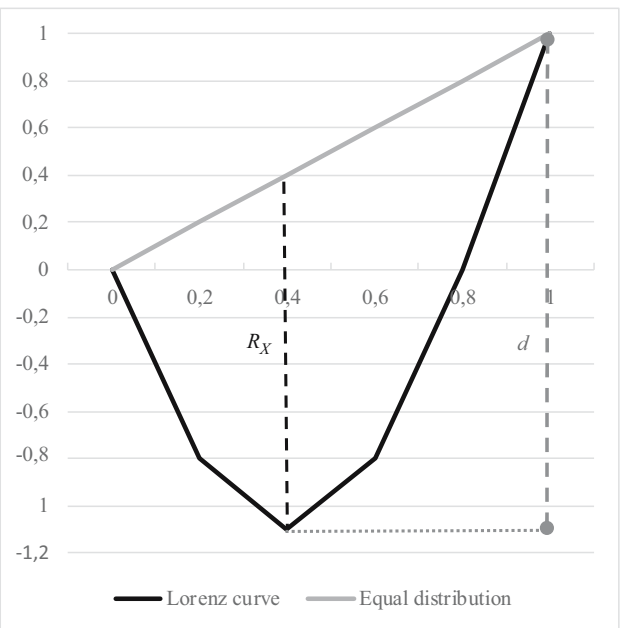

Fig. 1. Distribution (-8,-3, 3, 8, 10).

equal to 1.5. The distance between the highest and lowest value of the Lorenz curve equals 2.1. Dividing 1.5 by 2.1 gives a value of 0.71. Exactly the same value can be achieved using Equation (3).

2.3. No Over-Normalisation

The normalisation of the Robin Hood index could be too rough in the way that ratio T^+/T might be too large. To prove that no over-normalisation is done by dividing Equation (1) by this ratio (*proposition C*), consider the corresponding distribution with maximum inequality $Z = \left(-T^-, 0, \ldots, 0, T^+\right)$ used by Raffinetti et al. (2015) to normalise the Gini coefficient.

Proof of proposition C: Not only in terms of the Gini coefficient, but also in terms of the classical Robin Hood index, Z matches to maximum inequality. After all, the longest vertical distance from the Lorenz curve of Z to the equality line is previously proven to be identical to the classical Robin Hood index of Z and from Equation (1) equals $\frac{T^+-\mu}{T} = \frac{T^+}{T} - \frac{1}{N}$. As $N \to \infty$ this approaches $\frac{T^+}{T}$, which is already shown to be larger than the classical Robin Hood index of X.

For distribution Z the normalised Robin Hood index is equal to $R_Z^* = \frac{T^+ - \frac{\left(T^+ - T^-\right)}{N}}{T^+} = 1 - \frac{1}{N}\frac{T}{T^+}$. Obviously as $N \to \infty$, R_Z^* will approximate 1, which means that over-normalisation is not the case.

2.4. Normalising (Zero or Negative Total)

Up to now the total income T of distribution $X = \left(x_1, \ldots, x_N\right)$ was assumed to be positive. For the normalised Robin Hood index, however, it is no problem if T is zero, that is $T^- = T^+ (> 0)$. In this (rare) case R^* equals 1. In the special case where all values are zero, R and $R*$ are undefined, but since this refers to an equal distribution, the (normalised) Robin Hood index can be defined as 0. When T is negative ($T^- > T^+$) the normalised Robin Hood as formulated in Equation (3) cannot be applied, because to deduce it, the total was assumed to be positive. A solution for this problem can be found in the mirrored distribution of X. For this, first notice that the maximum vertical distance of the Lorenz curve of X to the equality line is equal to that of the Lorenz curve of $Y = -X$ (*proposition D*).

Proof of propostion D: For every $l \leq N$ the value of the Lorenz curve f of X is at least $\frac{l}{N}$, meaning that the Lorenz curve f lies above or on the equality line. To see this, let $k < N$ be the last unit in the (ranked) distribution X for which $x_k \leq \mu$. Then, since T and μ are equal to the negation of the total T_Y and the mean μ_Y of distribution Y respectively, for every $l \leq k$

$$y_l \geq \mu_Y \text{ and } f(l) = \frac{\sum_{i=1}^{l} x_i}{T} = \frac{\sum_{i=1}^{l}\left(-y_i\right)}{-T_Y} \geq \frac{l\mu_Y}{N\mu_Y} = \frac{l}{N}$$

and for every $l > k$

$$f(l) - \frac{l}{N} = \frac{\sum_{i=1}^{l} x_i}{T} - \frac{l}{N} > \frac{\sum_{i=1}^{l} \mu}{T} - \frac{l}{N} = 0.$$

Opposite to the situation in which the Lorenz curve lies beneath the equality line ($T > 0$), the maximum vertical distance of f to the equality line is found at the point from where on

the slope of f only takes on values smaller than 1. For the slope at this point, say at unit k, analogous to Equation (4) applies $\frac{\Delta f}{\Delta i} = \frac{Nx_k}{T} < 1$ which comes down to $y_k < \mu_Y$. This exactly corresponds to the unit at which the vertical distance of the Lorenz curve of Y to the equality line is at the largest, as seen before in Equation (5).

This means that the normalised Robin Hood of X (with negative total) can logically be defined as the normalisation of that of $Y = -X$. Normalising the Robin Hood index of Y means dividing it by $\frac{T_Y^+}{T_Y}$ and since $T_Y^+ = \sum_{i=1}^{N} max(0, y_i) = \sum_{i=1}^{N} max(0, -x_i) = \left|\sum_{i=1}^{N} min(0, x_i)\right| = T^-$ the normalised index for distribution X is where $T^- > T^+$ equals

$$R_Y^* = \frac{\sum_{i \in D_Y} (y_i - \mu_y)}{T_Y^+} = \frac{\sum_{i \in D_Y} (-x_i + \mu)}{T^-} = \frac{\sum_{i \in D^*} (\mu - x_i)}{T^-} = \frac{\sum_{i \in D} (x_i - \mu)}{T^-} \tag{6}$$

Combining Equations (3) and (6) the definition of the normalised Robin Hood index for a non-zero distribution X is:

$$\frac{\sum_{i \in D} (x_i - \mu)}{max(T^+, T^-)} \tag{7}$$

with mean μ and T^+, T^- and D as aforementioned.

Example In the distribution $X =$ (-8;-3;0;3) with positive as well as negative values the sum of the absolute negatives exceeds the sum of the positives. Applying Equation (3) would return a normalised Robin Hood index of 7/3. The longest vertical distance R_X from the Lorenz curve of X to the equality line is the same as that of the curve mirrored in the equality line (see Figure 2). This mirrored Lorenz curve belongs to the distribution $Y =$

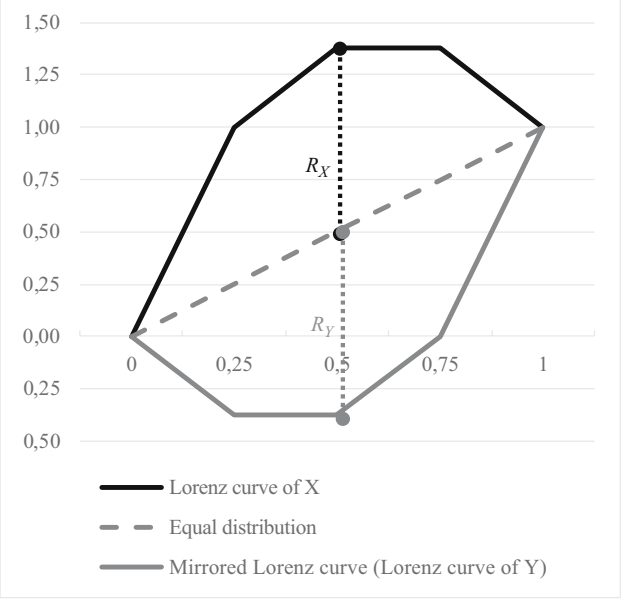

Fig. 2. Distribution (-8, -3, 0, 3).

(-3;0;3;8), that is, to the distribution $-X$ in which the total of the positive values exceeds the total of the absolute negative incomes. The normalised Robin Hood R_Y^* derived from Equation (3) is 7/11, which is exactly the same value achieved from Equation (7).

In the special case where all incomes of a distribution $X = (x_1, \ldots, x_N)$ are negative, the income inequality can be calculated by the mirrored distribution $Y = (y_1, \ldots, y_N) = (|x_1|, \ldots, |x_N|)$ and using the classical Robin Hood index (1).

2.5. *Properties of the Normalised Robin Hood Index*

For a distribution $X = (x_1, \ldots, x_N)$ with positive as well as negative values (i.e., $T^+ > 0$ and $T^- > 0$) a *scale 'invariance' property* can be formulated for the normalised Robin Hood index:

If $(T^+ \neq T^-)$ then for every constant $\alpha \neq 0$ the normalised Robin Hood index of $Y = \alpha X$ is equal to $\frac{\sum_{i \in D}(x_i - \mu)}{max(T^+, T^-)}$.

Other features the normalised Robin Hood index (just like the classical index) meets are for instance *symmetry* (swapping the income of two households leaves the index unchanged) and *population size independency* (merging two or more identical distributions does not influence the outcome of the index). The *Pigou-Dalton criterion* only holds for incomes shifted from the rich (individuals having an income above the mean) to the poor (income below the mean), and vice versa. Incomes transferred within the rich (poor) are not signaled by the (normalised) Robin Hood index. The normalised Robin Hood index satisfies *boundedness*, such that for every distribution the same upper and lower limit apply. This makes comparison of subpopulations possible and gives meaning to the level of inequailty. The traditional Robin Hood index meets the property of boundedness in distributions with solely nonnegatives. *Decomposition* of an inequality measure is a desirable but not necessary feature. Habib (2012) developed a method to decompose the traditional Robin Hood index. This method can easily be applied to the normalised Robin Hood index as well.

3. Transferring Equivalised Incomes

Multi-person households mostly have more income than a single person. However, it matters a lot how many people within a household have to live on a certain income. It therefore makes no sense to determine income inequality without taking into account the size of the household. To make households of different sizes and composition comparable, incomes are equivalised.

Income is equivalised by dividing the household income by a factor that expresses the economies of scale when running a joint household. Single-person households have been chosen as the standard. The factor is set to 1 for these households. For multi-person households the factor depends on the equivalence scale that is chosen. Various alternative equivalence scales have been developed worldwide. International scales that are frequently used within the OECD countries are the modified equivalence scale and the square root scale (OECD 2013). Using the square root scale, a single person with a disposable income of EUR 10,000 and a couple with a disposable income of EUR 14,100

are at the same level of prosperity: after equivalisation, the purchasing power for both households is EUR 10,000.

With equivalised incomes, Robin Hood's job is a bit more complex. The transfer of income from rich to poor should be done in such a way that afterwards every household has the same equivalised income and the total unequivalised income of the population remains the same. For clarification, consider the following situation.

Example A couple A has a disposable income of EUR 1,500, their single-person neighbor B has no income at all. Assuming for the sake of simplicity an equivalence factor of 1.5 for couples, the average equivalised income is EUR 500. If Robin Hood were to transfer equivalised incomes, couple A would have to hand over EUR 500 to individual B. This is half the total equivalised income, which implies the Robin Hood index $R = 1/2$. Counting back this means that couple A has EUR 750 to spend and individual B EUR 500. Has Robin Hood put EUR 250 in his own pocket?

By shifting not with 'fictional' (equivalised) but with genuine money, Robin Hood can prevent defamation. Because the couple A shares a household, they do not count for 2 but for 1.5. The total disposable income of EUR 1,500 must therefore be distributed in such a way that couple A has 1.5 times as much as individual B. Robin Hood calculates that A has to hand over EUR 600 to B, after which the couple has EUR 900 to spend. This means that 2/5 of the total income has been transferred in order to get equal equivalised incomes.

In general, this means that the total of non-equivalised incomes must be evenly distributed over a population of size equal to the sum of the equivalence factors. Based on the traditional Robin Hood index, the proper transfer of equivalised incomes $X = \left(x_1, \ldots, x_N\right)$ is expressed by:

$$R_{equi} = \frac{\sum_{i \in B} (x_i - \vartheta) e_i}{\sum_{i=1}^{N} y_i} \tag{8}$$

where y_i is the unequivalised income of household i, e_i the equivalence factor, ϑ the mean of all unequivalised incomes of the population of size $M = \sum_{i=1}^{N} e_i$:

$$\vartheta = \frac{\sum_{i}^{N} y_i}{M},$$

and B the subpopulation of households with $y_i > \vartheta$. It is straightforward to derive an expression for R_{equi} in case of negative (and positive) incomes and the total $T_Y = \sum_{i=1}^{N} y_i > 0$:

$$R^{*}_{equi} = \frac{\sum_{i \in B} (x_i - \vartheta) e_i}{T_Y^{+}} \tag{9}$$

whit T_Y^{+} the sum of all positive unequivalised incomes. If the total is negative, Equation (7) can be applied.

4. An Application

In this section, the impact and relevance of normalising the Robin Hood index is shown by giving some examples of the Income and Wealth Statistics (IWS) of the Netherlands. Furthermore, a comparison with the Gini coefficient is made here. Income inequality is hardy affected by the normalisation of the Robin Hood index or the Gini coefficient. Over the years, the normalised figures for equivalised disposable incomes (see CBS Statline 2020a, where normalisation in accordance with Equation (8) and Raffinetti et al. (2015) respectively is applied) were slightly smaller than the not-normalised income inequalities. For instance, the difference in 2017 was less than 0.2%. Normalisation has more impact on wealth inequality. This is because the wealth of almost 20% of the households in 2017 is negative: their liabilities (mortgage debts and consumer credit) transcend their assets (mainly bank balances, shares, real estate, and business capital). A negative income is much less common (0.5% in 2017). For both inequality measures, the impact of normalisation grows until 2014, after which it decreases (see Figure 3). This had to do with the economic climate in this period. As a result of the economic crisis that started at the end of 2008, more and more households were faced with negative wealth, especially due to falling house prices. From 2014, the Dutch economy recovered, house prices rose and the number of households with negative wealth decreased again. Note that compared to the Robin Hood index, normalising wealth inequality with the Gini coefficient has more effect. Normalisation reduced the wealth inequality of the Robin Hood index by 4% in 2017. Using the Gini coefficient, this figure was twice as large (7.9%). This is simply because the factor used to normalise the Gini coefficient, that is, $\left(T^{+}+T^{-}\right)/T$ in large populations (see Raffinetti et al. 2015), is larger than that of the Robin Hood index derived in Equation (2).

For certain groups of households in which relatively many negative wealth values occur, like households having a young main earner or households in which the main earner

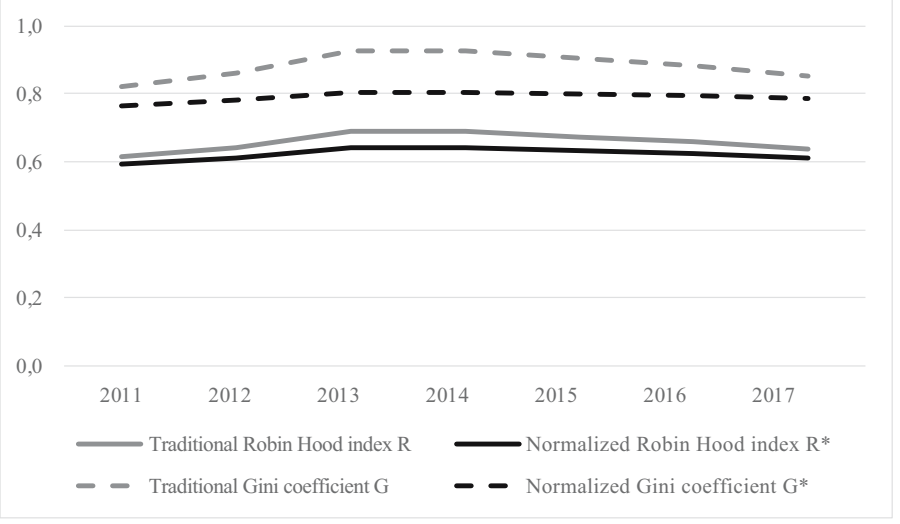

Fig. 3. Wealth inequality of households in the Netherlands.
1) G∗ in accordance with Raffinetti et al. 2015.
2) Descriptive statistics can be found on CBS StatLine (2020b).

has a non-Western migration background, the traditional Robin Hood index and the Gini coefficient take on values larger than 1 (see Table 1). Normalisation provides wealth inequality values smaller than 1. Note that normalisation enables to compare inequalities of different populations. Young households (main earner younger than 25 years) seem to have a much higher ineqality than the group with a non-Western migration background, according to standard inequality scales. However, after normalisation they appear to be quite comparable. The many negative wealth cases (mostly student loans in the young group) did indeed bias the inequality scale to a non-interpretable level.

Table 1. Wealth Households by Characteristics of Main Earner 2017.

	Households with negative wealth (%)	R	R*	G	G*
Age					
younger than 25	42	1,353	0,810	2,039	0,870
25 to 45	32	0,855	0,710	1,188	0,844
45 to 65	16	0,593	0,572	0,811	0,757
65 and over	4	0,516	0,513	0,710	0,700
Migration background					
Dutch	17	0,603	0,581	0,817	0,761
Western	20	0,708	0,675	0,920	0,838
non-Western	32	1,032	0,808	1,403	0,902

1) Traditional (G) and normalised Gini coefficient G* in accordance with Raffinetti et al. 2015.

2) Descriptive statistics can be found on CBS StatLine (2020b).

5. Conclusion and Discussion

This article proposed a method to normalise the Robin Hood index in order to deal with negative values in an income or wealth distribution. The normalised index expresses the share of the total positive amount of income or wealth (instead of the total amount, as in the traditional Robin Hood index) that has to be transferred from the rich half to the poor half in order to achieve perfect equality. The method provides an expression for normalisation, even in extreme distributions with zero or negative total. A proper (normalised) way to shift equivalised incomes from the rich to the poor is also incorporated. An application to the prosperity distributions of households in the Netherlands shows that normalisation is especially necessary for wealth inequality, since households with negative wealth are much more common than households with negative income. The development over time of income and wealth inequality after normalisation appears to be the same as before.

Although the Gini coefficient is widely used, its concept is more abstract than that of the Robin Hood index. The simple perception of the Robin Hood index makes it an accessible and understandable measure of income (or wealth) inequality. Moreover, in contrast with similarly easy measures that compare the top and bottom of a distribution (like the 80/20 ratio) or the share of the rich in the total wealth (see e.g., Piketty 2014), the Robin Hood index takes into account each individual value of the population. Another disadvantage of measures such as the 80/20 ratio is their inability to cope with negative values. This

underlines the significance of the normalised Robin Hood as a measure for inequality to serve a broad public, certainly since it meets several desirable features including symmetry, population size independency and decomposability. As the Pigou-Dalton criterion is only partly met, additional analyses of inequality using the (normalised) Gini coefficient are recommended.

In accordance with international standards, the disposable income does not include quaternary income components. Specifically, this means that both the social benefits received in kind (such as free education and medical care) and the benefits of collective goods (infrastructure and the like) are not taken into account in the disposable income. Therefore caution is required when comparing income inequality in the Netherlands with that in other countries. If the quaternary benefits are at a lower level elsewhere, the observed differences can quickly lead to distorted insights.

The ways in which wealth is measured internationally may differ even more (Balestra and Tonki 2018). In the Netherlands, for instance, when determining the mortgage debt, any accumulated assets with regard to savings and investment mortgages have not been included because the underlying data are lacking. Whether or not pension entitlements and other social security entitlements are counted as assets may also differ among countries, as well as equivalisation of wealth. The figures on income and wealth inequality are therefore primarily intended to monitor developments within a country, especially if normalisation is applied.

6. References

Balestra, C., and R. Tonki. 2018. *Inequalities in household wealth across OECD countries: Evidence from the OECD Wealth Distribution Database.* OECD Statistics Working Paper Series. Available at: www.oecd.org. (accessed September 2019).

CBS StatLine. 2020a. Inequlity of income and wealth (in Dutch). Available at: http://opendata.cbs.nl/statline/#/CBS/nl/dataset/84476NED/table?dl = 49EF5 (accessed January 2021).

CBS StatLine. 2020b. Welfare of households; key figures. Available at: http://opendata.cbs.nl/statline/#/CBS/en/dataset/83739ENG/table?dl = 49EF4 (accessed January 2021).

Chen, C.-N., T.-W. Tsaur, and T.-S. Rhai. 1982. "The Gini coefficient and negative income." *Oxford Economic Papers* 34 (3): 473–478. http://dx.doi.org/10.1093/oxfordjournals.oep.a041563.

Habib, E.A. 2012. "On the decomposition of the Schutz coefficient; an exact approach with an application." *Electronic Journal of Applied Statistical Analysis* 5(2) : 187–198. DOI: http://dx.doi.org/10.1285/i20705948v5n2p187.

OECD. 2013. *OECD Framework for statistics on the distribution of household income, consumption and wealth.* Available at: www.oecd.org (accessed September 2019).

OECD. 2014. *Focus on Inequality and Growth*. Paris: OECD Publishing.

OECD. 2015. *In It Together: Why Less Inequality Benefits All*. Paris: OECD Publishing.

Pietra, G. 1915. "Delle relazioni tra indici di variabilità." *Atti del reale Istituto Veneto di Scienze, Lettere ed Arti* 74(2).

Pietra, G. 2014. "On the relations between variability indices (Note I)." *METRON 72: 5–16*. DOI: https://doi. org/10.1007/s40300-014-0034-3.

Piketty, T. 2014. *Capital in the Twenty-First Century*. Harvard University Press.

Raffinetti, E., E. Siletti, and A. Vernizzi. 2015. "On the Gini coefficient normalization when attributes with negative values are considered." *Stat Methods Appl* 24: 507–512. DOI: http://dx.doi.org/10.1007/s10260-014-0293-4.

Ricci, U. 1916. "L'indice di variabilità e la curva dei redditi." *Giornale degli Economisti e Rivista di Statistica* LIII: 177–228. serie terza, anno XXVII. Available at: https://www.jstor.org/stable/23225478 (accessed October 2021).

Schutz, R. 1951. "On the measurement of income inequality." *The American Economic Review* 41: 107–122. Available at: On the Measurement of Income Inequality (jstor.org) (accessed April 2021).

United Nations. 2011. Canberra Group, Handbook on Household Income Statistics (second edition). New York and Geneva. Available at: Canbera_Handbook_2011_-WEB.pdf (unece.org) (accessed April 2021).

Received October 2019
Revised May 2020
Accepted April 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 1059–1078, http://dx.doi.org/10.2478/JOS-2021-0045

Estimation of Domain Means from Business Surveys in the Presence of Stratum Jumpers and Nonresponse

Mengxuan Xu[1], *Victoria Landsman*[1], *and Barry I. Graubard*[2]

Misclassified frame records (also called stratum jumpers) and low response rates are characteristic for business surveys. In the context of estimation of the domain parameters, jumpers may contribute to extreme variation in sample weights and skewed sampling distributions of the estimators, especially for domains with a small number of observations. There is limited literature about the extent to which these problems may affect the performance of the ratio estimators with nonresponse-adjusted weights. To address this gap, we designed a simulation study to explore the properties of the Horvitz-Thompson type ratio estimators, with and without smoothing of the weights, under different scenarios. The ratio estimator with propensity-adjusted weights showed satisfactory performance in all scenarios with a high response rate. For scenarios with a low response rate, the performance of this estimator improved with an increase in the proportion of jumpers in the domain. The smoothed estimators that we studied performed well in scenarios with non-informative weights, but can become markedly biased when the weights are informative, irrespective of response rate. We also studied the performance of the 'doubled half' bootstrap method for variance estimation. We illustrated an application of the methods in a real business survey.

Key words: Doubled half bootstrap; propensity-adjusted weights; weight smoothing.

1. Introduction

Firm size plays an instrumental role in different stages of business surveys. The information about a firm's size (usually in the form of employment or annual sales), geographic area and business sector obtained from administrative data are often used to define sampling strata at the design stage of the survey. Additionally, firm size is believed to be related to the probability to participate in the survey. Since business surveys may be prone to low response rates (Cook et al. 2009), the information about firm size is useful for nonresponse adjustments of the sample weights. Firm size was also suggested to be an important proxy variable to a firm's occupational health and safety performance (Nordlöf et al. 2015). For this reason, there is a lot of interest to evaluate and compare various performance metrics between firms of different sizes. Frequently, the target parameter is expressed as the domain mean where the domains are defined by the levels of firm size,

[1] Institute for Work and Health, 400 University Avenue, Suite 1800, Toronto, Ontario M5G 1S5, Canada. Emails: mengxuan.xu@mail.utoronto.ca and vlandsman@iwh.on.ca
[2] National Cancer Institute, Division of Cancer Epidemiology and Genetics, Bethesda, MD 20852, U.S.A. Email: graubard@mail.nih.gov
Acknowledgments: The authors thank Drs. Ben Amick III, Sheilah Hogg-Johnson, and Ms. Colette Severin for providing the OLIP data and the inspiration for the jumpers problem, which motivated this study. The authors also express their gratitude to the Associate Editor and three anonymous referees whose valuable comments help to considerably improve the quality of this manuscript.

usually collected at the time of survey. The information about firm size obtained from administrative data that is typically used for defining the sampling strata and nonresponse adjustments, may contradict the information obtained from survey participants. Such situations may occur, for instance, when a `small` firm expanded and became a `large` firm in the time period between the survey design and data collection. Firms for which the two sources of information do not agree have been referred to in the literature as 'misclassified frame records' or 'stratum jumpers' (MacNeil and Pursey 2002, Beaumont and Rivest 2009). This article aims to address challenges in estimating the domain means of organizational performance measures from business surveys with misclassified frame records and low response rates.

Firms are usually sampled with sampling fractions proportional to their size. Therefore, jumpers that were considered as `small` firms at the design stage would typically have very large sample weights. In contrast, the correctly classified `large` firms are frequently sampled with certainty or with probability close to 1, and, therefore, have weights close to or equal to 1. When the domains of interest are defined by the levels of the firm size variable collected at the time of the survey, the presence of just a few jumpers in the domain of `large` firms may cause a dramatic increase in the coefficient of variation (CV) of the sample weights in the domain. Moreover, if the probability of response is proportional to firm size, nonresponse adjustments applied to the sample weights may further increase the CV of the weights in the domain.

Highly variable sample weights are known to lead to inefficient weighted estimators for means and totals (Korn and Graubard 1999, 172–173). Also, design-based variance estimators have been shown to underestimate the variance of the weighted estimators with highly variable weights in case-control designs used in epidemiological studies (Li et al. 2011; Landsman and Graubard 2013). In addition, one might expect a skewed sampling distribution of the ratio weighted estimators for domains with a small number of observations (Lee 1995). All of these factors combined are expected to slow the rate of convergence of consistent design-based variance estimators and negatively affect the coverage probabilities of the confidence intervals (CIs) around the parameters of interest.

The winsorization method is often used to treat influential values of outliers in survey data (Chen et al. 2017). This method requires finding a threshold, above which values of the variable of interest or the sampling weights are 'trimmed' down to the threshold to reduce their influence. Usually, the value of the threshold is selected to minimize the design mean square error of the estimator with the winsorized weights, although other methods have been proposed more recently (Favre-Martinoz et al. 2015). The winsorization method is less appealing in business surveys with multiple variables of interest, because different cutoff values imply different winsorized weights for each variable. For this reason, we did not pursue winsorization in this study. Weight smoothing is another approach that has been proposed to address the problem of highly variable weights (e.g., Pfeffermann and Sverchkov 1999; Beaumont 2008). Beaumont and Rivest (2009) devised two weight smoothing procedures to address the problem of the influential weights of the jumpers for estimating a population total from a stratified random sample with known (fixed) selection probabilities. One of the two procedures is a special case of the model-based approach proposed in the work of Beaumont (2008), and the second procedure can be viewed as a mixture of winsorization and smoothing approaches. The authors further showed, using a

real data example, that the Horvitz-Thompson (H-T) estimator for the total with smoothed weights may be superior to the standard H-T estimator with the original (base) weights.

To the best of our knowledge, the performance of ratio estimators with nonresponse-adjusted weights, with and without smoothing of the weights, has not been studied previously in the context of business surveys with misclassified frame records. Design-based variance estimation in this context can also be challenging, especially when using smoothed weights. Beaumont and Rivest (2009) used a Rao-Wu rescaled bootstrap estimator (Rao and Wu 1988; Rao et al. 1992) with smoothed weights, but did not consider nonresponse.

In this article we describe the exploratory work designed to address these gaps in the analysis of business surveys. The rest of the article is organized as follows. In Section 2, we formalize the problem and define a target parameter for the case of full response. For extension beyond this case, nonresponse adjustments are briefly outlined in Section 3. In Section 4, we describe the two weight smoothing methods proposed by Beaumont and Rivest (2009) that we adapted to the estimation of domain means. Section 5 describes the variance estimation using the 'doubled half' bootstrap method (Antal and Tillé 2014). In Section 6, we illustrate the application of the methods to real data from an organizational performance survey. A detailed description of the simulation study and our main findings are summarized in Section 7. We conclude with a discussion about our findings in Section 8.

2. Problem Statement

Suppose that U is a finite population (of size N) of all firms in a given jurisdiction, and $Z_{\texttt{surv}}$ is a firm size variable with D categories defined at the time of the survey (hence, the subscript `surv`). Thereby, $Z_{\texttt{surv}}$ defines the partition of U into D domains. The domain means μ_d ($d = 1,\ldots, D$) of a study variable y are the target parameters. To estimate μ_d, a stratified random sample $\tilde{S}$ (of size $\tilde{n}$) with $H > 1$ strata is usually drawn from U. In business surveys, the strata are typically defined by the categories of firm size and other relevant variables (e.g., geographic region and business sector).

The values of the variable $Z_{\texttt{surv}}$ are usually unobserved for the firms in the finite population and will only be obtained for the responding firms. The proxy variable for the firm size, $Z_{\texttt{des}}$, is usually derived from the available information (e.g., administrative data) at the design stage (hence, the subscript `des`) and this is the firm size variable that is used to form the strata. Throughout this article, $Z_{\texttt{des}}$ has two values: `small` and `large`. As a result, multiple strata of `small` and `large` firms are created by all possible combinations of the two categories of $Z_{\texttt{des}}$ and the categories of additional variables used to defined the strata. For simplicity of presentation, in this section and in Section 4 we assume that $Z_{\texttt{surv}}$ also has two values, `small` and `large`, defined in the same way as the two levels of $Z_{\texttt{des}}$. In Sections 6 and 7, we consider the case where $Z_{\texttt{surv}}$ has five different values.

The two categories of $Z_{\texttt{surv}}$ define the partition of $\tilde{S}$ into two mutually exclusive and collectively exhaustive domains: $\tilde{S}_{\texttt{small}}$ – the domain of `small` firms and $\tilde{S}_{\texttt{large}}$ – the domain of `large` firms. Assuming full response, $\mu_d (d \in \{\texttt{small}, \texttt{large}\})$ can be estimated from the corresponding sample domain $\tilde{S}_d$ using a standard Horvitz-Thompson type ratio estimator

$$\hat{\mu}_d = \sum_{i \in \tilde{S}_d} b_i y_i / \sum_{i \in \tilde{S}_d} b_i \tag{1}$$

where b_i is a base sampling weight of firm i determined by the stratified design.

Discrepancies between the values of $Z_{\texttt{des}}$ and $Z_{\texttt{surv}}$ might occur, especially if the information on the frame, used to define $Z_{\texttt{des}}$, was outdated (e.g., a large gap in time between the design and implementation of the survey). In this article, we assume that no firms changed their size from `large` to `small`, and focus our attention on the situation when firms changed their size from `small` to `large`, as this situation is more challenging for practitioners. Under this assumption, the sample domain of `large` firms can be written as $\tilde{S}_{\texttt{large}} = \tilde{S}_J \cup \tilde{S}_L$, where

$$\tilde{S}_J = \left\{ i \in \tilde{S} | Z_{\texttt{des},i} = \texttt{small}, Z_{\texttt{surv},i} = \texttt{large} \right\}$$

and

$$\tilde{S}_L = \left\{ i \in \tilde{S} | Z_{\texttt{des},i} = \texttt{large}, Z_{\texttt{surv},i} = \texttt{large} \right\}$$

Firms in $\tilde{S}_J$ have been referred to in the literature as 'misclassified frame records', or 'stratum jumpers'.

Probability proportional to size (PPS) sampling, often used in stratified designs, assigns small sampling fractions, or, equivalently, large b_i, to `small` firms, including the firms in $\tilde{S}_J$. Firms in $\tilde{S}_L$, on the other hand, are usually assigned sampling fractions with values close to or equal to 1, or, equivalently, values of b_i which are close to or equal to 1. As a result, the weights of the firms in $\tilde{S}_J$ can be considerably higher than the weights of the firms in $\tilde{S}_L$, causing a dramatic increase in the CV of the weights in $\tilde{S}_{\texttt{large}}$. This is problematic, as the domain mean estimator (1) with highly variable weights may lead to estimates with large variances if the correlation between the weights and the variables of interest is weak (Rao 1966).

3. Nonresponse Adjustments

In practice, survey data is available for only a subset of respondents S rather than for the entire selected sample $\tilde{S}$. In such cases, adjustments to the base weights are used to reduce the nonresponse bias. The main idea behind various adjustment procedures proposed in the literature is to estimate a firm's response probability, p_i, using all of the available data for the firms in $\tilde{S}$. The nonresponse-adjusted weights are defined as $w_i = b_i \hat{p}_i^{-1}$, where $\hat{p}_i$ is the estimated response probability of firm i, and may be further calibrated to match known population totals (Haziza and Lesage 2016). For simplicity, we do not consider calibration in this article and define the propensity-adjusted weights w_i as final sample weights. Then, the population domain mean μ_d can be estimated from a sample of respondents S_d using a ratio estimator with nonresponse-adjusted weights w_i

$$\hat{\mu}_d = \sum_{i \in S_d} w_i y_i / \sum_{i \in S_d} w_i \tag{2}$$

Different procedures can be applied to estimate p_i, depending on data availability and research goals. If the only data available for the firms in $\tilde{S}$ is the data that was used to form the sampling strata, the response probability can be estimated by $\hat{p}_{ih} = n_h / \tilde{n}_h$, where n_h is

the number of respondents and $\tilde{n}_h$ is the sample size of stratum h ($h = 1,\ldots, H$) in $\tilde{S}$. This method has been referred to as the 'cell adjustment' method (Little and Vartivarian 2003, Valliant et al. 2013) and it assumes that nonresponse is 'completely at random' within each stratum. We denote the 'cell-adjusted' weights by $w_{\texttt{str}}$ throughout the article.

If additional information is available for the firms in the selected sample $\tilde{S}$ (e.g., by linking $\tilde{S}$ to administrative data), the response probabilities can be estimated by fitting a response regression model using all the relevant covariates from the linked data. In this case, the response probability is approximated by $\hat{p}_i = p(\mathbf{x}_i, \hat{\beta})$, where $\mathbf{x}_i$ is the vector of observed covariates and $\hat{\beta}$ are the regression coefficients that were estimated from fitting a response regression model to $\tilde{S}$. The resulting adjusted weights are called the propensity-adjusted weights (Kim and Kim 2007). This method assumes that nonresponse is 'completely at random' given $\mathbf{x}$. We denote the propensity-adjusted weights by $w_{\texttt{pa}}$ throughout the article.

A combination of jumpers and low response rate in the data may create additional challenges for estimating domain means. On the one hand, nonresponse adjustments are expected to further increase the CV of the weights in the domain of `large` firms, especially if the response probability is proportional to the firm's size. On the other hand, low response rates imply small effective sample sizes of the domains, which can slow down the rate of convergence of the estimators.

4. Sample Weight Smoothing Techniques

In this section, we adapt the two sample weight smoothing procedures, proposed by Beaumont and Rivest (2009), to the estimation of domain means in the presence of jumpers.

For this purpose, we assume that each domain (defined by the levels of $Z_{\texttt{surv}}$) is further partitioned into T mutually exclusive and collectively exhaustive subdomains: $S_{\texttt{large}} = \cup_{t=1}^{T} S_{\texttt{large},t}$ and $S_{\texttt{small}} = \cup_{t=1}^{T} S_{\texttt{small},t}$, where $S_{\texttt{large}}$ and $S_{\texttt{small}}$ are the corresponding subsets of respondents of $\tilde{S}_{\texttt{large}}$ and $\tilde{S}_{\texttt{small}}$, respectively. These subdomains can be formed by the categories of additional stratifying variables (e.g., geographic region and business sector) often used to design the sample. Analogously, $S_{\texttt{large}} = S_J \cup S_L$, where S_J and S_L are the subsets of respondents of $\tilde{S}_J$ and $\tilde{S}_L$ (defined in Section 2), respectively. Both smoothing procedures are applied independently to each subdomain $S_{\texttt{large},t} = S_{J,t} \cup S_{L,t}$, where $S_{J,t}$ and $S_{L,t}$ are the corresponding tth subdomains of S_J and S_L ($t = 1,\ldots, T$).

The first smoothing procedure defines a new weight $w_{\texttt{BR1},ti}$ for each $i \in S_{\texttt{large},t}$ as

$$w_{\texttt{BR1},ti} = \begin{cases} \dfrac{\sum_{j \in S_{\texttt{large},t}} w_{tj}}{|S_{\texttt{large},t}|} & \text{if } S_{J,t} \neq \emptyset \\ w_{ti} & \text{otherwise} \end{cases} \tag{3}$$

where w_{tj} is the final sample weight and $|S_{\texttt{large},t}|$ is the size of $S_{\texttt{large},t}$. It should be noted that in the presence of jumpers in $S_{\texttt{large},t}$, the smoothed weight $w_{\texttt{BR1},ti}$ is constant for all firms $i \in S_{\texttt{large},t}$ and is equal to the average of the original weights of the jumpers and `large` firms in this subdomain. In other words, this procedure completely removes the variability of the weights in $S_{\texttt{large},t}$ if jumpers were found in it, thus implying that all the

firms in $S_{J,t}$ or $S_{L,t}$ represent the same number of firms in the finite population. The weights $w_{\texttt{BR1}}$ are referred to as BR1-weights throughout the rest of this article (BR stands for Beaumont and Rivest, who first proposed the smoothing procedure).

The BR1-weights are attractive because they are easy to implement and are suitable for multipurpose studies (like the study analyzed in Section 6) since these weights do not depend on a specific variable of interest. On the other hand, the BR1-weights of the firms in $S_{L,t}$ may become very large if the weights of the firms in $S_{J,t}$ are extremely influential (as is the case in the real data example presented in Subsection 6.3). To address this problem, Beaumont and Rivest (2009) suggested a modified smoothing procedure which consists of two steps: a trimming step and an adjustment step. In the trimming step, the weights of the jumpers in $S_{J,t}$ are trimmed using the average of the weights in $S_{\texttt{large},t}$ as a trimming threshold. In other words, the weights in $S_{J,t}$ are assigned the value $w_{\texttt{BR1},t}$ as defined in Equation (3). The weights in $S_{L,t}$ remain unchanged. The procedure is repeated for each $t = 1,\ldots, T$ with $S_{J,t} \neq \emptyset$. In the adjustment step, a correction factor C_t is computed as follows

$$C_t = \begin{cases} \dfrac{\sum_{i\in S_{J,t}} w_{ti} + \sum_{i\in S_{L,t}} w_{ti} + \sum_{i\in S_{\texttt{small},t}} w_{ti}}{\sum_{i\in S_{J,t}} w_{\texttt{BR1},ti} + \sum_{i\in S_{L,t}} w_{ti} + \sum_{i\in S_{\texttt{small},t}} w_{ti}} & \text{if } S_{J,t} \neq \emptyset \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

The factor C_t is further used to adjust the weights of all the firms in $S_{\texttt{large},t}$ and all the firms in the corresponding subdomain of small firms (i.e., $S_{\texttt{small},t}$) as follows

$$w_{\texttt{BR2},ti} = \begin{cases} C_t w_{\texttt{BR1},ti} & \text{if } i \in S_{J,t} \\ C_t w_{ti} & \text{if } i \in S_{L,t} \text{ or } i \in S_{\texttt{small},t} \end{cases} \tag{5}$$

We refer to these modified weights as BR2-weights. Since the original weights of the jumpers are at least as large as the BR1-weights, the coefficient C_t is at least 1 (with equality if either $S_{J,t} = \emptyset$ or $S_{L,t} = \emptyset$). In practice, the value of C_t is expected to be only slightly greater than 1 given that very few jumpers (if any) can be found in $S_{\texttt{large},t}$. In this case, it follows from (5) that the modified smoothing procedure will only slightly affect the weights of the firms in $S_{L,t}$ and $S_{\texttt{small},t}$, while still reducing the weights of the jumpers in $S_{J,t}$.

It is important to emphasize that the validity of the two weight smoothing procedures in Equations (3) and (5) relies on the assumption of independence between the variable of interest y and the weights in a given subdomain before smoothing. In other words, the weights are required to be *noninformative* in that subdomain. For this reason, these smoothing procedures may not be suitable for propensity-adjusted weights, which may depend on additional covariates that can be associated with y, even after conditioning on $Z_{\texttt{surv}}$ and other characteristics that define a subdomain (e.g., geographic area and business sector). In such cases, one possible approach could be to apply the smoothing procedures to the 'cell-adjusted' weights $w_{\texttt{str}}$ defined in Section 2. This option may not be feasible if only one final set of nonresponse-adjusted weights has been released to an analyst, which is often the case in practice. The independence condition will also be violated if the distribution of the y-variable of the jumpers is different from the

distribution of the y-variable of the rest of the firms in the subdomain that the jumpers moved into. Using the `BR1`- and `BR2`-weights in situations where the conditional independence assumption is violated may lead to biased estimators. In general, parametric modeling of the weights might be necessary to smooth the weights; see Chen et al. (2017) for a detailed review.

We use $\hat{\mu}_{\texttt{pa}}$, $\hat{\mu}_{\texttt{str}}$, $\hat{\mu}_{\texttt{BR1}}$ and $\hat{\mu}_{\texttt{BR2}}$ to denote the corresponding domain mean estimators obtained from Equation (2) by replacing the weights w with the weights $w_{\texttt{pa}}$, $w_{\texttt{str}}$, $w_{\texttt{BR1}}$, and $w_{\texttt{BR2}}$, respectively. The subscript d is added to emphasize the domain mean estimation where necessary.

5. Variance Estimation

The linearization variance estimator of the weighted mean with propensity-adjusted weights is given in the work of Kim and Kim (2007). Expanding this estimator to obtain the linearization variance estimator of the domain mean estimator $\hat{\mu}_d$ with nonresponse-adjusted weights, defined in the previous sections, can be challenging due to the presence of jumpers and weight smoothing. Replication methods, in particular bootstrap variance estimators, are recommended in these cases (Mashreghi et al. 2016). Beaumont and Rivest (2009) used the rescaled bootstrap method (Rao and Wu 1988; Rao et al. 1992) to estimate the variance of the population total estimator obtained from a stratified sample of firms. The formula for bootstrap (rescaled) weights that account for sampling without replacement from a finite population is given in Beaumont and Patak 2012. In this study, we used the 'doubled half' bootstrap method devised by Antal and Tillé (2014) for sampling designs without replacement.

5.1. Bootstrap Variance Estimator

Let $\tilde{S}$ be a stratified random sample from the finite population of firms. Assume that the information about the response status of each firm is contained in $\tilde{S}$ as a binary variable `response`. This variable takes value 1 for respondents and 0 otherwise. The following steps describe the implementation of the method for estimating the variances of $\hat{\mu}_d$, calculated using the nonresponse-adjusted weights defined in the previous sections, with or without smoothing.

1. Obtain a bootstrapped sample $\tilde{S}^*$ from $\tilde{S}$ by applying the 'doubled half' method to each stratum of $\tilde{S}$.
2. The sample $\tilde{S}^*$ inherits the response status contained in the variable `response` from the parent sample $\tilde{S}$.
3. Estimate the response probabilities $\hat{p}_i^*$ from $\tilde{S}^*$ (for example, using a logistic regression model).
4. A subset of $\tilde{S}^*$, for which the value of `response` is equal to 1, is used as a bootstrap replicate S^* of a sample of respondents S. Calculate $w^*_{\texttt{pa},i}$, $w^*_{\texttt{str},i}$, $w^*_{\texttt{BR1},i}$ and $w^*_{\texttt{BR2},i}$ for each $i \in S^*$.
5. Estimate the bootstrap domain means μ_d^* using one of the estimators: $\hat{\mu}^*_{\texttt{pa}}$, $\hat{\mu}^*_{\texttt{str}}$, $\hat{\mu}^*_{\texttt{BR1}}$ and $\hat{\mu}^*_{\texttt{BR2}}$.

6. Repeat steps 1–5 R times to obtain a sequence of bootstrap estimates

$$\hat{\mu}_d^{*(1)}, \hat{\mu}_d^{*(2)}, \ldots, \hat{\mu}_d^{*(R)}$$

for each of the four estimators.

7. Estimate $V(\hat{\mu}_d)$ as

$$\hat{V}(\hat{\mu}_d) = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\mu}_d^{*(r)} - \bar{\hat{\mu}}_d)^2 \tag{6}$$

where $\bar{\hat{\mu}}_d = \frac{1}{R}\sum_{r=1}^{R}\hat{\mu}_d^{*(r)}$.

6. Application to an Organizational Performance Survey

In this section we illustrate domain mean estimation using data from the Ontario Leading Indicators Project (OLIP) survey. OLIP is a cross-sectional business survey designed for auditing the organizational performance in preventing work-related injury and illness using reliable and validated indicators (measures) (Institute for Work and Health 2011). The target population consists of firms in Ontario, Canada, that were registered with the Workplace Safety and Insurance Board (WSIB, http://www.wsib.on.ca/), an organization responsible for workers' compensation. The number of full-time equivalents (FTEs), derived from the WSIB payroll information, was used as the measure of firm size at the design stage. Only firms with at least one FTE in 2009, and in one of the following industries were inluded: agriculture, manufacturing, service, education, municipal, healthcare, pulp and paper, construction, transportation, electrical and utilities (Institute for Work and Health 2011).

Respondents were asked to classify the size of their firm, $Z_{\texttt{surv}}$, into five categories:

$$Z_{\texttt{surv},i} = \begin{cases} \texttt{very small} & \text{if firm } i \text{ has 1 to 4 employees} \\ \texttt{small} & \text{if firm } i \text{ has 5 to 19 employees} \\ \texttt{medium} & \text{if firm } i \text{ has 20 to 99 employees} \\ \texttt{large} & \text{if firm } i \text{ has 100 to 299 employees} \\ \texttt{very large} & \text{if firm } i \text{ has 300 or more employees} \end{cases}$$

Considering each level of $Z_{\texttt{surv}}$ as a domain in the target population, the focus here is on estimating the population domain means and their 95% confidence intervals (CI) for six variables of interest: Safety Practices (SP), Health and Safety Leadership (HSL), Ergonomic Practices (EP), Disability Prevention (DP), People-Oriented Culture (POC), and Organizational Performance Measure (OPM). These variables were derived from the key OHS measures collected by the OLIP study (Institute for Work and Health 2013). Each measure was presented in the questionnaire as a scale with a variable number of items. Each item had five possible responses, with values from 0 to 4, with higher values corresponding to better performance. The individual value of a final variable of interest was obtained by taking the average of individual responses on the items within a given scale, resulting in a continuous variable in the range [0; 4].

The selected sample ($\tilde{S}$) was designed as a stratified simple random sample, with the strata formed by two levels of the firm size $Z_{\texttt{des}}$ derived from the FTE data in 2009 (`small` if the firm has 1 to 19 employees and `large` if the firm has 20 or more employees), five geographical regions in Ontario, and ten business sectors as reported on the WSIB data registry, resulting in a total of 100 strata. For agriculture, manufacturing and service sectors, 300 firms were randomly sampled from a stratum. In the other seven sectors, 150 firms were sampled randomly from the stratum. Otherwise, all firms in the stratum were sampled. Base weights (b) were calculated from these sampling fractions. The total size of $\tilde{S}$ was $\tilde{n} = 12{,}767$, of which $n = 1{,}639$ were respondents, indicating a low response rate ($\approx$ 13%). Finally, $\tilde{S}$ was linked to the WSIB administrative data at the firm level, which allowed us to obtain information about compensation claims activity for all firms in $\tilde{S}$.

6.1. Nonresponse Adjustment

The nonresponse-adjusted weights $w_{\texttt{str}}$ and $w_{\texttt{pa}}$ have been calculated as described in Section 3. Results from the logistic regression model, fitted with and without base weights (b) to obtain $w_{\texttt{pa}}$, were nearly identical (Kim and Kim 2007). Three design variables ($Z_{\texttt{des}}$, geographic region and business sector) along with various covariates related to firms' claims activity were included in the response regression model. The estimated regression coefficient of $Z_{\texttt{des}}$ was 0.76, which indicates higher response probabilities for `large` firms, after adjusting for all the other factors. Firms in municipal, electrical and utilities, and construction sectors had the highest negative coefficients (-1.15, -1.15 and -1.14, respectively), which indicates the lowest response probabilities for these sectors. Interestingly, firms with higher claim rates showed higher response probabilities: the regression coefficient was equal to 0.14. All five coefficients had p-values smaller than 0.001.

We validated the nonresponse-adjusted weights by comparing the weighted estimators from $\tilde{S}$ using the weights b (complete data case) with the weighted estimators from S using the weights b (and thus completely ignoring the nonresponse), $w_{\texttt{str}}$, and $w_{\texttt{pa}}$, respectively, for a number of variables in $\tilde{S}$ (Lohr et al. 2016). Both nonresponse-adjusted estimators showed closer correspondence with the complete case estimates when compared to the unadjusted estimator (the details are available from the corresponding author).

6.2. Application of Smoothing Procedures for D = 5

In OLIP, the five levels of $Z_{\texttt{surv}}$ define $D = 5$ domains of interest: `very small`, `small`, `medium`, `large`, and `very large`, which can also be denoted using the notation in Section 4 as $S_{\texttt{very small}}$, $S_{\texttt{small}}$, $S_{\texttt{medium}}$, $S_{\texttt{large}}$, and $S_{\texttt{very large}}$, respectively. Five geographic regions and ten business sectors partition each of the five domains into 50 subdomains. The firms that were registered as small firms in the WSIB data ($Z_{\texttt{des}} =$ `small`) but reported more than 19 employees in the survey were declared jumpers. Using this definition, 45 jumpers were found in $S_{\texttt{medium}}$, 12 in $S_{\texttt{large}}$, and 9 in $S_{\texttt{very large}}$ – for a total of 66 jumpers.

Weight smoothing procedures described in Section 4 were adapted to each subdomain of `medium`, `large`, and `very large` firms in which jumpers were found. For example, for a given t, if jumpers were found in $S_{\texttt{medium},t}$ and $S_{\texttt{large},t}$, but not in $S_{\texttt{very large},t}$, the `BR1`-weights were computed by applying Equation (3) independently to $S_{\texttt{medium},t}$ and

$S_{\texttt{large},t}$. The corresponding BR1-weight was then used as a trimmimg threshold to complete the trimming step of the BR2-procedure as described in Section 4. Next, letting $S_{\texttt{medium},t} = S_{J_1,t} \cup S_{M,t}$ and $S_{\texttt{large},t} = S_{J_2,t} \cup S_{L,t}$, the correction factor was obtained analogously applying to Equation (4) as

$$\frac{\sum_{i\in S_{J_1,t}} w_{ti} + \sum_{i\in S_{J_2,t}} w_{ti} \sum_{i\in S_{M,t}} w_{ti} + \sum_{i\in S_{L,t}} w_{ti} + \sum_{i\in S_{\texttt{very small},t}} w_{ti} + \sum_{i\in S_{\texttt{small},t}} w_{ti}}{\sum_{i\in S_{J_1,t}} w_{\text{BR1},ti} + \sum_{i\in S_{J_2,t}} w_{\text{BR1},ti} \sum_{i\in S_{M,t}} w_{ti} + \sum_{i\in S_{L,t}} w_{ti} + \sum_{i\in S_{\texttt{very small},t}} w_{ti} + \sum_{i\in S_{\texttt{small},t}} w_{ti}}$$

Finally, the weights of the firms in $S_{\texttt{very small},t}$, $S_{\texttt{small},t}$, $S_{\texttt{medium},t}$ and $S_{\texttt{large},t}$ were updated as in Equation (5). The weights of the firms in $S_{\texttt{very large},t}$ remained unchanged since no jumpers were found in this subdomain.

In this example we defined jumpers as firms that were registered as small ($Z_{\texttt{des}} =$ small) but reported their size in the survey as medium, large or very large ($Z_{\texttt{surv}} =$ medium, large, very large). It is worth noting that this definition of jumpers is not fixed and can be adjusted depending on the particular problem. For example, we performed a sensitivity analysis using another definition, by which the firms defined as ($Z_{\texttt{des}} =$ small, $Z_{\texttt{surv}} =$ medium) were not declared jumpers (45 firms in this example). The resulting means in the domain of medium firms were virtually identical under both definitions (data not shown), since the weights of these 45 firms were not extreme for the vast majority of cases. The flexibility in defining jumpers may also be useful if the categories of the variable $Z_{\texttt{surv}}$ do not allow a clear-cut partition into small and large firms as defined by the categories of the variable $Z_{\texttt{des}}$. In general, we recommend exploring the distribution of the weights in the domains defined by $Z_{\texttt{surv}}$ before making a decision about the partition of the firms into small and large to confirm that the proposed definition of jumpers is sensible for a given problem. Once jumpers have been identified, the smoothing procedures described in the previous sections should be applied to the domains (defined by $Z_{\texttt{surv}}$) that contain jumpers.

6.3. Estimation of Domain Means by Firm Size

In this section we illustrate the estimation of the domain means of large and very large firms. The number of respondents in each of these domains were 260 and 166, respectively.

The CV of the base and nonresponse-adjusted weights, with and without smoothing, are shown in Table 1. The smoothed weights, w_{BR1} and w_{BR2}, were applied to the cell-adjusted

Table 1. The coefficient of variation (CV, %) of base (b), propensity-adjusted (w_{pa}), cell-adjusted (w_{str}), and two sets of smoothed weights (w_{BR1}, w_{BR2}) in the domain of large and very large firms in the OLIP sample. The smoothed weights were calculated by applying the smoothing procedures defined in Equations (3) and (5) to the cell-adjusted weights (w_{str}).

	Weights				
Domain	b	$w_{\texttt{pa}}$	$w_{\texttt{str}}$	w_{BR1}	w_{BR2}
Large	156.3	341.5	513.2	246.6	586.6
Very large	265.6	642.3	581.6	294.1	353.4

weights $w_{\texttt{str}}$ to ensure that the conditional independence assumption is met. The numbers in the table indicate the impact of the jumpers on the CV of the base weights, an additional increase in the CV caused by the nonresponse adjustment, and the impact of the two smoothing procedures on reduction in the CV.

Figures 1 and 2 (Online supplemental data) show the estimated means and their 95% CIs for the domain of `large` and `very large` firms, respectively. To ensure that the lower and upper bounds of the CIs lie within the range [0; 4], we applied a `logit`-type transformation $g(\hat{\mu}) = \frac{\hat{\mu}}{4-\hat{\mu}}$, similarly to a commonly used approach for proportions (Lachin 2011). The resulting CIs for $\hat{\mu}$ are almost identical to the symmetric CIs obtained without a transformation, except for a few cases where the estimated upper bounds of the symmetric CIs were above 4. In these few cases, the resulting CIs are noticeably asymmetric.

Overall, both nonresponse-adjusted estimators (without smoothing) produced very similar estimates for the majority of study variables in both domains. This may reflect the fact that the design variables were major predictors for response. The discrepancies between the estimates with and without smoothing of the weights can be seen in some *y*-variables, most noticeably in the OPM variable in the domain of `very large` firms. A substantially lower domain mean estimate $\hat{\mu}_{\texttt{BRI}}$ was obtained for EP in the domain of `very large` firms (see Online supplemental data, Figure 2). With further checking we found that this dramatic difference in the estimated mean was caused by a single observation that corresponded to a `very large` firm with a very small value of EP and a very large `BR1`-weight. It can also be seen from the graph that the estimator with `BR2`-weights nicely addressed this problem. In the next section a simulation study is used to compare the performance of the weighted estimators in terms of bias and efficiency.

7. Simulation Study

The extensive simulation study described in this section was designed with two goals in mind. First, we hypothesized that two parameters of the finite population, namely, (1) $p_{\texttt{jump}}$ the percent of jumpers among `small` firms, and (2) δ – the difference between the mean of the *y*-values of the jumpers and the mean of the *y*-values of the firms in the subdomain they jumped into, may have a considerable impact on the performance of the weighted estimators. In particular, we anticipated that the smoothed estimators will be biased when $\delta > 0$, since in this case the assumption of independence between the weights and the *y*-variable within a subdomain fails. We also expected that the increase in $p_{\texttt{jump}}$ may improve the performance of the estimators. Therefore, we created different scenarios to explore the performance of the estimators for various combinations of δ and $p_{\texttt{jump}}$. Second, we aimed to test the performance of Antal-Tille's bootstrap variance estimator, described in Section 5 and used in the analysis of the OLIP data. We also studied the error rate of the CIs constructed from the bootstrap estimates.

7.1. Data Generation

7.1.1. Generation of a Finite Population

The finite population of firms was reconstructed from the selected sample $\tilde{S}$ in the OLIP study by repeating each row in $\tilde{S}$ as many times as its base weight *b*. A

continuous variable for FTEs in 2009 was used to construct the variables $Z_{\texttt{des}}$ and $Z_{\texttt{surv}}$ as defined in Section 6. Five geographic regions in Ontario, $\{G_1,\ldots,G_5\}$, were used together with the two categories of $Z_{\texttt{des}}$ to form 10 design strata, which can be enumerated in pairs: for example, the pair (`small`, G_1) corresponds to the stratum of `small` firms in region G_1. Let N and N_h ($h = 1,\ldots,10$) be the sizes of the finite population and the stratum h, respectively, as defined by design. The values of the y-variable were generated independently within each stratum h from a normal distribution with mean μ_h and a common standard deviation of 0.1. The values of μ_h were chosen such that the `large` firms have higher mean values than the `small` firms. For illustration, we estimate the mean of the y-variable in the domain of `very large` firms ($Z_{\texttt{surv}} = \texttt{very large}$).

7.1.2. Generation of Jumpers in the Population

First, $p_{\texttt{jump}}$ (%) of the firms from all the `small` firms in the population ($Z_{\texttt{des}} = \texttt{small}$) were randomly assigned to be jumpers. For simplicity, we assumed that all the generated jumpers changed their firm size from `small` to `very large` and changed the original value of $Z_{\texttt{surv}}$ for these firms to `very large`. Next, the values of the y-variable of the jumpers were re-generated from a normal distribution with mean $\mu_{h*} + \delta$ and standard deviation 0.1, where $\mu_h{}^*$ corresponds to the mean used to create the y-values of `large` firms from the same geographic region and $\delta \geq 0$ is a predefined constant. We set $\delta \in \{0, 0.15, 0.25, 0.5\}$ to reflect various scenarios when the mean of the y-variable of the jumpers is up to 12.5% greater than the mean of the y-variable of the `large` firms. The assumption of conditional independence between the y-variable and the weights in a given subdomain holds when $\delta = 0$. Scenarios with $\delta > 0$ reflect the fact that jumpers may be very different from the firms in the subdomain they jumped into. The assumption of conditional independence is not met in this case and the results will serve as sensitivity analyses. We set $p_{\texttt{jump}} \in \{1\%, 2\%, 5\%, 7\%, 10\%\}$ to reflect that in most practical cases the percent of jumpers in the population is not expected to be greater than 10% (Favre-Martinoz et al. 2015). Since approximately 88% of the firms in the artificially created population have $Z_{\texttt{des}} = \texttt{small}$, these values reflect most practical scenarios fairly well. All possible combinations of $p_{\texttt{jump}}$ and δ define 20 scenarios of interest.

7.1.3. Obtaining a Sample of Respondents

Business surveys are prone to nonresponse with variable response rates. US government business surveys that require mandatory participation, can have a response rate as high as 80% or higher (U.S. Department of Labor 2019). Low response rates (under 50%) might be expected in smaller surveys, in which participation is not mandatory (Cook et al. 2009). In this study, we decided to explore the properties of the three weighted estimators under a low response rate (20%) and a target response rate (80%). A sample of respondents was selected from the finite population in a two-step process. First, a full sample $\tilde{S}$ was selected as a stratified random sample without replacement with sample inclusion probabilities equal to $\tilde{n}_h/N_h$ for predetermined $\tilde{n}_h$ (see Table 2). The values of $\tilde{n}_h$ were selected to allow sufficient variation in sample weights between the strata of `small` and `large` firms. In the

Table 2. Numerical values of the parameters used in the data generation process: N_h – the hth stratum size in the population; $\tilde{n}_h$ – the hth stratum size in the selected sample $\tilde{S}$; $b_h = N_h/\tilde{n}_h$ (rounded to a whole number); μ_h – the mean of the y-variable in the hth stratum in the population.

Design stratum	N_h	$\tilde{n}_h$	b_h	μ_h
(small, G_1)	62,938	252	250	2.3
(small, G_2)	36,826	246	150	2.0
(small, G_3)	29,164	243	120	1.8
(small, G_4)	24,273	405	60	2.1
(small, G_5)	13,960	465	30	2.5
(large, G_1)	9,468	947	10	3.6
(large, G_2)	4,651	1,163	4	3.4
(large, G_3)	4,864	973	5	3.2
(large, G_4)	2,734	1,367	2	3.5
(large, G_5)	1,478	1,478	1	3.8

Note: (small, G_1) refers to a stratum of small firms in the geographic region G_1.

second step, the sample of respondents was selected using Poisson sampling with two sets of response probabilities p_i that correspond to the two response rates:

$$\text{logit}(p_i) = \begin{cases} -3.5 + 0.5z_i + g_{1i} + g_{2i} + g_{4i} + g_{5i} & \text{for response rate} = 20\% \\ -0.45 + 0.5z_i + g_{1i} + g_{2i} + g_{4i} + g_{5i} & \text{for response rate} = 80\% \end{cases} \quad (7)$$

where z is the firm size variable; the variables g_1, g_2, g_4, g_5 are dummy variables for the geographic regions G_1, G_2, G_4, G_5, respectively (G_3 was defined as a reference); and $\text{logit}(p_i) = \frac{p_i}{1-p_i}$.

7.2. Performance of the Weighted Estimators

We examined the performance of the weighted estimators with propensity-adjusted weights, with and without smoothing, for estimating the mean of the y-variable in the domain of very large firms ($Z_{\text{surv}} =$ very large) under 20 scenarios of interest and two response rates: 20% and 80%. The estimated response probabilities $\hat{p}_i$, used to construct the propensity-adjusted weights w_{pa}, were obtained by fitting the logistic regression model with firm size and region as regressors. The smoothing procedures were applied to the propensity-adjusted weights.

The simulations were repeated 5,000 times for each scenario. In scenarios with $p_{\text{jump}} = 0\%$, the average sample size of the domain of very large firms is 140 and 290 under a 20% and an 80% response rate, respectively. In scenarios with $p_{\text{jump}} = 1\%$, the average number of jumpers equals 2 and 12 under a 20% and an 80% response rate, respectively. In scenarios with $p_{\text{jump}} = 10\%$, the average number of jumpers equals 17 and 115 under a 20% and an 80% response rate, respectively.

The CV of the weights, relative bias (RB) and relative efficiency (RE) were computed for each scenario. The reported CV of the weights is the average of the 5,000 ratios of the standard deviation of the weights to their mean obtained from each simulated set. The RE of the estimators with smoothed weights was computed relative to the estimator with propensity-adjusted weights.

7.2.1. CV of the Weights

It follows from the results in Table 3, that the CV of $w_{\texttt{pa}}$ is decreasing with the increase in $p_{\texttt{jump}}$. In contrast, the CV of $w_{\texttt{BR2}}$ is increasing with the increase in $p_{\texttt{jump}}$. The CV of $w_{\texttt{BR1}}$ stays almost constant for different values of $p_{\texttt{jump}}$. Both BR1- and BR2-weights have lower CV than the propensity-adjusted weights. The CV of the weights is similar for both response rates as expected.

7.2.2. RB and RE of the Estimators when $\delta = 0$

The RB and RE of the estimators when $\delta = 0$ are summarized in Tables 4–5. The results in Table 4 show that the RB of $\hat{\mu}_{\texttt{pa}}$ and $\hat{\mu}_{\texttt{BR1}}$ is close to zero for all values of $p_{\texttt{jump}}$ under a 20% response rate. The RB of the two estimators is essentially equal to zero under an 80% response rate. The RB of $\hat{\mu}_{\texttt{BR2}}$ changes from negative to positive with the increase in $p_{\texttt{jump}}$. This pattern is similar for both response rates with a somewhat noticeable shift in the value of $p_{\texttt{jump}}$ at which the RB crosses zero. Though the values of the RB of $\hat{\mu}_{\texttt{BR2}}$ are higher than the values of the RB of the other two estimators, the increase in bias is not alarming. As follows from the results in Table 5, $\hat{\mu}_{\texttt{BR1}}$ is more efficient than $\hat{\mu}_{\texttt{pa}}$ for any value of $p_{\texttt{jump}}$ and for both response rates. Similarly, $\hat{\mu}_{\texttt{BR2}}$ is more efficient than $\hat{\mu}_{\texttt{pa}}$ for any value of $p_{\texttt{jump}}$ that we studied under a 20% response rate. However, under a higher response rate, $\hat{\mu}_{\texttt{BR2}}$ is less efficient than $\hat{\mu}_{\texttt{pa}}$ for $p_{\texttt{jump}} > 5\%$.

7.2.3. RB and RE of the Estimators when $\delta > 0$

Tables 4 and 5 also contain the RB and RE of the estimators when $\delta = 0.15$. The full results of the sensitivity analyses that compare the RB and RE of the estimators for $\delta \in \{0, 0.15, 0.25, 0.5\}$ are presented in Online Supplemental data, Figures 3–7: Figure 3 displays the RBs of $\hat{\mu}_{\texttt{pa}}$, Figures 4–5 display the RBs of $\hat{\mu}_{\texttt{BR1}}$ and $\hat{\mu}_{\texttt{BR2}}$, and Figures 6–7 display the REs of $\hat{\mu}_{\texttt{BR1}}$ and $\hat{\mu}_{\texttt{BR2}}$, respectively. For lower values of $p_{\texttt{jump}}$ ($< 5\%$), the RB of $\hat{\mu}_{\texttt{pa}}$ increases with the increase in δ under a 20% response rate. However, as $p_{\texttt{jump}}$ increases, the bias stabilizes around 0. Under a higher response rate, the bias of $\hat{\mu}_{\texttt{pa}}$ is relatively small for all values of δ. The bias of $\hat{\mu}_{\texttt{BR1}}$ stays almost constant for different values of $p_{\texttt{jump}}$ and increases with the increase in δ. The bias of $\hat{\mu}_{\texttt{BR2}}$ also increases with the increase in δ, but tends to improve for higher values of $p_{\texttt{jump}}$. Response rate has little effect on the bias of the smoothed estimators. The RE of the smoothed estimators tends to increase as δ increases. The bias constitutes a major part of this increase. The informativeness of the weights in the subdomains, in which the smoothing was applied, is the reason for this bias: the distribution of the y-variable depends on the values of the design variable $Z_{\texttt{des}}$ when $\delta > 0$, thus violating the conditional independence assumption.

7.3. Performance of Antal-Tille's Bootstrap Variance Estimator

To evaluate the performance of Antal-Tille's bootstrap estimator, we implemented the method with $R = 500$ bootstraps for each of the 500 samples simulated as described in Subsection 7.1. The reduction in the number of simulations was needed to reduce the time of computation. To check that the 500 simulated sets are sufficient, we compared the values of the sampling standard errors (SSEs) (defined as the standard deviation of the estimated domain means) of the estimators from 500 versus 5,000 simulated sets and

Table 3. Average coefficient of variation (CV, %) of w_{pa}, w_{BR1} and w_{BR2} under a 20% and an 80% response rate.

Response rate	Weight	p_{jump}, %				
		1	2	5	7	10
20%	$w_{\texttt{pa}}$	400.5	471.0	433.1	392.3	345.2
	$w_{\texttt{BR1}}$	85.3	87.1	83.0	79.4	74.5
	$w_{\texttt{BR2}}$	90.1	108.4	161.1	185.5	205.4
80%	$w_{\texttt{pa}}$	329.8	314.1	247.1	221.0	194.2
	$w_{\texttt{BR1}}$	64.7	66.9	64.2	64.2	63.6
	$w_{\texttt{BR2}}$	73.6	93.6	138.4	150.3	151.8

Note: p_{jump} – percent of jumpers among `small` firms in the finite population.

Table 4. Relative bias (RB, %) of $\hat{\mu}_{pa}$, $\hat{\mu}_{BR1}$ and $\hat{\mu}_{BR2}$ under a 20% and an 80% response rate.

Response rate	δ	Estimator	p_{jump}, %				
			1	2	5	7	10
20%	0	$\hat{\mu}_{\texttt{pa}}$	−0.052	0.108	0.127	0.136	0.097
		$\hat{\mu}_{\texttt{BR1}}$	−0.048	0.138	0.143	0.146	0.061
		$\hat{\mu}_{\texttt{BR2}}$	−0.737	−0.839	−0.499	−0.184	0.168
	0.15	$\hat{\mu}_{\texttt{pa}}$	−0.567	−0.216	−0.001	−0.055	−0.056
		$\hat{\mu}_{\texttt{BR1}}$	−2.231	−2.718	−3.125	−3.196	−3.251
		$\hat{\mu}_{\texttt{BR2}}$	−2.844	−3.266	−2.578	−1.933	−1.144
80%	0	$\hat{\mu}_{\texttt{pa}}$	0.007	−0.008	−0.019	0.003	0.008
		$\hat{\mu}_{\texttt{BR1}}$	−0.024	−0.009	−0.020	−0.050	−0.008
		$\hat{\mu}_{\texttt{BR2}}$	−0.665	−0.592	0.150	0.441	0.700
	0.15	$\hat{\mu}_{\texttt{pa}}$	−0.127	−0.060	−0.058	0.016	−0.002
		$\hat{\mu}_{\texttt{BR1}}$	−2.135	−2.672	−2.955	−2.874	−2.719
		$\hat{\mu}_{\texttt{BR2}}$	−2.556	−2.594	−1.313	−0.594	0.073

Note: δ – difference between the mean of the y-values of the jumpers and the mean of the y-values of the rest of the firms in the subdomain, in the finite population; p_{jump} – percent of jumpers among `small` firms in the finite population.

Table 5. Relative efficiency (RE, %) of $\hat{\mu}_{BR1}$ and $\hat{\mu}_{BR2}$ under a 20% and an 80% response rate.

Response rate	δ	Estimator	p_{jump}, %				
			1	2	5	7	10
20%	0	$\hat{\mu}_{\texttt{BR1}}$	69	74	75	75	75
		$\hat{\mu}_{\texttt{BR2}}$	21	26	37	48	72
	0.15	$\hat{\mu}_{\texttt{BR1}}$	126	184	317	386	492
		$\hat{\mu}_{\texttt{BR2}}$	131	188	208	180	140
80%	0	$\hat{\mu}_{\texttt{BR1}}$	72	70	72	76	76
		$\hat{\mu}_{\texttt{BR2}}$	53	59	74	140	266
	0.15	$\hat{\mu}_{\texttt{BR1}}$	401	745	1,717	2,143	2,653
		$\hat{\mu}_{\texttt{BR2}}$	503	675	429	211	134

Note: δ – difference between the mean of the y-values of the jumpers and the mean of the y-values of the rest of the firms in the subdomain, in the finite population; p_{jump} – percent of jumpers among `small` firms in the finite population.

obtained virtually identical values in both cases. Next, we computed the estimated standard errors (ESEs) (defined as the average of the 500 estimated standard errors) and the ratios SSE/ESE. Ratios close to 1 indicate good performance of the bootstrap variance estimator, whereas ratios greater than 1 indicate variance underestimation. In addition, the 95% CIs around the estimated domain means were constructed using the 2.5% and 97.5% percentiles of the bootstrap distribution obtained for each simulated dataset. The error rate (ER), defined as the number of times in which the true value of the domain mean was not included in the corresponding CI, was computed.

In the absence of jumpers (i.e., $p_{\texttt{jump}} = 0\%$), the SSE/ESE ratios of $\hat{\mu}_{\texttt{pa}}$ are 1.0123 and 1.0228 under a 20% and an 80% response rates, respectively, which are close to 1, as expected. The SSE/ESE ratios for $p_{\texttt{jump}} \in \{1\%, 2\%, 5\%, 7\%, 10\%\}$ are presented in Table 6. As we can see from these results, the values of the SSE/ESE ratio for $\hat{\mu}_{\texttt{BR2}}$ are close to 1 under both response rates. Also, the SSE/ESE ratios are very close to 1 for all three estimators under an 80% response rate, thus confirming a good performance of the bootstrap variance estimator in these cases. Under a 20% response rate, the variances of $\hat{\mu}_{\texttt{pa}}$ and $\hat{\mu}_{\texttt{BR1}}$ are underestimated for low values of $p_{\texttt{jump}}$, regardless of the value of δ.

In the absence of jumpers (i.e., $p_{\texttt{jump}} = 0\%$), the ER values of $\hat{\mu}_{\texttt{pa}}$ are 5.4% and 5.8% under a 20% and an 80% response rates, respectively, which are close to the nominal 5%, as expected. The ERs for $p_{\texttt{jump}} \in \{1\%, 2\%, 5\%, 7\%, 10\%\}$ are presented in Table 7. The ERs of $\hat{\mu}_{\texttt{pa}}$ improve considerably and get closer to the nominal 5% with the increase in $p_{\texttt{jump}}$ and under a higher response rate. The ERs of the smoothed estimators seem to be seriously affected by the value of δ under both response rates. The increase in bias and distortion of the sampling distribution (see Online supplemental data, Figures 8–11) can potentially explain these abnormally high ERs.

Table 6. The SSE/ESE ratio of $\hat{\mu}_{pa}$, $\hat{\mu}_{BR1}$ and $\hat{\mu}_{BR2}$ under a 20% and an 80% response rate.

Response rate	δ	Estimator	p_{jump}, %				
			1	2	5	7	10
20%	0	$\hat{\mu}_{\texttt{pa}}$	1.4297	1.3439	1.1343	1.1031	1.0310
		$\hat{\mu}_{\texttt{BR1}}$	1.3851	1.2969	1.1424	1.0841	1.0205
		$\hat{\mu}_{\texttt{BR2}}$	1.0348	1.0000	1.0000	1.1062	1.0626
	0.15	$\hat{\mu}_{\texttt{pa}}$	1.4042	1.1689	1.1315	1.0511	1.0492
		$\hat{\mu}_{\texttt{BR1}}$	1.3947	1.3222	1.1511	1.0566	1.0340
		$\hat{\mu}_{\texttt{BR2}}$	1.0000	0.9786	1.1074	1.0351	1.0288
80%	0	$\hat{\mu}_{\texttt{pa}}$	1.0745	0.9828	1.0266	1.0216	0.9543
		$\hat{\mu}_{\texttt{BR1}}$	1.0596	0.9833	1.0044	1.0200	0.9249
		$\hat{\mu}_{\texttt{BR2}}$	0.9718	1.0000	1.0452	1.0502	0.9763
	0.15	$\hat{\mu}_{\texttt{pa}}$	1.0676	0.9888	1.0000	1.0044	1.0408
		$\hat{\mu}_{\texttt{BR1}}$	1.0776	1.0066	1.0000	0.9860	1.0000
		$\hat{\mu}_{\texttt{BR2}}$	1.0503	1.0225	1.1019	0.9920	1.0493

Note: SSE – sampling standard error of the estimator [= standard deviation of the 500 estimated domain means]; ESE – estimated standard error of the estimator [= mean of the 500 estimated standard errors of the bootstraps]. δ – difference between the mean of the y-values of the jumpers and the mean of the y-values of the rest of the firms in the subdomain, in the finite population; p_{jump} – percent of jumpers among `small` firms in the finite population.

Table 7. Error rate (ER, %) of $\hat{\mu}_{pa}$, $\hat{\mu}_{BR1}$ and $\hat{\mu}_{BR2}$ under a 20% and an 80% response rate.

Response rate	δ	Estimator	p_{jump}, %				
			1	2	5	7	10
20%	0	$\hat{\mu}_{\text{pa}}$	12.2	9.2	6.6	11.0	7.4
		$\hat{\mu}_{\text{BR1}}$	12.6	10.4	10.0	11.4	10.0
		$\hat{\mu}_{\text{BR2}}$	21.2	23.0	7.0	8.2	6.6
	0.15	$\hat{\mu}_{\text{pa}}$	32.0	15.4	8.0	5.8	9.2
		$\hat{\mu}_{\text{BR1}}$	48.0	46.4	47.6	58.0	69.0
		$\hat{\mu}_{\text{BR2}}$	81.6	79.0	49.0	31.8	16.2
80%	0	$\hat{\mu}_{\text{pa}}$	6.6	4.4	6.4	6.8	3.8
		$\hat{\mu}_{\text{BR1}}$	7.4	4.8	5.8	7.6	3.4
		$\hat{\mu}_{\text{BR2}}$	37.4	22.0	7.4	11.4	19.2
	0.15	$\hat{\mu}_{\text{pa}}$	13.0	6.0	6.2	6.8	5.6
		$\hat{\mu}_{\text{BR1}}$	76.4	94.4	99.2	100.0	100.0
		$\hat{\mu}_{\text{BR2}}$	96.4	93.2	45.6	15.2	5.6

Note: ER – error rate of the CIs constructed from the percentiles of the bootstrap distribution. δ – difference between the mean of the y-values of the jumpers and the mean of the y-values of the rest of the firms in the subdomain, in the finite population; p_{jump} – percent of jumpers among `small` firms in the finite population.

8. Discussion

In this article, we focused on challenges in estimating the means of domains (defined by the level of firm size) from a business survey in the presence of misclassified frame records (jumpers) and a low response rate. We studied the impact of jumpers on the performance of the weighted ratio estimator with propensity-adjusted weights with and without smoothing of the weights. Our findings and conclusions are based on the results of an extensive simulation study with numerous scenarios of practical relevance.

Our results demonstrate satisfactory performance of the weighted ratio estimator with propensity-adjusted weights for all scenarios under a 20% response rate, in which the percent of jumpers among `small` firms in the finite population was greater than or equal to 7%, and all scenarios under an 80% response rate. Antal-Tille's variance estimator was very close to the sampling variance in these cases and the error rates of the corresponding CIs were satisfactory. For smaller percentages of jumpers and a low response rate, we observed underestimation of variances of the propensity-adjusted domain mean estimator and higher than nominal error rates.

The underestimated variances can, at least partially, be explained by highly variable weights in domains with a very small number of jumpers. Similar findings have been observed in a different context (Li et al. 2011; Landsman and Graubard 2013). Small-sized population domains of `large` and/or `very large` firms, typically seen in populations of businesses, combined with a low response rate, may imply a very small number of respondents in the sample domain that can be used for the estimation. In such cases, the asymptotic properties of the weighted ratio estimator may not apply, resulting in a distorted sampling distribution and high error rates for CI coverage. As expected, the results improve considerably under a higher response rate.

The validity of the two smoothed weighted estimators relies on the assumption of conditional independence between the variable of interest and the nonresponse-adjusted

weights (before smoothing) given the variables that define a subdomain (e.g., firm size collected at the time of survey, geographic region, and business sector). Adapting additional model-assisted methods of improving survey-weighted estimates in the presence of highly variable weights (Chen et al. 2017) to business surveys with misclassified records and exploring their properties empirically is an important direction for future research.

It is important to note that the sampling design in the OLIP and the simulated data, like many other business surveys with voluntary response, can be considered to be a special case of a two-phase sampling design, where a stratified random sample of firms is selected in the first phase and a sample of respondents is self-selected in the second phase. The proposed bootstrap variance estimator does not fully account for the randomness in the second phase, because the response indicator is assumed to be a fixed variable. In general, if the sampling fractions are non-negligible in the first phase (as is the case in the OLIP and the simulated data), both variance components should be accounted for in the variance estimator to avoid variance underestimation (Kim et al. 2006; Beaumont et al. 2015). To address this problem, Kim and Yu (2011) proposed a multiplicative adjustment to the second phase weights and implemented it for the jackknife variance estimator. However, the jackknife variance estimator is not appropriate for our data due to jumpers and literature on this topic in the framework of bootstrap variance estimator appears to be sparse. At the same time, the results of our simulation study demonstrated a very close correspondence of the estimated standard error (ESE) and sampling standard error (SSE) for the propensity-adjusted estimator in the scenario without jumpers, under both response rates. This implies that the missed variance component might be of a small order in the application that we studied. Still, development and implementation of an improved variance estimator in the bootstrap setting is an important direction for future work.

Given the complexity of the jumpers problem, it is instructive to compare among several estimates rather than sticking to a standard approach. We hope that the analysis of the OLIP data serves as a useful demonstration of this strategy. Although we find it encouraging to see that the various estimates produced fairly similar results for some variables of interest in the application that we studied, we acknowledge that it may not be the case in other applications, since the validity of these estimators relies on strong assumptions that were not verified as part of the current analysis. If an analyst observes substantial discrepancies between the estimates, we would encourage them to explore this problem further and try to understand the reason for it (e.g., outliers, informativeness of the weights), which might help to decide which of the estimates would be the most reliable in a given situation. We would also like to emphasize that our simulation results imply that a small number of jumpers combined with a low response rate is the most difficult scenario for making a valid statistical inference from a business survey. If a practitioner suspects that this is going to be the case in their data, the estimator with nonresponse-adjusted weights (without smoothing) would probably be the safest choice, based on our findings.

9. References

Antal, E. and Y. Tillé. 2014. "A New Resampling Method for Sampling Designs Without Replacement: the Doubled Half Bootstrap." *Computational Statistics* 29: 1345–1363. DOI: https://doi.org/10.1007/s00180-014-0495-0.

Beaumont, J.-F. 2008. "A New Approach to Weighting and Inference in Sample Surveys." *Biometrika* 95: 539–553. DOI: http:s//doi.org/10.1093/biomet/asn028.

Beaumont, J.-F., A. Beliveau, and D. Haziza. 2015. "Clarifying Some Aspects of Variance Estimation in Two-Phase Sampling." *Journal of Survey Statistics and Methodology* 3: 524–542. DOI: https://doi.org/10.1093/jssam/smv022.

Beaumont, J.-F., and Z. Patak. 2012. "On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling." *International Statistical Review* 80: 127–148. DOI: https://dx.doi.org/10.1111/j.1751-5823.2011.00166.x.

Beaumont, J.-F., and L.-P. Rivest. 2009. "Dealing with Outliers in Survey Data." In *Handbook of Statistics*, vol. 29A, *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, 247–279. Amsterdam: Elsevier. DOI: https://doi.org/10.1016/S0169-7161(08)00011-4.

Chen, Q., M.R. Elliott, D. Haziza, Y. Yang, M. Ghosh, R.J.A. Little, J. Sedransk, M. Thompson. 2017. "Approaches to improving survey-weighted estimates". *Statistical Science* 32(2): 227–248. DOI: https://doi.org/10.1214/17-STS609.

Cook, S., P. LeBaron, L. Flicker, and T.S. Flanigan. 2009. "Applying Incentives to Establishment Surveys: A Review of the Literature." In Proceedings of the Section on Survey Research Methods: American Statistical Association, May 14–17, 2009. 5639–5647. Available at: http://www.asasrms.org/Proceedings/y2009/Files/400022.pdf (accessed March 2019).

Favre-Martinoz, C., D. Haziza, and J.-F. Beaumont. 2015. "A Method of Determining the Winsorization Threshold, With an Application to Domain Estimation." *Survey Methodology* 41: 57–77.

Haziza, D., and E. 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.

Institute for Work and Health. 2011. "Benchmarking Organizational Leading Indicators for the Prevention and Management of Injuries and Illnesses: Final Report." Available at: https://www.iwh.on.ca/sites/iwh/files/iwh/reports/iwh_report_benchmarking_organizational_leading_indicators_2011.pdf (accessed March 2019).

Institute for Work and Health. 2013. "Measures in the Ontario Leading Indicators Project (OLIP) Survey." Available at: https://www.iwh.on.ca/sites/iwh/files/iwh/reports/iwh_project_olip_about_the_measures_august_2013.pdf (accessed March 2019).

Kim., J.K., and J.J. Kim. 2007. "Nonresponse Weighting Adjustment Using Estimated Response Probability." *The Canadian Journal of Statistics* 35: 501–514. DOI: https://doi.org/10.1002/cjs.5550350403.

Kim, J.K., A. Navarro, and W.A. Fuller. 2006. "Replication Variance Estimation for Two-Phase Stratified Sampling." *Journal of the American Statistical Association* 101: 312–320.

Kim, J.K., and C.L. Yu. 2011. "Replication Variance Estimation Under Two-phase Sampling." *Survey Methodology* 37: 67–74.

Korn, E.L., and B.I. Graubard. 1999. *Analysis of Health Surveys.* New York: Wiley. DOI: https://doi.org/10.1002/9781118032619.fmatter.

Lachin, J.M. 2011. *Biostatistical Methods: The Assessment of Relative Risks.* 2nd edition, Hoboken, NJ: Wiley. DOI: https://doi.org/10.1002/sim.1167.

Landsman, V., and B.I. Graubard. 2013. "Efficient Analysis of Case-control Studies With Sample Weights." *Statistics in Medicine* 32: 347–360. DOI: https://doi.org/10.1002/sim.5530.

Lee, H. 1995. "Outliers in Business Surveys." In *Business Survey Methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott, 503–526. New York: Wiley. DOI: https://doi.org/10.1002/9781118150504.ch26.

Li, Y., B.I. Graubard, and R. DiGaetano. 2011. "Weighting Methods for Population-based Case-control Studies With Complex Sampling." *JRSS(C)* 60: 165–185. DOI: https://doi.org/10.1111/j.1467-9876.2010.00731.x.

Lohr, S.L., M.K. Riddles, and D. Morganstein. 2016. "Tests for Evaluating Nonresponse Bias in Surveys." *Survey Methodology* 42: 195–218.

Little, R.J., and S. Vartivarian. 2003. "On Weighting the Rates in Non-response Weights." *Statistics in Medicine* 22: 1589–1599. DOI: https://doi.org/10.1002/sim.1513.

MacNeil, D., and S. Pursey. 2002. "Dealing With Industry Misclassifications in the Unified Enterprise Survey." In Proceedings of the Survey Methods Section: SSC Annual Meeting, May 2002: 51–56.

Mashreghi, Z., D. Haziza, and C. Léger. 2016. "A Survey of Bootstrap Methods in Finite Population Sampling." *Statistics Surveys* 10: 1–52. DOI: https://doi.org/10.1214/16-SS113.

Nordlöf, H., K. Wijk, and K.-E. Westergren. 2015. "Perceptions of Work Environment Priorities: Are there any Differences by Company Size? – An Ecological Study." *Work* 52: 697–706. DOI: https://doi.org/10.3233/WOR-152123.

Pfeffermann, D., and M. Sverchkov. 1999. "Parametric and Semi-parametric Estimation of Regression Models Fitted to Survey Data." *Sankhya* B 61: 166–186.

Rao, J.N.K. 1966. "Alternative Estimators in PPS Sampling for Multiple Characteristics." *Sankhya: The Indian Journal of Statistics* 28: 47–60.

Rao, J.N.K. and C.F.J. Wu. 1988. "Resampling Inference With Complex Survey Data." *Journal of the American Statistical Association* 83: 231–241. DOI: https://doi.org/10.1080/01621459.1988.10478591.

Rao, J.N.K., C.F.J. Wu, and K. Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology* 18: 209–217.

Valliant, R., J.A. Dever, and F. Kreuter. 2013. "Basic Steps in Weighting." In *Practical Tools for Designing and Weighting Survey Samples*, edited by R. Valliant, J.A. Dever, and F. Kreuter, 307–348. New York: Springer. DOI: https://doi.org/10.1007/978-3-319-93632-1_13.

U.S. Department of Labor; Bureau of Labor Statistics; Office of Survey Methods Research. Household and establishment survey response rates. Chart 2. Establishment surveys: overall unit response rates. Washington DC. Available at: https://www.bls.gov/osmr/response-rates/home.htm (accessed October 2019).

Received March 2019
Revised October 2019
Accepted May 2021

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 1079–1081, http://dx.doi.org/10.2478/JOS-2021-0046

Book Review

Alina Matei[1]

Yves Tillé. *Sampling and Estimation from Finite Populations*. 2020 New York: Wiley, ISBN: 978-0-470-68205-0, 448 pages.

In 2020, Wiley published the book by Yves Tillé entitled 'Sampling and Estimation from Finite Populations'. This is the English version of the book 'Théorie des sondages: échantillonnage et estimation en populations finies' (second edition) originally published in French (Tillé 2019). The book is the outgrowth of 30 years of experience of the author in survey sampling research and practice. Yves Tillé authored more than 70 peer-reviewed journal articles and six books on survey sampling.

Featuring a broad range of topics, the book contains 16 chapters and offers both a classical and a modern view on sampling and estimation, from the history of survey sampling to nonresponse treatment. The book is broken down into two main parts; the first one is dedicated to sampling methods, and the second one to estimation problems. In the sampling part, simple and systematic sampling designs, stratified sampling, unequal probability sampling designs, balanced sampling, cluster and two-stage sampling are presented in detail. Additional topics on spatial sampling, sampling coordination, and multiple survey frames are more generally discussed and included in a single separate chapter. In the estimation part, estimators such as the ratio, the difference, the regression, the poststratified and the calibration estimators are fully covered.

Focusing on the sampling part, the book addresses two topics not fully discussed by the existing books; these are unequal probability sampling designs and balanced sampling. Yves Tillé provides a deep insight into unequal probability sampling designs that make use of first-order inclusion probabilities computed using auxiliary information. The strategy used to estimate a total is given by an unequal probability sampling design (without replacement) and the Horvitz-Thompson estimator. Chapter 5 presents a number of sampling schemes, including systematic sampling with unequal probabilities, Poisson sampling, the Rao-Sampford method, the Brewer method, as well as a few less popular sampling designs, such as order sampling, the pivotal method, and Deville's systematic sampling. All these are given together with variance estimation/approximation of the Horvitz-Thompson estimator. The entropy of a sampling design with unequal probabilities (Hájek 1981) is discussed in Subsection 5.3 (for equal probability sampling designs, see Subsection 3.11). This measure of randomness of a sampling design is then considered in connection with Poisson sampling and maximum entropy design (also known as conditional Poisson sampling). The choice of an optimal unequal probability sampling design is discussed, and the

[1] Institute of Statistics, University of Neuchâtel, Switzerland. Email: alina.matei@unine.ch

author concludes on page 110 that 'unfortunately, there is no ideal method'. Chapter 6 is dedicated to balanced sampling that (approximately) recovers totals of known auxiliary variables from the sample. It offers a major review on this sampling design, mainly focusing on the cube method introduced by Deville and Tillé (2004). The cube method is revisited in Chapter 8, where it is used to provide a sample that is both spread out geographically and balanced on auxiliary information; this method is called 'local cube method' (see Grafström and Tillé 2013).

Chapters 9, 10, and 11 focus on estimation and present design-based estimators that make use of auxiliary information. These are classical estimators such as the ratio, the regression and the postratified estimators. Chapter 12 offers a significant review on the calibration estimator (Deville and Särndal 1992). Together with the article of Särndal (2007), it currently represents one of the most important reviews on this topic from the design-based point of view. The chapter overviews the existing distances and calibration functions, discusses the main algorithm based on the Newton-Raphson procedure, and the use of the bounds in calibration. The chapter ends with a look at the generalized calibration, a very useful method to correct the bias due to non-ignorable nonresponse. This is briefly reconsidered in Chapter 16, which is dedicated to nonresponse treatment.

The book is mainly concerned with the design-based approach of inference. Chapter 13 copes, however, with model-based approach, where the inference is based on a superpopulation model conditionally on the selected sample. A regression model without intercept, relating the variable of interest to the set of auxiliary variables, that it is the usual model in survey sampling, is used as the superpopulation model. Model-based and design-based approaches are not presented as competitors. One finds in this chapter two topics advocated in the previous chapters: calibration estimator and balanced sampling. Model-based approach is shown to support an important design-based estimator, the calibration estimator, which is unbiased under the advocated model. On the other hand, the balanced sampling reduces the anticipated variance of the Horvitz-Thompson estimator of the population total, under a mixed approach, that is both model and design-based.

Chapter 14 on 'Estimation of Complex Parameters' and Chapter 15 on 'Variance Estimation by Linearization' investigate less common topics. Estimation of Lorenz curves, quantile share ratios and Gini indexes given in Chapter 14 are innovative topics in survey sampling books. The author provides an impressive work in Chapter 15 on variance estimation by linearization for complex statistics (such as logistic regression coefficients and Gini index), creating possibly the most extended review on this topic to date.

The book includes comprehensive introductory chapters, making it accessible to a broad audience, including survey statisticians, practitioners and researchers. Given that a set of exercises with summary solutions are available, the book is also an excellent support for advanced courses in survey sampling. Some chapters are more technical, and require more knowledge in survey sampling theory. Without any doubt, the book represents a salient contribution to survey sampling theory. I hope that it will soon be included in the list of the most influential books on survey sampling, such as Särndal et al. (1992) and Lohr (2019).

References

Deville, J.-C., and C.-E. Särndal. 1992. "Calibration estimators in survey sampling." J*ournal of the American Statistical Association* 87: 376–382. DOI: https://doi.org/10.1080/01621459.1992.10475217.

Deville, J.-C., and Y. Tillé. 2004. "Efficient balanced sampling: the cube method." *Biometrika* 91: 893–912. DOI: https://doi.org/10.1093/biomet/91.4.893.

Grafström, A., and Y. Tillé. 2013. "Doubly balanced spatial sampling with spreading and restitution of auxiliary totals." *Environmetrics* 14(2): 120–131. DOI: https://doi.org/10.1002/env.2194.

Hájek, J. 1981. *Sampling from a Finite Population*. Marcel Dekker, New York.

Lohr, S. 2019. *Sampling: Design and Analysis*. CRC Press, second edition.

Särndal, C.-E. 2007. "The calibration approach in survey theory and practice." *Survey Methodology* 33 (2): 99–119.

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

Tillé, Y. 2019. *Théorie des sondages: échantillonnage et estimation en populations finies*. Dunod, Paris, deuxième edition.

sciendo

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 1083–1089, http://dx.doi.org/10.2478/JOS-2021-0047

Editorial Collaborators

The editors wish to thank the following referees and guest editors of theme issues who have generously given their time and skills to the Journal of Official Statistics during the period 1 October 2020 to 30 September 2021. An asterisk indicates that the referee served more than once during the period.

Abe, Naohito, Hitotsubashi Daigaku Institute of Economic Research, Tokyo, Japan
Adiguzel, Feray, Erasmus University Rotterdam, Rotterdam, the Netherlands
Afrifa-Yamoah, Ebenezer, Norwegian University of Science and Technology, Trondheim, Norway
Akande, Olanrewaju, Duke University, Durham, North Carolina, U.S.A.
Alaimo Di Loro, Pierfrancesco, Sapienza University of Rome, Rome, Italy
Amknecht, Paul, IMF, Williamsburg, Virginia, U.S.A.
Andersson, Per Gösta, Stockholm University, Stockholm, Sweden
Ashmead, Robert, Ohio Colleges of Medicine Government Resource Center, Columbus, Ohio, U.S.A.*
Astill, Gregory, USDA, Kansas City, Missouri, U.S.A.
Axelson, Martin, Statistics Sweden, Örebro, Sweden
Bacchini, Fabio, Italian National Institute of Statistics, Rome, Italy*
Bach, Ruben, University of Mannheim, Mannheim, Germany*
Baena, Daniel, UPC, Barcelona, Spain
Baldacci, Emanuele, Eurostat, Bertrange, Luxembourg
Balk, Bert, Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands
Bauder, Donald, U.S. Census Bureau, Washington D.C., U.S.A.*
Bavdaž, Mojca, University of Ljubljana, Ljubljana, Slovenia
Beaumont, Jean-Francois, Statistics Canada, Ottawa, Canada
Benassi, Federico, Italian National Institute of Statistics, Rome, Italy
Benedetti, Roberto, University of Chieti Pescara, Pescara, Italy*
Bentley, Alan, Statistics New Zealand, Wellington, New Zealand
Beręsewicz, Maciej, Poznań University of Economics and Business, Wielkopolska, Poland*
Beresovsky, Vladislav, National Center for Health Statistics, Hyattsville, Maryland, U.S.A.*
Berg, Emily, Iowa State University of Science and Technology, Ames, Iowa, U.S.A.
Bertarelli, Gaia, University of Pisa, Pisa, Italy
Bethlehem, Jelke, Leiden University, Hazerswoude-Rijndijk, the Netherlands
Białek, Jacek, University of Lodz, Lodz, Poland*
Biffignandi, Silvia, University of Bergamo, Bergamo, Italy
Bijlsma, Ineke, Maastricht University, Maastricht, the Netherlands*

D'Elia, Enrico, Italian National Institute of Statistics, Rome, Italy
De Haan, Jan, Statistics Netherlands, The Hague, the Netherlands
De Waal, Ton, Statistics Netherlands, The Hague, the Netherlands*
De Wolf, Peter-Paul, Statistics Netherlands, The Hague, the Netherlands
Di Gennaro, Luca, National Statistics Office, Valletta, Malta*
Di Iorio, Francesca, University of Naples Federico II, Naples, Italy
Dillman, Don, Washington State University, Pullman, Washington, U.S.A.
Dixon, John, U.S. Bureau of Labor, Washington D.C., U.S.A.*
Donze, Laurent, University of Freibourg, Fribourg, Switzerland*
D'Orazio, Marcello, Italian National Institute of Statistics, Rome, Italy*
Draper, Rohan, Statistics Sweden, Solna, Sweden
Drechsler, Jörg, Institute for Employment Research, Nuremberg, Germany
Dumpert, Florian, German Federal Statistical Office, Wiesbaden, Germany*
Durand, Claire, University of Montreal, Quebec, Canada
Durr, Jean-Michel, Caos-consulting, Clermont-Ferrand, France
Earp, Morgan, U.S. Bureau of Labor, Washington D.C., U.S.A.*
Eck, Daniel, University of Illinois, Champaign, Illinois, U.S.A.
Elison, Joanne, University of Southampton, Southampton, UK*
Elliott, Duncan, Office for National Statistics, Newport, UK
Eltinge, John, U.S. Bureau of Labor, Oakton, Virginia, U.S.A.
Ertz, Florian, University of Trier, Trier, Germany
Evangelista, Rui, Eurostat, Luxembourg, Luxembourg
Evans, Thomas, U.S. Bureau of Labor, Washington D.C., U.S.A.*
Fabrizi, Enrico, Catholic University, Piacenza, Italy
Falorsi, Piero, Former Italian National Statistical Institute, Rome, Italy
Farrugia, Naomi, National Statistics Office Malta, Valletta, Malta*
Fleck, Susan, U.S. Bureau of Labor, Washington D.C., U.S.A.*
Flower, Tanya, Office for National Statistics, Newport, UK
Fortier, Susie, Statistics Canada, Ottawa, Canada*
Fox, Kevin, University of New South Wales, Sydney, Australia
Fuller, Wayne, Iowa State University, Ames, Iowa, U.S.A.
Garcia Trejo, Yazmin, U.S. Census Bureau, Suitland, Maryland, U.S.A.*
Gelman, Andrew, Columbia University, New York, U.S.A.
Giesen, Deirdre, Statistics Netherlands, Heerlen, the Netherlands*
Giessing, Sarah, German Federal Statistical Office, Wiesbaden, Germany
Girardin, Valérie, Université de Caen Normandie, Caen, France
Gołata, Elżbieta, Poznań University of Economics and Business, Poznań, Poland
Gonzalez, Jeffrey, U.S. Department of Agriculture, Washington D.C., U.S.A.
Graham, Patrick, Statistics New Zealand, Christchurch, New Zealand*
Gravem, Dag, Statistics Norway, Oslo, Norway
Guha, Saurav, Indian Agricultural Statistics Research Institute, New Delhi, India
Gummer, Tobias, GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany
Gweon, Hyukjun, University of Waterloo, Waterloo, Ontario, Canada
Haas, Georg-Christoph, Institute for Employment Research, Nuremberg, Germany
Haraldsen, Gustav, Statistics Norway, Kongsvinger, Norway*

Madson, Gabriel, RTI International, Durham, North Carolina, U.S.A.
Malmros, Jens, Stockholm University, Stockholm, Sweden
Maples, Jerry, U.S. Census Bureau, Washington D.C., U.S.A.
Marchetti, Stefano, University of Pisa, Pisa, Italy*
Marella, Daniela, Sapienza University of Rome, Rome, Italy
Massing, Natascha. GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany
Mavletova, Aigul, National Research University Higher School of Economics, Mocow, Russia*
McConway, Kevin, Open University, Milton Keynes, UK
Mecatti, Fulvia, University of Milan-Bicocca, Milan, Italy
Modugno, Lucia, Bank of Italy, Rome, Italy
Mohler, Peter, University of Mannheim, Mannheim, Germany*
Morgan, Peter, Cardiff University, Cardiff, UK*
Morrison, Rebecca, National Science Foundation, Arlington, Virginia, U.S.A.
Mothashami, Gholamreza, Ferowsi University of Mashhad, Iran
Mukhopadhyay, Pushpal, SAS Institute Inc., Cary, North Carolina, U.S.A.*
Mule, Vincent, U.S. Census Bureau, Suitland, Maryland, U.S.A.*
Nagaraja, Chaitra. Fordham University, New York, New York, U.S.A.
Neri, Laura, University of Siena, Siena, Italy*
Neuert, Cornelia, GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany*
Newlands, Gemma, Norwegian Business School, Oslo, Norway*
Nichols, Elizaeth, U.S. Census Bureau, Washington D.C., U.S.A.
Norberg, Anders, Statistics Sweden, Solna, Sweden
Oancea, Bogdan, University of Bucharest, Romania
O'Hanlon, Niall, IMF, Washington D.C., U.S.A.
Okonek, Taylor, University of Washington, Seattle, Washington, U.S.A.
Ongena, Yfke, University of Groningen, Groningen, the Netherlands*
Opsomer, Jean, Westat, Rockville, Maryland, U.S.A.
Österholm, Pär , Örebro University, Örebro, Sweden
Pannekoek, Jeroen, Statistics Netherlands, The Hague, the Netherlands*
Park, Mingue, Korea University, Seoul, Republic of Korea
Pascale, Joanne, U.S. Census Bureau, Washington D.C., U.S.A.
Persson, Andreas, Statistics Sweden, Örebro, Sweden*
Peycheva, Darina, Bulgarian Academy of Sciences, Sofia, Bulgaria
Peytcheva, Emilia, RTI International, Research Triangle Park, North Carolina, U.S.A.
Phillips, Keith, Federal Reserve Bank of Dallas, Texas, U.S.A.*
Pijpers, Frank, Statistics Netherlands, The Hague, the Netherlands*
Pinheiro Jacob, Guilherme, Brazilian Institute of Geography and Statistics, Manaus, Brazil
Polidoro, Federico, Italian National Institute of Statistics, Rome, Italy*
Pratesi, Monica, University of Pisa, Pisa, Italy
Proietti, Tommaso, University of Rome, Rome, Italy
Quartier-la-Tente, Alain Quartier-la-Tente, INSEE, Paris, France*
Radjabov, Botir, GOPA Luxembourg, Luxembourg
Rambaldi, Alicia, University of Queensland, Brisbane, Australia
Ranalli, Giovanna, University of Perugia, Perugia, Italy

White, Kirk, Economic Research Service, Washington D.C., U.S.A.*
Willenborg, Leon, Statistics Netherlands, The Hague, the Netherlands
Williams, Douglas, U.S. Bureau of Labor, Washington D.C., U.S.A.*
Wu, Changbao, University of Waterloo, Waterloo, Ontario, Canada
Yan, Ting, Westat, Rockville, Maryland, U.S.A.
Yang, Daniel, U.S. Bureau of Labor, Washington D.C., U.S.A.
Yu, Erica, U.S. Bureau of Labor, Washington D.C., U.S.A.*
Zabala, Felipa, Statistics New Zealand, Wellington, New Zealand*
Zeugner, Stefan, European Commission, Brussels, Belgium
Zhang, Chan, University of Michigan, Ann Arbor, Michigan, U.S.A.*
Zhang Mark (Xichuan), Australian Bureau of Statistics, Belconnen, Australia
Zimmermann, Thomas, Federal Statistical Office, Wiesbaden, Germany

Journal of Official Statistics, Vol. 37, No. 4, 2021, pp. 1091–1094, http://dx.doi.org/10.2478/jos-2021-0048

Index to Volume 37, 2021

Contents of Volume 37, Numbers 1–4

Author Index

Book Reviews